Naturwissenschaftliches Denken im Lehramtsstudium
– Vergleich klassischer und adaptiver Leistungsmessung

Volker Brüggemann, Volkhard Nordmeier

Freie Universität Berlin, Didaktik der Physik, Arnimallee 14, 14195 Berlin
volker.brueggemann@fu-berlin.de, volkhard.nordmeier@fu-berlin.de

Kurzfassung

Die Projekte Ko-WADiS und ValiDiS untersuchen die Kompetenzentwicklung naturwissenschaftlichen Denkens bei Lehramtsstudierenden. Das im Rahmen der ersten Projektphase (Ko-WADiS) entwickelte Testinstrument zur Erfassung dieser Kompetenzen befindet sich aktuell in der Validierungsphase (ValiDiS). Da das bisherige Format im Einsatz sehr zeitaufwändig ist und eine geringe Messgenauigkeit aufweist, wurde eine zweite Version entwickelt: Ein computeradaptiver Multi-Stage-Test. Dieses Testformat ermöglicht im Vergleich zu papierbasierten Instrumenten kürzere Befragungen bei gleichbleibender Messgenauigkeit.
In diesem Beitrag werden das methodische Vorgehen und die Ergebnisse einer Simulationsstudie sowie einer Pilotierungsstudie des adaptiven Formats zusammengefasst. Die Ergebnisse sind vielversprechend: Durch Vergleich beider Versionen konnte eine deutliche Steigerung der Messeffizienz bei adaptiver Testanwendung nachgewiesen werden.

1. Der Ko-WADiS-Test

Der Ko-WADiS-Test ist ein Instrument zur Erfassung von Kompetenzen des naturwissenschaftlichen Denkens (Hartmann et al. 2015; Straube 2016). Er wurde in den Projekten Ko-WADiS und ValiDiS entwickelt und erprobt, um die Entwicklung dieser Kompetenzen bei Lehramtsstudierenden im Verlauf des Studiums zu beobachten.

Im Kontext der Projekte beziehungsweise des Ko-WADiS-Tests wird naturwissenschaftliches Denken definiert als eine Problemlösefähigkeit, angelehnt an vorangegangene Arbeiten von Mayer (2007) zum Problemlösen in den Naturwissenschaften sowie Upmeier zu Belzen und Krüger (2010) zur Arbeit mit naturwissenschaftlichen Modellen. Den dort gelegten theoretischen Grundlagen folgend wird die Kompetenz als Teilbereich der Erkenntnisgewinnung aufgefasst.

Die Operationalisierung im Testinstrument erfolgte anhand von sieben verschiedenen Kompetenzfacetten (Straube, 2016). Jede davon stellt eine unterschiedliche Handlung oder auch Arbeitsphase in naturwissenschaftlichen Untersuchungsprozessen dar (Fragestellungen formulieren, Hypothesen bilden, Untersuchungen planen und durchführen, Daten auswerten, den Zweck von Modellen erkennen, Modelle testen und Modelle abändern), die als exemplarisch für die zu messende Kompetenz angesehen werden. Eine Unterscheidung von Leistungs-/Niveaustufen wurde weder im Kompetenzmodell noch in der Aufgabenstruktur vorgenommen.

Die konkrete Messung erfolgt anhand von dichotomen Multiple-Choice Aufgaben. Von diesen wurde ein Pool an Aufgaben konstruiert, in dem sich für jede der Facetten mehrere Aufgaben wiederfinden: Jeweils drei verschiedene Aufgaben pro Facette und beteiligtem Studienfach (Biologie, Chemie und Physik), also 63 Aufgaben insgesamt (ebd.).

Für projektinterne Befragungen wird das Instrument in einem Multimatrix-Design eingesetzt, um die mehrfache Befragung mit identischen Aufgaben in Längsschnittstudien zu vermeiden. Jedes der neun verwendeten Testhefte beinhaltet dabei 21 Aufgaben, je eine pro Fach und Facette.

Die Auswertung des bisher gewonnene Datensatzes ($N > 10.000$) und anhaltende Validierungsstudien legen nahe, dass Instrument und Aufgaben die valide Messung naturwissenschaftlichen Denkens ermöglichen. Die Messgenauigkeit unterscheidet sich je nach Testheft und liegt im Gesamtdatensatz bei einer EAP/PV-Reliabilität von 0.544 (Hartmann et al., 2015). Dieser Wert ist für sich gesehen als gering zu bezeichnen und ermöglicht zwar eine vorsichtige Beurteilung von Gruppen, nicht jedoch die Diagnose von einzelnen Personen. Er ist jedoch in Einklang mit weiteren rein schriftlichen Testinstrumenten für ähnliche Konstrukte (vgl. Neumann, 2011: 0.55; Terzer, 2013: 0.46; Wellnitz, 2012: 0.59). Da das Instrument nach Abschluss des laufenden Projektes veröffentlicht und weiterhin eingesetzt werden soll, auch an Standorten mit kleinen Stichprobengrößen, muss an dieser Stelle nachgebessert werden. Eine Möglichkeit, für die keine vollständige Neukonzi-

pierung der Aufgaben notwendig ist, bietet die Erstellung einer adaptiven Testversion.

2. Adaptive Tests

Adaptive Testinstrumente passen die Schwierigkeit der gestellten Aufgaben an die jeweiligen Proband*innen an: Nach der Bearbeitung erster Aufgaben durch eine/n Proband*in wird die latente Fähigkeit dieser Person geschätzt, hierfür werden typischerweise ein- bis dreiparametrische Modelle der Item-Response-Theory in Kombination mit zuvor normierten Aufgabenpools verwendet. Die so gewonnene Schätzung der Personenfähigkeit wird genutzt, um als nächstes die bestmöglich passende Aufgabe aus dem Pool zu ziehen (Frey, 2012).

Durch mehrfache Anwendung des Algorithmus kann die Fähigkeitsschätzung im Laufe der Befragung immer genauer durchgeführt und die Auswahl der Aufgaben verfeinert werden. Sobald das Abschlusskriterium des Testinstruments erfüllt wird, wird die Befragung beendet. Dies kann beispielsweise das Erreichen einer maximalen Testlänge oder einer erwünschten Schätzgenauigkeit sein. Das Instrument reagiert somit auf die einzelnen Personen und passt sich an – es ist adaptiv und maximiert die gewonnene Information pro bearbeiteter Aufgabe. Vergleichsstudien zwischen klassischen und adaptiven Versionen einzelner Instrumente zeigen, dass die adaptive Form auf diese Weise zu deutlichen Erhöhungen der Testeffizienz führt (vgl. Weiss, 1982; Hendrickson, 2007).

Die genauere Beschreibung verschiedener adaptiver Testverfahren und die Konzipierung der adaptiven Umsetzung für das vorliegende Testinstrument sind beschrieben in Brüggemann & Nordmeier (2018).

3. Simulationsstudie

Nach der Erstellung von Infrastruktur und Algorithmus des adaptiven Multistage-Tests (MST) war die Frage nach der genauen Struktur und Länge des Tests zunächst offengeblieben, denn es war nicht bekannt, inwieweit das neue Testverfahren sich auf die erreichbare Messgenauigkeit auswirken würde. Der Literatur folgend war mit einer Steigerung der Messgenauigkeit zu rechnen. Zeitgleich sollte aber zur Belastungsreduzierung auf Seiten der Proband*innen die Länge des Instruments reduziert werden, was der Messgenauigkeit zulasten geht. Um hier die richtige Balance zu finden, mussten mehrere Teststrukturen verglichen werden. Da weder beliebig viele Proband*innen noch unbegrenzte Zeit zur Verfügung standen, wurden für diese Vergleiche Simulationsstudien durchgeführt, um die notwendigen Stichprobengrößen zeitnah realisieren zu können. Ermöglicht wurde dieses Vorgehen durch den sehr großen bereits vorhandenen Datensatz.

3.1. Methodik

Die Grundidee der Simulationsstudie ist unkompliziert: Alle infrage kommenden MST-Strukturen wurden vollständig konzipiert und inklusive einzeln passender Aufgabenblöcke erstellt. Anstelle von einer wirklichen Bearbeitung durch Proband*innen aus der Zielgruppe wurden die Befragung mittels jeder einzelnen Struktur an derselben virtuellen Stichprobe aus dem bereits vorhandenen Datensatz simuliert.

Diese Stichproben wurden auf Basis der Längsschnittdatensätze des Projekts erstellt. Dazu wurden in den Simulationen die Tests von einer Gruppe zufällig ausgesuchter Proband*innen des Datensatzes durchlaufen, die gegebenen Antworten wurden aus den Realdaten ausgelesen. Da die Datenmatrix aus den realen Befragungen unvollständig war, (~75% missing by Design) wurden die Ergebnisse aus der Ko-WADiS-Längsschnittstudie mittels eines zweiparametrischen IRT-Modells imputiert.

3.2. Ergebnisse

Abbildung 1 zeigt die erreichten EAP/PV-Reliabilitäten eines simulierten klassischen Tests (FIT) sowie drei verschiedener MST Versionen. Dabei wurden für jeden dieser Tests auch verschiedene Gesamtlängen geprüft. Der FIT wurde durch die Auswahl der – psychometrisch betrachtet – besten Aufgaben des gesamten Pools erstellt. Es handelt sich dabei also um die nach aktuellen Erkenntnissen bestmögliche nicht-adaptive Version des Instruments, die als Maßstab in der Simulation dienen sollte.

Um die Güte der Ergebnisse einzuschätzen, wurden zudem die Daten von echten Befragungen reproduziert: Für alle Proband*innen wurden die Bearbeitungen der von ihnen ausgefüllten, klassischen Testhefte simuliert. Hierbei wurden die Antworten in der Simulation anhand der in Realstudien geschätzten Fähigkeitsparameter erzeugt. Die in Wirklichkeit erreichten Reliabilitäten der einzelnen Hefte konnten in der simulierten Messung mit Abweichungen <1% reproduziert werden.

Durch Vergleiche der in echten Messungen und in den Simulationen geschätzten Fähigkeiten konnte gezeigt werden, dass die Simulationen die EAP/PV Reliabilitäten systematisch um 0.02 höher einschätzten. Vermutlich geschah dies in Folge der in den Simulationen als perfekt angenommenen (und damit überschätzten) Modellpassung, auf deren Grundlage ein Teil der Antworten generiert wurde.

Zusammengefasst konnte durch die Zusammenstellung eines optimierten Testheftes im Vergleich zu den bisherigen Versionen die erwartete Messgenauigkeit um 27% gesteigert werden. Es wird erwartet, dass eine Umsetzung als MST im 1-2-2 Design mit 15 Aufgaben Gesamtlänge dieselbe Genauigkeit erreichen kann. Das stellt eine zusätzliche Reduzie-

rung der Testlänge um 28% dar. Das Instrument wurde diesen Erwartungen gemäß final umgesetzt.

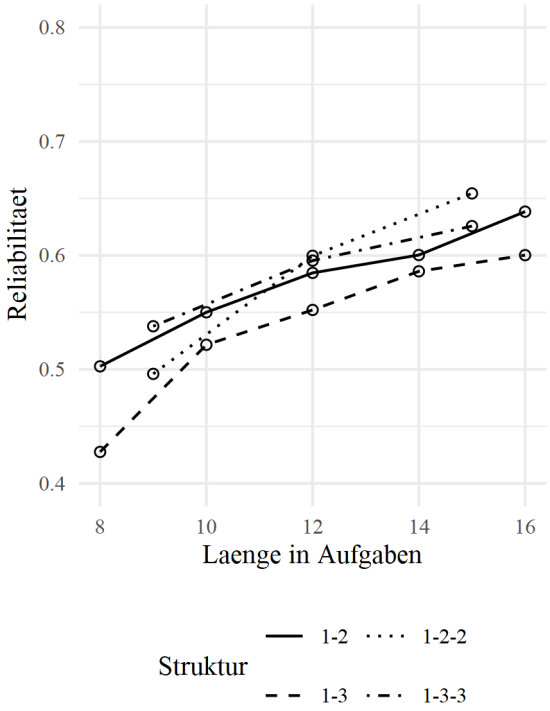

Abb. 1: Simulationsergebnisse; EAP/PV-Reliabilität des klassischen Tests (FIT) und verschiedener Multistage-Test (MST) -Strukturen für verschiedene Testlängen. Die Nummerierung steht für die Anzahl der differenzierten Schwierigkeitsbereiche in den aufeinanderfolgenden Stufen der Tests.

4. Pilotierungsstudie

Um das neue Testformat auch in einer realen Situation zu evaluieren, wurde im ersten Quartal 2019 eine Pilotierungsstudie durchgeführt.

Bei der eigentlichen Zielgruppe des Instruments handelt es sich um die Lehramtsstudierenden der drei naturwissenschaftlichen Fächer. Durch die kleinen Studierendenzahlen in diesen Studiengängen konnte aber keine ausreichend große Stichprobe für die Pilotierung gewonnen werden. Zudem gab es bei praktisch allen infrage kommenden Proband*innen an den teilnehmenden Standorten das Problem, dass die Aufgaben des Instruments durch vorige Befragungen in den Längsschnitterhebungen des Projekts bekannt waren.

Studierende des Sachunterrichts im Grundschullehramt arbeiten jedoch ebenfalls mit naturwissenschaftlichen Inhalten, sofern sie hier ihren Studienschwerpunkt setzen. Somit kommen sie ebenfalls für den Testeinsatz in Frage und wurden bereits in früheren Studien mit dem Instrument untersucht (Straube, 2016). Durch das Ausweichen auf die alternative Probandengruppe konnte an der Freien Universität Berlin eine Stichprobe von N = 283 gewonnen werden. Es wurde damit auch das Problem von Proband*innen umgangen, denen Teile der Testaufgaben bereits im Voraus bekannt gewesen wären.

Die Pilotierung wurde am Standort in Kleingruppen (maximal 30 Personen) durchgeführt. Sie fand unter Aufsicht von Testleiter*innen statt, die mit Befragungszweck und Instrument vertraut waren. Vor Beginn der einzelnen Erhebungen wurde jeweils explizit darauf aufmerksam gemacht, dass es sich um ein adaptives Testinstrument handelt. Dieses Vorgehen wurde gewählt, da

- entgegen üblicher Befragungsformate am Standort keine Antwortkorrektur möglich war und
- die Anpassung der Aufgabenschwierigkeiten zu Motivationsverlusten führen kann (Frey et al., 2009).

4.1. Ergebnisse und Diskussion

Vor der eigentlichen Auswertung der Daten erfolgte eine Betrachtung der ebenfalls erhobenen Bearbeitungszeiten aller Aufgaben. Ziel war die Identifikation von solchen Aufgaben, die sich durch auffällig lange oder kurze Bearbeitung auszeichneten und von Personen, die systematisch signifikant schnellere Aufgabenbearbeitungen aufwiesen als der Rest der Stichprobe. Bei fünf der 283 Personen wurde ein solches Verhalten festgestellt und als durchgängiges Rateverhalten interpretiert, weshalb sie für alle weiteren Auswertungen ausgeschlossen wurden. Auffällige Aufgaben wurden nicht entdeckt.

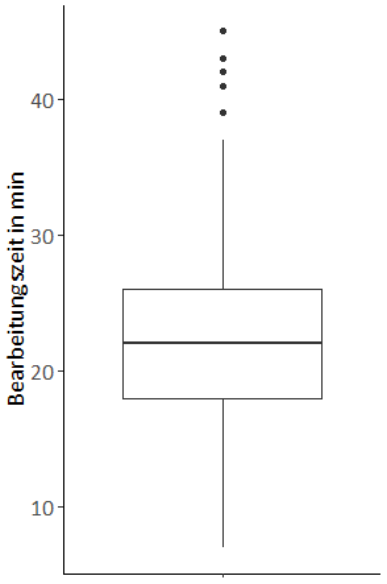

Abb. 2: Gemessene Gesamtbearbeitungszeiten des adaptiven Instruments in der Pilotierungsphase nach Datenbereinigung (siehe erster Absatz Abschnitt 4.1). Punkte stellen Ausreißer dar.

Die Bearbeitungszeit der Befragung lag im Mittel bei 22 Minuten (mit einer Standardabweichung von 6 Minuten, siehe Abb. 2). Für das papierbasierte Instrument werden Bearbeitungszeiten von 35 bis 45 Minuten angelegt. Diese Werte sich allerdings rein ‚anekdotisch' und damit möglicherweise verfälscht sowie gruppenbezogen. Sie entsprechen also dem Zeitpunkt, zu dem bei dem papierbasierten Testeinsatz jeweils die meisten Proband*innen fertig waren, nicht das arithmetische Mittel der Bearbeitungszeit. Für den Vergleich beider Formate wurde daher die mittlere ‚anekdotische' Zeitangabe (40 min) abgeglichen mit der Zeitmarke im adaptiven Test, zu dem die Mehrheit der Stichprobe fertig war: 28 Min. (eine Standardabweichung später als der Mittelwert). Die Bearbeitungszeit wurde durch die neue Testversion also um etwa 30% reduziert. Das Ergebnis deckt sich mit der Reduzierung der pro Person bearbeiteten Aufgaben von 21 im Papierformat zu 15 im adaptiven Test.

Bei der Messung wurde eine Messgenauigkeit von 0.62 (EAP/PV-Reliabilität) erreicht. Sie lag damit über der des Papierinstruments, jedoch unterhalb der Prognose der zuvor durchgeführten Simulationen.

Werden die Unterschiede zwischen den Pilotierungsdaten zur adaptiven Version und den vorliegenden Daten des Papierinstruments zusammenfassend betrachtet (Tabelle 1), zeigt sich eine Reduzierung der Testlänge um ~30% sowie eine gleichzeitige Erhöhung der Messgenauigkeit um ~13%. Der Informationsgewinn pro Item (oder auch die zeitliche Ökonomie des Instruments) konnte also deutlich gesteigert werden.

	Testlänge	EAP-Reliabilität
FIT, Ist-Stand	21	0.544
Simulierter FIT, optimiert	21	0.6
1-2-2 MST, simuliert	15	0.65
1-2-2 MST, gemessen	15	0.62

Tabelle 1: Erreichte Messgenauigkeiten des papierbasierten (FIT) und der adaptiven Multistage-Version (MST) des Instruments in Simulationsstudien und Pilotierung.

Die vorliegenden Ergebnisse sind vermutlich durch die Auswahl der Stichprobe leicht verzerrt. Verglichen zur ursprünglich angepeilten Population war die mittlere Personenfähigkeit der Proband*innen um 0,7 Standardabweichungen geringer (diese Schätzung basiert auf den Daten früherer Erhebungen mit dem Papierinstrument). Die Diskrepanz zwischen angenommener und realer Stichprobenverteilung schränkt die Messgenauigkeit des Instruments ein, da die ursprünglichen Fähigkeitsannahmen stark in die Zusammenstellung der verwendeten Items einflossen. Es wird daher von einer Reduzierung der Messgenauigkeit mit schwachem Effekt ausgegangen.

5. Ausblick

Grundsätzlich wird die Erprobung des adaptiven Testformats als erfolgreich betrachtet. In der Pilotierungsstudie wurde eine deutliche Effizienzsteigerung erreicht, was Ziel des Vorhabens war. Die Messgenauigkeit fiel dabei aber noch etwas niedriger aus als gewünscht. Diesbezüglich ist abzuwarten, welche Resultate eine weitere Erhebung in der eigentlichen Zielgruppe liefert. Zum aktuellen Zeitpunkt ist jedoch nicht klar, ob eine zweite Pilotierungsphase realisierbar ist. Aus diesem Grund wird noch diskutiert, die Testlänge von 15 auf 18 Aufgaben anzuheben. Im Aufgabenpool wären ausreichend passende Aufgaben vorhanden und somit könnte die Messgenauigkeit ausreichend angehoben werden. Das Gegenargument besteht weiterhin in der höheren Belastung der Proband*innen.

Nach Abschluss des Projekts ValiDiS soll das adaptive Instrument zusammen mit der papierbasierten Variante bis Mitte 2020 veröffentlicht werden. (Der Ort der Veröffentlichung des adaptiven Tests ist aktuell noch offen, zur Diskussion steht u. a. die Plattform tet.folio, auf der der adaptive Test bisher für Entwicklungszwecke eingesetzt wird.)

Die Projekte Ko-WADiS und ValiDiS wurden im Rahmen des wissenschaftlichen Transferprojekts „Kompetenzmodelle und Instrumente der Kompetenzerfassung im Hochschulsektor – Validierungen und methodische Innovationen" (KoKoHs) durch das Bundesministerium für Bildung und Forschung (BMBF) gefördert.

6. Literaturverzeichnis

Brüggemann, Volker; Nordmeier, Volkhard (2018): Naturwissenschaftliches Denken im Lehramtsstudium. Computeradaptive Leistungsmessung. In: *PhyDid B*, S. 191–196. Online verfügbar unter http://www.phydid.de/index.php/phydid-b/article/view/892

Frey, Andreas (2012): Adaptives Testen. In: Helfried Moosbrugger und Augustin Kelava (Hg.): Testtheorie und Fragebogenkonstruktion. 2., aktualisierte und überarbeitete Auflage. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg (Springer-Lehrbuch), S. 275–293.

Frey, Andreas; Hartig, Johannes; Moosbrugger, Helfried (2009): Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests. In: *Diagnostica* 55 (1), S. 20–28. DOI: 10.1026/0012-1924.55.1.20.

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: Barbara Koch-Priewe, Anne Köker, Jürgen Seifried und Eveline Wuttke (Hg.): Kompetenzerwerb an Hochschulen: Modellierung und Messung. Zur Professionalisierung angehender Lehrerinnen und Lehrer sowie frühpädagogischer Fachkräfte. Bad Heilbrunn: Verlag Julius Klinkhardt, S. 39–58.

Hendrickson, Amy (2007): An NCME Instructional Module on Multistage Testing. In: *Educational Measurement: Issues and Practice* 26 (2), S. 44–52. DOI: 10.1111/j.1745-3992.2007.00093.x.

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. 1st ed. Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg (Springer-Lehrbuch), S. 177–187.

Neumann, Irene (2011): Beyond physics content knowledge. Modeling competence regarding nature of science inquiry and nature of scientific knowledge. Berlin, Logos.

Straube, Philipp (2016): Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-) Studierenden im Fach Physik. Berlin, Logos.

Terzer, Eva (2013). Modellkompetenz im Kontext Biologieunterricht – Empirische Beschreibung von Modellkompetenz mithilfe von Multiple-Choice. Dissertation. Humboldt-Universität zu Berlin, Berlin. Mathematisch-Naturwissenschaftliche Fakultät I. Online verfügbar unter https://edoc.hu-berlin.de/bitstream/handle/18452/17303/terzer.pdf

Upmeier zu Belzen, Annette; Krüger, Dirk (2010): Modellkompetenz im Biologieunterricht. In: *Zeitschrift der Didaktik der Naturwissenschaften* 16, S. 41–57.

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 der naturwissenschaftlichen Erkenntnisgewinnung. Berlin, Logos.