Volker Brüggemann Freie Universität Berlin
Volkhard Nordmeier

Einsatz adaptiver Testformate und Umsetzung im Projekt ValiDiS

Theoretische Rahmung

Mit dem Ko-WADiS-Test liegt ein Instrument vor, dass für die Kompetenzmessung im Bereich naturwissenschaftlichen Denkens entwickelt und erprobt wurde (Hartmann et al., 2015; Straube, 2016). Diese Kompetenz wird definiert nach Modellen von Mayer (2007) sowie Upmeier zu Belzen und Krüger (2010) und ist im Bereich der Erkenntnisgewinnung zu verorten.

Im Testinstrument wird sie anhand von sieben verschiedenen Facetten operationalisiert, die jeweils eine typische Handlung in naturwissenschaftlichen Untersuchungen darstellen (Fragestellungen formulieren, Hypothesen bilden, Untersuchungen planen und durchführen, Daten auswerten, den Zweck von Modellen erkennen, Modelle testen und Modelle abändern). Die Messung findet anhand von dichotomen Multiple-Choice Aufgaben statt. Bei den bisherigen Erhebungen wurde ein Testheft eingesetzt, das für jede der Facetten jeweils eine Aufgabe mit Kontexten aus den Fächern Biologie, Chemie und Physik, also insgesamt 21 Items beinhaltet. Für die im Projekt ValiDiS andauernden Längsschnittstudien werden mehrere solcher Testhefte in einem Multimatrix-Design verwendet. Der Itempool des gesamten Instruments umfasst 63 Aufgaben, die auf die Fächer und Facetten gleichverteilt sind.

Die bisher im Projekt erfassten Daten (N > 10.000) weisen auf eine valide Auslegung der Testdaten zur Kompetenzerfassung hin. Problematisch ist hierbei noch die Testgenauigkeit, die je nach Testheft schwankt und im Gesamtdatensatz bei einer EAP/PV-Reliabilität von 0.55 liegt. Betrachtet man andere Testinstrumente, die ähnliche Kompetenzkonstrukte durch rein papierbasierte Messung erfassen wollen, zeigen sich vergleichbare Reliabilitäten (vgl. Wellnitz, 2012; Woitkowski, 2015). Dennoch ist die Messgenauigkeit als gering zu bezeichnen. Für die projektinternen Forschungsvorhaben (in den Projekten Ko-WADiS und ValiDiS; vgl. auch Straube, 2016) wurden bzw. werden daher auch nur Gruppenwerte erhoben, die verlässliche Diagnose von Einzelpersonen ist nicht möglich. 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 Neukonzipierung der Aufgaben notwendig ist, bietet die Erstellung einer adaptiven Testversion.

Adaptive Tests passen sich in ihrer Schwierigkeit an die Proband*innen an. 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 und die gewonnen Information pro Aufgabe zu maximieren (SARI et al., 2016; Frey, 2012).

Durch mehrfache Wiederholung dieses Vorgangs kann der Test die Schätzung und Item-Auswahl verfeinern und somit auf einzelne Proband*innen adaptiv reagieren. Vergleichende Studien zeigen, dass adaptive Testverfahren gegenüber linearen Instrumenten (klassische Papiertests, auch FIT für *Fixed-Item-Test* genannt) die zeitliche Testökonomie deutlich erhöhen können (vgl. Weiss, 1982). Im Projekt ValiDiS wurde daher ein adaptiver Multistage-Test auf Grundlage des bestehenden Itempools entwickelt. Die Konzeption und

die Simulationsstudien zur Findung eines optimalen Testalgorithmus wurden 2018 durchgeführt und abgeschlossen (Brüggemann & Nordmeier, 2019).

Pilotierungsstudie – Vorgehen

Um das neue Testformat zu evaluieren, wurde im ersten Quartal 2019 eine Pilotierungsstudie durchgeführt. Die Zielgruppe waren Lehramtsstudierenden der drei naturwissenschaftlichen Fächer. Aufgrund der kleinen Studierendenzahlen in diesen Studiengängen sowie der ‚Vorbelastung' durch vorherige Befragungen mit der klassischen Version des Instruments konnte aber zunächst keine ausreichend große Stichprobe für die Pilotierung gewonnen werden. Da sich auch Studierende des Sachunterrichts im Grundschullehramt mit naturwissenschaftlichen Inhalten befassen, sofern sie Naturwissenschaften als Studienschwerpunkt gewählt haben, kamen sie ebenfalls für den Testeinsatz in Frage und wurden auch schon in früheren Studien untersucht (Straube, 2016). Es konnte hier eine Stichprobe von N = 283 Studierenden gewonnen werden, die zuvor noch nicht mit den Items des Instruments konfrontiert worden waren.

Die Pilotierung fand in Gruppen á maximal 30 Personen und unter Aufsicht von mit dem Instrument vertrauten Personen statt. Ein Zeitlimit wurde nicht gegeben. Die Teilnehmer*innen wurden vor der Befragung explizit darauf aufmerksam gemacht, dass es sich um ein adaptives Testinstrument handelte. Dieses Vorgehen wurde gewählt, da

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

Pilotierungsstudie – Ergebnisse

Im ersten Schritt der Auswertung wurden die Bearbeitungszeiten der Items analysiert. Es zeigte sich bei fünf der 283 Datensätze ein eindeutiges Rateverhalten (Aufgaben wurden systematisch und signifikant schneller bearbeitet als vom Gruppenmittel), weshalb sie für alle weiteren Auswertungen ausgeschlossen wurden.

Die mittlere Bearbeitungszeit der Befragung lag bei 22 Minuten (mit einer Standardabweichung von 6 Minuten; Abb. 1). Erfahrungswerten nach wird die Zeit für das papierbasierte Instrument bei 35 bis 45 Minuten angelegt. Hierbei ist anzumerken, dass diese Werte anekdotisch und damit vermutlich gruppenbezogen sind, also den Zeitpunkt darstellen, zu dem jeweils die meisten Proband*innen fertig waren. Für den Vergleich wird daher die mittlere anekdotische Zeitangabe (40 min) vergleichen mit der Zeitmarke im adaptiven Test, zu dem die Mehrheit der Stichprobe fertig war: 28 min, eine Standardabweichung später als der Mittelwert. Das stellt eine Verkürzung der Bearbeitungszeit um 30% dar und deckt sich mit der Reduzierung der pro Person bearbeiteten Aufgaben von 21 im Papierformat zu 15 im adaptiven Test.

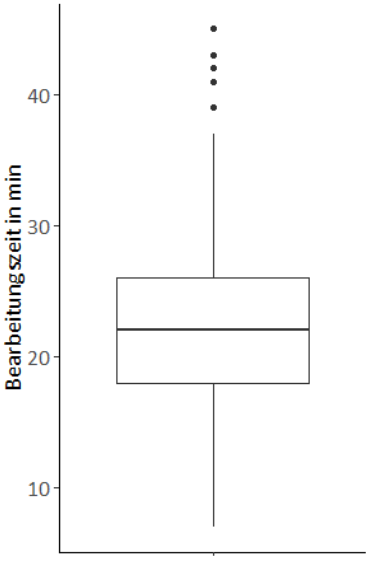

Abb. 1: Bearbeitungszeiten des adaptiven Tests (Pilotierung)

Die erreichte EAP-Reliabilität der Messung betrug 0.62. Sie lag damit über der des Papierinstruments und geringfügig unter der Prognose der zuvor durchgeführten Simulationen[1].

	Papierversion, Längsschnittdaten	optimierte Papierversion, simuliert	adaptive Version, Simulation	adaptive Version, Pilotierung
Testlänge (Anz. Items)	21	21	15	15
EAP-Reliabilität	0.55	0.6	0.63	0.62

Tabelle 1: Übersicht der Testlängen und Messgenauigkeiten (simuliert und real)

Betrachtet man zusammenfassend die Unterschiede zwischen den Pilotierungsdaten zur adaptiven Version und den vorliegenden Daten des Papierinstruments (Tabelle 1), zeigt sich eine Reduzierung der Testlänge um ~30% sowie eine Erhöhung der Messgenauigkeit um ~13%. Der Informationsgewinn pro Item (oder auch die zeitliche Ökonomie des Instruments) konnte um 60% gesteigert werden.
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 hier von einer Reduzierung der Messgenauigkeit mit schwachem Effekt ausgegangen.

Ausblick
Den Ergebnissen der Pilotstudie folgend wird die Umsetzung ins adaptive Testformat zunächst als erfolgreich eingestuft. Daher soll es nach Abschluss des Projekts ValiDiS zusammen mit der papierbasierten Variante bis Mitte 2020 veröffentlicht werden. Der Ort der Veröffentlichung ist aktuell noch offen, zur Diskussion steht u. a. die Plattform tet.folio, auf der der adaptive Test zurzeit für Entwicklungszwecke eingesetzt wird.
Ebenso muss noch über die endgültige Länge des Testinstruments entschieden werden. Im Sinne einer höheren Messgenauigkeit ist zu prüfen, ob die adaptive Testversion auf die ursprüngliche Testlänge erweitert werden sollte. Unabhängig von der projektinternen Entscheidung wird durch die Veröffentlichung von Testalgorithmus, Itempool, Datenbank und Simulationsskripten aber in Zukunft die Möglichkeit bestehen, das Instrument in seiner Länge zu variieren. Je nach Kontext kann so zwischen Belastung der Proband*innen und notwendiger Messgenauigkeit abgewogen werden.

[1] In einer früheren Veröffentlichung (Brüggemann & Nordmeier, 2019) wurde eine höhere Prognose der Messgenauigkeit aus Simulationen angegeben. Die Änderungen sind dadurch begründet, dass die Simulationen zwischenzeitlich wiederholt wurden. Die neuen Prognosen basieren auf einer stärker eingeschränkten Datenbasis, die der Zielgruppe besser entsprechen sollte. Durch die Neuberechnung ergab sich die konservativere Einschätzung.

Literatur

Brüggemann, Volker; Nordmeier, Volkhard (2019): Adaptive Leistungsmessung naturwissenschaftlichen Denkens. In: Christian Maurer (Hg.): Naturwissenschaftliche Bildung als Grundlage für berufliche und gesellschaftliche Teilhabe. GDCP Jahrestagung in Kiel 2018. Jahrestagung der Gesellschaft für Didaktik der Chemie und Physik. Kiel, 17.-20.09.2018. Universität Regensburg (39), S. 404–407.

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, zuletzt geprüft am 10.10.2019.

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, zuletzt geprüft am 10.10.2019.

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. Dissertation. Freie Universität Berlin, Berlin. Fachbereich Physik, zuletzt geprüft am 14.10.2019.

Upmeier zu Belzen, Annette; Krüger, Dirk (2010): Modellkompetenz im Biologieunterricht. In: *Zeitschrift der Didaktik der Naturwissenschaften* 16, S. 41–57, zuletzt geprüft am 10.10.2019.

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. Berlin: Logos Verlag (Biologie lernen und lehren, Band 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).