Naturwissenschaftliches Denken im Lehramtsstudium
– computeradaptive 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 Durchführung von Leistungstests zur Kompetenzmessung hat häufig das Problem eines sehr hohen Zeitaufwands. Als Folge werden Kompetenzen zwar in Forschungsprojekten umfänglich erhoben und untersucht, jedoch kaum zur Evaluation und inhaltlichen Sicherung der Lehrqualität an Hochschulen herangezogen. Computeradaptive Messverfahren ermöglichen eine weitaus effizientere (das bedeutet bei gleicher Messgenauigkeit zeitlich kürzere) Messung und somit einen praktikablen Einsatz von Leistungstests in Evaluationsszenarien.
Im ersten Teil dieses Beitrags werden das Projekt ValiDiS sowie das dort eingesetzte Instrument zur Kompetenzmessung im Bereich *naturwissenschaftlichen Denkens* vorgestellt. Im Anschluss daran werden adaptive Testverfahren vorgestellt. Im dritten und vierten Teil werden die bereits erfolgte Umsetzung eines computeradaptiven Multistage-Tests sowie die geplante Testerprobung skizziert.

1. Einleitung

Im Projekt ValiDiS (Kompetenzmodellierung und -erfassung: Validierungsstudie zum wissenschaftlichen Denken im naturwissenschaftlichen Studium) wird die Kompetenzentwicklung im Bereich *naturwissenschaftlichen Denkens* im Lehramtsstudium untersucht. Es schließt damit an das Projekt KoWADiS (Hartmann et al., 2015a) an. Im Zuge beider Projekte wird naturwissenschaftliches Denken als eine Kompetenz aufgefasst, die sich im Bereich der Erkenntnisgewinnung verorten lässt. Zu beobachten ist diese Kompetenz demnach in wissenschaftlichen Untersuchungs- und Modellierungsprozessen (Straube, 2016). *Naturwissenschaftliches Denken* wird im verwendeten Kompetenzmodell in sieben Handlungsfacetten eingeteilt (Tab. 1), die aus der Kombination von bestehenden Modellen zu Erkenntnisgewinnungsprozessen (Mayer, 2007) und Modellierungsprozessen (Upmeier zu Belzen & Krüger, 2010) gewonnen wurden.

Forschungsfragen formulieren	Hypothesen generieren	Untersuchungen planen	Daten auswerten und interpretieren
Zweck von Modellen bestimmen	Testen von Modellen	Ändern von Modellen	

Tabelle 1: Handlungsfacetten naturwissenschaftlichen Denkens

Für die Messung dieser Kompetenz wurde eigens ein Testinstrument entwickelt (Straube, 2016). Dabei handelt es sich um einen papierbasierten Leistungstest mit Multiple-Choice-Aufgabenformat. Der Aufgabenpool umfasst 63 Items, von denen in verschiedenen Fragebogenversionen stets 21 bearbeitet werden. Das Instrument befindet sich derzeit in der Validierungsphase.

1.1 Aktuelle Validierungsstudien und Ergebnisse

Um die Validität der Testwertinterpretation zu sichern, werden im Projekt ValiDiS eine Reihe verschiedener Evidenzquellen genutzt. Dazu gehören unter anderem Längsschnittstudien, Interventionen, Known-Groups-Vergleiche sowie Untersuchungen zur Prognose in Bezug auf reale Problemsituationen. Im Folgenden werden lediglich die für diesen Beitrag relevanten Studien erwähnt.

Im Längsschnitt wird das Instrument über den Verlauf des Bachelor- und Masterstudiums eingesetzt. Dort wird eine EAP-Reliabilität von .544 (Hartmann, zu Belzen, Krüger & Pant, 2015b) erreicht[1]. Es handelt sich derzeit nur um vorläufige Ergebnisse; die Beobachtung im Längsschnitt ist noch nicht in allen Kohorten abgeschlossen. Endgültige Testdaten sowie die inhaltliche Auswertung der Längsschnittstudie stehen daher noch aus. Im Vergleich unterschiedlicher Kohorten im Querschnitt zeigt sich

[1] Dieser Wert ist, wenn man sich nach einschlägigen Faustregeln für die Güte von Testinstrumenten richtet, eher als niedrig einzuordnen. Betrachtet man jedoch andere Leistungstests im Kompetenzbereich Erkenntnisgewinnung, so entspricht er den erreichten Genauigkeiten ähnlicher Instrumente (vgl. z.B. Terzer, 2012 Wellnitz, 2012; Woitkowski, 2015).

aber bereits ein erwartungskonformes Bild der Leistungsverteilungen, so zum Beispiel ein Ansteigen der Leistung im fortschreitenden Studium (Straube, 2016; Hartmann et al., 2015b).

Auch in den noch andauernden Interventionsstudien ließen sich bisher theoretische Erwartungen bestätigen. Die vorläufigen Ergebnisse lassen zudem auf eine ausreichende Sensitivität des Instruments hoffen, um Kompetenzverläufe im Rahmen von einzelnen Lehrveranstaltungen aufzulösen. Hiermit wird der zukünftige Einsatz des Tests in der Lehrevaluation ermöglicht.

Im Rahmen einer kompetenzorientierten Ausbildung an Hochschulen ist es wünschenswert, die eingesetzten Lehrveranstaltungen im Hinblick auf die erreichte Kompetenzförderung zu untersuchen (Wissenschaftsrat, 2008) – nicht nur in Bezug zur strukturellen Güte oder zur Eignung des jeweiligen Lehrpersonals. Leider besteht im Bereich der Kompetenzmessung nach wie vor ein Mangel an einsetzbaren und erprobten Instrumenten (Zlatkin-Troitschanskaia et al., 2015). Das im Projekt ValiDiS nun vorliegende Instrument bietet die Gelegenheit, genau dies in den Naturwissenschaften fachübergreifend zu tun.

1.2 Herausforderungen

Für die Implementierung des Testinstruments in Evaluationsszenarien besteht derzeit eine praktische Hürde, die es zu bewältigen gilt: Der Zeitaufwand einer Messung liegt je nach Proband*innengruppe bei mind. 35 Minuten. Soll eine Veranstaltung in Bezug auf die erreichte Kompetenzförderung evaluiert werden, so müsste der Test in einem Prä-Post-Design mindestens zweimal durchgeführt werden. Werden noch die bei solchen Erhebungen nötigen Zeiträume für die Informierung und Fragen zum ersten Zeitpunkt eingerechnet, so bewegt sich der Zeitaufwand im Bereich von 90 Minuten. Eine freiwillige Teilnahme der Studierenden (zu Kosten ihrer Freizeit) erscheint in diesem Rahmen unwahrscheinlich. Es ist also wünschenswert, den Zeitaufwand pro Messung zu verkürzen, um den Einsatz seitens der Hochschullehrenden ökonomischer und die Teilnahme wahrscheinlicher zu gestalten.

Hierzu kann der Test aber nicht einfach verkürzt und als Kurzversion mit weniger Aufgaben verwendet werden. Wie zuvor erwähnt, fällt die bisher erreichte Messgenauigkeit dem Inhaltsbereich entsprechend aus, sollte aber in keinem Fall weiter gemindert werden. Um eine weiterhin akzeptable Genauigkeit der Messung zu gewährleisten, muss also die Testeffizienz gesteigert werden.

2. Adaptive und klassische Testverfahren

Eine Möglichkeit, um Messinstrumente effizient zu gestalten, ist die Nutzung verschiedener Testverfahren. Eine dieser Optionen sind adaptive, insbesondere computeradaptive Testverfahren (CAT). Inwieweit diese eine Steigerung der Testeffizienz ermöglichen, soll im Folgenden skizziert werden.

2.1 Lineare Testverfahren

In klassischen Testverfahren, zum Beispiel bei papierbasierten Fragebögen, wird allen Befragten eine feste Anzahl von Aufgaben in einer festen Reihenfolge präsentiert. Die Menge, Reihenfolge und auch die Auswahl der Aufgaben werden vor der Befragung festgelegt und sind in den meisten Fällen auch für alle Befragten gleich. Werden die so gewonnenen Daten mittels der Item-Response-Theory ausgewertet, so tritt folgender Effekt auf: Im Aufgabenpool zeigt sich eine Spanne von unterschiedlichen Aufgabenschwierigkeiten. Je nach Art und Konstruktion der Fragen kann diese sehr breit sein – ob gewünscht oder nicht. Ebenso verhält es sich mit der Verteilung der Testleistungen (und damit der latenten Fähigkeiten) aller Teilnehmer*innen. Einzelne Aufgaben geben aber nur ein hohes Maß an Information über die einzelnen Personen, wenn Aufgabenschwierigkeit und die Ausprägung der latenten Fähigkeit einander entsprechen. Um eine Befragung möglichst effizient zu gestalten, müssten dementsprechend allen Teilnehmer*innen die geeigneten Aufgaben zugewiesen werden. Im Umkehrschluss folgt daraus, dass in vielen klassischen Testformaten ausnahmslos alle Proband*innen eine ganze Reihe von Aufgaben lösen, die für sie wenig geeignet sind und nur wenig Information liefern.

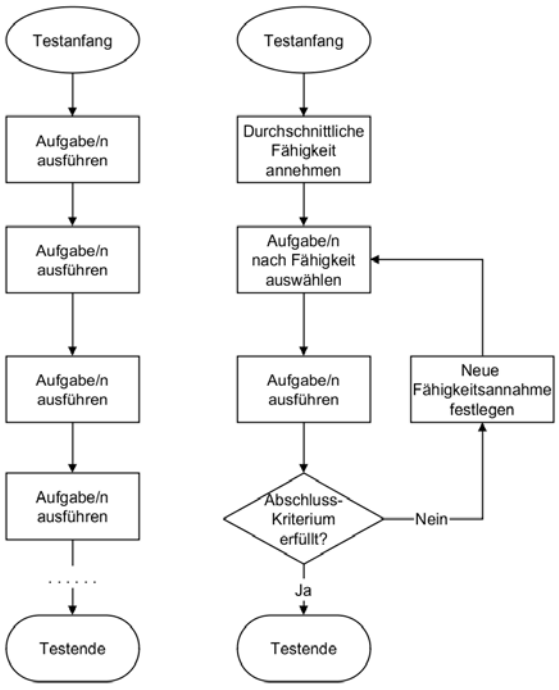

Abbildung 1: Beispielalgorithmen klassischer/linearer (links) und adaptiver (rechts) Testverfahren

2.2 Adaptive Testverfahren

Das Problem der Passung zwischen Proband*in und Aufgabe soll und kann durch computeradaptive Tests (CAT) umgangen werden (SARI et al., 2016). Hier wird, nachdem von einem/r Proband*in eine Aufgabe bearbeitet wurde, die latente Fähigkeit dieser Person geschätzt. Meist geschieht dies mittels ein- bis dreiparametrischer Modelle der Item-Response-Theory und auf Grundlage von zuvor normierten Aufgabenpools. 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 (Abbildung 1). Sobald ein vorher definiertes Abschlusskriterium erfüllt wird, ist der Test 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 (Weiss, 1982).

2.3 Multistage-Tests

Wie häufig adaptive Tests Fähigkeitsschätzungen durchführen, ist von Instrument zu Instrument unterschiedlich. ‚Echte' adaptive Tests, so wie sie ursprünglich entwickelt wurden, führen nach jeder einzelnen Aufgabe eine Berechnung durch. Daneben gibt es aber auch Multistage-Tests oder kurz MSTs (Hendrickson, 2007).

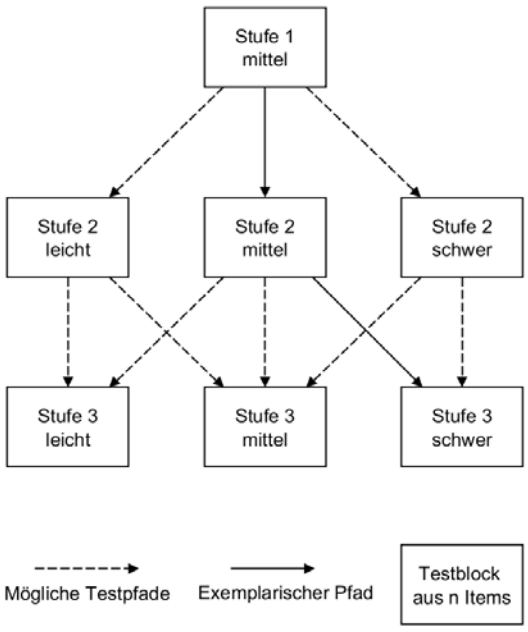

Abbildung 2: Multistage-Tests im 1-3-3 Design mit 3 Stufen und 3 Schwierigkeitsbereichen

Solche Tests bestehen aus einer Reihe von Aufgabenblöcken. Jeder Block besteht aus mehreren Aufgaben in einem bestimmten Schwierigkeitsbereich. Für jeden Bereich gibt es mindestens einen Aufgabenblock, sodass die Blöcke als verschieden schwere Versionen des gesamten Tests agieren und zusammen alle möglichen Fähigkeitsbereiche abdecken. Die Fähigkeitsschätzung wird in solchen Tests nach dem Absolvieren eines einzelnen Blocks durchgeführt, um danach zum nächsten Block weiterzuleiten.

Das Abschlusskriterium wird auch hier nicht zwingend immer gleich gewählt, richtet sich meist aber nach einer vorgesehenen Anzahl von absolvierten Aufgabenblöcken.

Die Gesamtstruktur, also die Menge an Blöcken und die möglichen Pfade zwischen diesen, ist nicht per se festgelegt. Simulations- und Vergleichsstudien verweisen aber auf das sogenannte 1-3-3-Design (Abbildung 2) als eine bewährte Möglichkeit (Zheng & Chang, 2014).

2.4 Vergleich der Testverfahren

Die Entscheidung zwischen linearen Verfahren, CATs und MSTs ergibt sich je nach Anforderung. Im Grunde vereinen MSTs die Vorteile beider anderer Verfahren (SARI et al., 2016): Sie ermöglichen wie lineare Tests die gleichzeitige Messung mehrerer Inhaltsbereiche, erlauben das Zurückspringen zu und Überarbeiten von vorherigen Antworten. Was Gesamtlänge und Messgenauigkeit betrifft, so sind sie eher im Bereich der CATs zu verordnen. Somit sind sie deutlich effizienter als lineare Verfahren.

Der wohl größte Nachteil computeradaptiver Verfahren gilt jedoch auch für MSTs: Es ist notwendig, eine umfassende Itemdatenbank zu generieren, die normiert und IRT-konform ist. Zudem muss je nach Komplexität des Tests und Algorithmus die entsprechend notwendige Infrastruktur aufgebaut werden. Die hierfür notwendigen Server stellen verglichen zu linearen Verfahren einen nicht unerheblichen Aufwand dar. Neben den noch eher geringen Materialkosten ist besonders die Einrichtung und Wartung durch Fachpersonal kostspielig. Auch hier liegen MSTs im Bereich zwischen den anderen Formaten: Es sind weniger Berechnungen nötig als bei CATs, womit auch die entsprechende Rechenleistung geringer ausfällt.

3. Umsetzung eines computeradaptiven Tests

Da für das Vorhaben im Projekt ValiDiS die Effizienzsteigerung als dringend notwendig angesehen wird, überwiegen hier die Vorteile einer adaptiven Version die entsprechenden Nachteile.

Es wurde daher beschlossen, eine solche Anpassung des Instruments umzusetzen.

Die grundlegende Entwicklung wurde im Februar 2018 abgeschlossen und erprobt. Die dafür notwendigen Arbeitsschritte und Designentscheidungen sollen im Folgenden dargelegt werden. Besonders anzumerken ist, dass die Problematik der notwendigen Infrastruktur und technischen Wartung umgangen werden konnte. Es war daher möglich, das neue Testformat in Form einer ‚Website' zu realisieren (mehr dazu unter Punkt 3.6).

3.1 Wahl des Testformats

Die Entscheidung des Testformats wurde nach zwei praktischen Gesichtspunkten zu Gunsten eines MSTs getroffen. Zum einen ermöglicht dieser die Bearbeitung mehrerer inhaltlicher Facetten/ Fachbereiche aus dem vorliegenden Itempool. Da das Konstrukt *Naturwissenschaftliches Denken* eindimensional modelliert wird (Straube, 2016), wäre dies nicht zwingend nötig gewesen. Es wurde trotzdem entschieden, dass eine Fähigkeitsschätzung aufgrund mehrerer inhaltlicher Facetten valider erscheint.

Zum anderen war zum Zeitpunkt der Testumsetzung keine ausreichende technische Infrastruktur vorhanden. Zum Zeitpunkt der Entscheidung wurde hier noch von einem nicht unerheblichen Aufwand in Ressourcen und Arbeitszeit ausgegangen. Um diesen zumindest einzugrenzen, erschien ein MST sinnvoller.

3.2 Wahl des Abschlusskriteriums

Es wurden in Bezug auf den Zweck der Neugestaltung zwei mögliche Abschlusskriterien diskutiert: Der Abbruch nach einer definierten Testlänge sorgt dafür, dass bei der Messung von ganzen Gruppen die Zeit möglichst ökonomisch genutzt wird – die Proband*innen benötigen bei gleicher Aufgabenzahl erfahrungsgemäß etwa gleich viel Zeit. Bricht der Test dagegen erst nach einer gewissen Genauigkeit der Schätzung ab, so kann dies vor allem bei Personen mit extremen Ausprägungen erheblich länger dauern. In den Randbereichen der Fähigkeitsverteilung wird weniger Information pro Aufgabe gewonnen. Im vorliegenden Fall würde die resultierende Verlängerung der Befragung noch verstärkt, da ganze Aufgabenblöcke nachgereicht würden. Dafür könnte so eine sehr hohe Genauigkeit für die Schätzung jeder einzelnen Person gesichert werden. Im Endeffekt wurde, wie meist bei MSTs, die Testlänge als Kriterium ausgesucht. Der Grund liegt in dem Fokus auf Gruppenmessung und -vergleich im Gegensatz zur Einzeldiagnostik.

3.3 Erstellung und Normierung des Aufgabenpools

In den Projekten Ko-WADiS und ValiDiS wurden von Anfang an sämtliche Daten mittels IRT ausgewertet. Dementsprechend lag bereits ein vollständig IRT-konformer Aufgabenpool vor. Durch die in Längsschnittbefragungen aggregierten Daten ist zudem bereits eine sichere Normierung der Aufgaben möglich. Die bestehenden Daten werden von den getesteten Modellen (ein- bis dreiparametrisch) am besten durch ein 2PL-Modell beschrieben. Das Konstrukt wird dabei eindimensional betrachtet (s. o.). Da keine signifikanten Hintergrundvariablen identifiziert werden konnten, wird kein Hintergrundmodell verwendet.

3.4 Festlegung der Teststruktur

Die Testlänge richtet sich nach der angestrebten Messgenauigkeit. Adaptive Verfahren erreichen schon bei halber Testlänge oft die gleiche Messgenauigkeit wie die entsprechenden linearen Testversionen (Frey & Ehmke, 2008). Da MSTs bei der Effizienz im Vergleich zu CATs leichte Einbußen verzeichnen, wurde als Ausgangspunkt eine konservativere Entscheidung getroffen und die Länge um etwa ein Viertel reduziert: Von bisher 21 auf 15 Aufgaben. (Je nach Ergebnissen der Pilotierung wird die Testlänge später angepasst.)

Die Pfadstruktur des Tests wurde genau wie die Testlänge anhand einer Recherche festgelegt. Vergleichende Studien kommen zu dem Ergebnis, dass drei Schwierigkeitsbereiche mit drei Teststufen (1-3-3 Format) für eine erfolgreiche Adaption ausreichend sind (Hendrickson, 2007). Daher wurde diese Struktur ausgewählt.

3.5 Konstruktion der Aufgabenblöcke

Die Blockgröße wurde den vorigen beiden Punkten entsprechend auf fünf Aufgaben festgelegt. Die Aufgaben wurden den sieben benötigten Blöcken wie folgt zugewiesen:

Die Schätzung der Aufgabenschwierigkeiten wurde anhand der bisher gesammelten Daten vorgenommen. Für die in den kommenden Befragungen angepeilte Zielgruppe (Physikstudierende) wurde danach aus den bisherigen Daten eine Vorhersage für die Fähigkeitsverteilung durchgeführt. Die Fähigkeitsausprägungen wurden in Terzile eingeteilt. Anhand dieser Grenzen wurde die Zuordnung der Aufgaben in einen leichten, mittleren und schweren Bereich entschieden (Abbildung 3).

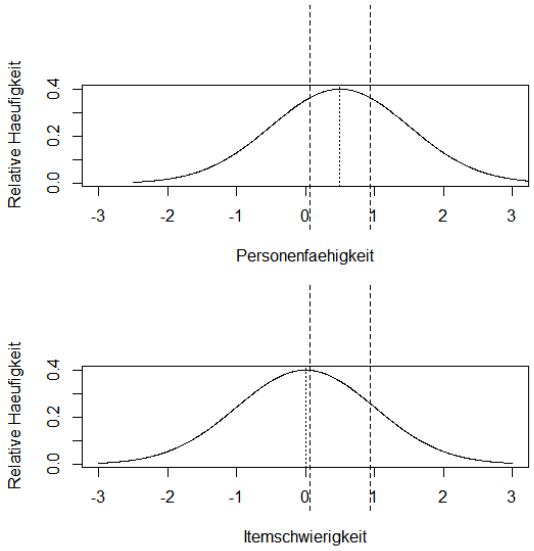

Abbildung 3: Festlegung der Schwierigkeitsbereiche

Im nächsten Schritt wurden die einzelnen Blöcke aus den zur Verfügung stehenden Aufgaben so konstruiert, dass a) die Aufgabenschwierigkeiten innerhalb jeden Blockes möglichst gleichverteilt sind, b) in jedem Block alle Inhaltsbereiche (Physik, Chemie, Biologie) mindestens einmal vertreten sind und c) verschiedene Handlungsfacetten (Tabelle 1) angesprochen werden.

3.6 Implementierung des Testalgorithmus

Ein besonderes Merkmal dieser Testumsetzung ist, dass die technischen Anforderungen stark vermindert werden konnten. Normalerweise würden zwischen den Stufen des Tests in Echtzeit Berechnungen durchgeführt, um die Fähigkeit zu schätzen. Hierzu ist es notwendig, dem mathematischen Modell entsprechend Serverleistung zur Verfügung zu stellen. In diesem Fall jedoch ist die Anzahl der möglichen Testpfade überschaubar: Nach der Durchführung des zweiten Blocks, also zur zweiten und letzten Weiterleitung innerhalb des Tests, wurden zehn Fragen beantwortet. Es ergeben sich damit 1024 mögliche Kombinationen aus richtigen oder falschen Antworten. Die resultierenden Fähigkeitsschätzungen wurden nun für alle Kombinationen im Voraus berechnet und gespeichert. Während der tatsächlichen Messung müssen sie daher nur noch ausgelesen und nicht neu errechnet werden. Somit war es möglich, das adaptive Instrument als Webumgebung zu erstellen. Dies erfolgte mit Unterstützung der Applikation tetfolio (https://tetfolio.fu-berlin.de/).

3.7 Technische Erprobung

Im Februar 2018 wurde eine erste technische Erprobung des neuen Testformats durchgeführt. Ziel war die Sicherstellung eines technisch fehlerfreien Ablaufs der Befragung. Es nahmen 32 Physik-Studierende im Mono- und Lehramtsstudium des Fachbereichs Physik am Standort FU Berlin teil. Eine statistische Auswertung der Daten war nicht Ziel dieser Erhebung (da die Schätzung eines IRT-Modells mit 32 Datensätzen nicht sinnvoll ist). Die durchschnittliche Bearbeitungszeit lag bei 25 Minuten.

4. Ausblick

Nach der ersten technischen Umsetzung ist nun zu prüfen, ob das gesetzte Ziel im Projekt erreicht werden kann. Dafür soll das adaptive Testinstrument pilotiert werden. Geplant ist eine einmalige Erhebung von Lehramtsstudierenden der naturwissenschaftlichen Fächer. Die benötigte Stichprobe wird auf n=350 geschätzt. Bei einer etwa erwarteten Gleichverteilung auf die drei Schwierigkeitsebenen sollen damit mindestens 100 Bearbeitungen für jedes Item garantiert werden. Diese Grenze wird angesetzt, um eine sichere Schätzung aller Item- und Personenparameter durchzuführen. Die daraus gewonnenen Informationen zur Reliabilität des eigenständig eingesetzten Tests werden als Kriterium verwendet.

Noch ist unklar, ob es sich bei der beschriebenen Version des Tests um die effizienteste Möglichkeit handelt. Wie unter 3.4 und 3.5 dargelegt, wurden Teststruktur und Aufgabenblöcke nach dem vorliegenden Stand internationaler Literatur zusammengestellt. Es bleiben dennoch mehrere offene Punkte:

a) Obwohl es in der Literatur am häufigsten zu finden ist, müssen die Blöcke eines MSTs nicht alle gleich lang gewählt werden. Der erste Block kann auch zu Kosten der späteren verlängert werden. Somit ist die erste Einordnung der Personen genauer, kann aber weniger gut nachgesteuert werden. Hier muss je nach konkretem Fall entschieden werden, wie zu verfahren ist.

b) Die Testlänge wurde sehr vorsichtig gewählt. Sofern möglich, sollte sie noch weiter eingeschränkt werden.

c) Je nach Fähigkeitsverteilung der Zielgruppe kann es sein, dass Aufgaben an den Grenzen der festgelegten Schwierigkeitsbereiche anders eingeordnet werden müssten. Es ist zu prüfen, ob dadurch bei extremen Gruppen signifikante Einbußen der Messgenauigkeit auftreten.

Um die Fragen zu beantworten, sind Vergleiche der alternativen Testzusammenstellungen notwendig. Durch die Vielzahl der möglichen Kombinationen würde dies bei Vergleichsstudien einen erheblichen Aufwand darstellen. Aus diesem Grund sollen stattdessen **Simulationen** eingesetzt werden. Der im Projekt vorliegende Datensatz wird als ausreichend angesehen, um für die verschiedenen denkbaren Strukturen Stichproben zu simulieren und einen

Vergleich anzustellen. Dieses Vorgehen wird als weitaus ökonomischer und in Bezug auf die Belastung der sonst nötigen Proband*innen als ethischer angesehen.

Erste Ergebnisse werden im Herbst 2018 erwartet.

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

5. Literaturverzeichnis

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

Frey, Andreas; Ehmke, Timo (2008): Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. In: Manfred Prenzel, Ingrid Gogolin und Heinz-Hermann Krüger (Hg.): Kompetenzdiagnostik. Zeitschrift für Erziehungswissenschaft. Wiesbaden: VS Verlag für Sozialwissenschaften (Zeitschrift für Erziehungswissenschaft Sonderheft, 8), S. 169–186.

Hartmann, Stefan; Upmeier zu Belzen, Annette; Krüger, Dirk (2015a): Ko-WADiS Schlussbericht.

Hartmann, Stefan; Upmeier zu Belzen, Annette; Krüger, Dirk; Pant, Hans Anand (2015b): Scientific Reasoning in Higher Education. In: Zeitschrift für Psychologie 223 (1), S. 47–53. DOI: 10.1027/2151-2604/a000199.

Hendrickson, Amy (2007): An NCME Instructional Module on Multistage Testing. In: Educational Measurement: Issues and Practice 26 (2).

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

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).

Terzer, Eva (2012): Modellkompetenz im Kontext Biologieunterricht - Empirische Beschreibung von Modellkompetenz mithilfe von Multiple-Choice Items. Dissertation. Humboldt-Universität zu Berlin, Berlin. Mathematisch-Naturwissenschaftliche Fakultät I.

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.

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).

Wissenschaftsrat (2008): Empfehlungen zur Qualitätsverbesserung von Lehre und Studium. Berlin. Online verfügbar unter https://www.wissenschaftsrat.de/download/archiv/8639-08.pdf, zuletzt geprüft am 04.05.2018.

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).

Zlatkin-Troitschanskaia, Olga; Shavelson, Richard J.; Kuhn, Christiane (2015): The international state of research on measurement of competency in higher education. In: Studies in Higher Education 40 (3), S. 393–411. DOI: 10.1080/03075079.2015.1004241.