Testvorstellung: Computeradaptive Leistungsmessung im Bereich naturwissenschaftlichen Denkens

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

In den Projekten Ko-WADiS und ValiDiS wurde die Kompetenzentwicklung naturwissenschaftlichen Denkens bei Lehramtsstudierenden untersucht. Zu diesem Zweck wurde in der ersten Projektphase ein Messinstrument entwickelt und in der zweiten Phase in Bezug auf seinen Einsatz validiert. Da das Instrument zwar valide Messungen ermöglicht, aber 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 die theoretischen Grundlagen adaptiven Testens und das Vorgehen der Testkonstruktion beschrieben. Zusätzlich werden Methodik sowie die Ergebnisse von Simulationsstudien und der Pilotierungsstudie des adaptiven Formats zusammengefasst. Im Vergleich beider Versionen konnte eine deutliche Steigerung der Messeffizienz (höhere Messgenauigkeit bei kürzerer Testzeit) durch die adaptive Testanwendung nachgewiesen werden.

1. Der Ko-WADiS-Test

Bei dem Ko-WADiS-Test handelt es sich um ein Messinstrument zur Erfassung von Kompetenzen des naturwissenschaftlichen Denkens (Straube, 2016; Hartmann, Mathesius, Stiller, Straube, Krüger & Upmeier zu Belzen, 2015;). Er wurde in den Projekten Ko-WADiS und ValiDiS entwickelt, um diese Kompetenzen bei Lehramtsstudierenden im Verlauf des Studiums erheben bzw. beobachten zu können.

Als theoretische Fundierung des Ko-WADiS-Tests dient der Ansatz, naturwissenschaftliches Denken als eine Problemlösefähigkeit zu interpretieren (nach Arbeiten von Mayer (2007) zum Problemlösen in den Naturwissenschaften sowie Upmeier zu Belzen und Krüger (2010) zur Arbeit mit naturwissenschaftlichen Modellen). Naturwissenschaftliches Denken wird danach als eine Kompetenz betrachtet, die einen Teilbereich der Erkenntnisgewinnung ausmacht.

Im Ko-WADiS-Testinstrument wird die Kompetenz durch sieben Kompetenzfacetten (Straube, 2016) operationalisiert. Sie stellen unterschiedliche Arbeitsphasen 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.

Der Test besteht aus einem Pool von 63 dichotomen Multiple-Choice-Items. Jede Facette ist durch 9 Items repräsentiert, von denen wiederum je 3 einen Kontext aus der Biologie, Chemie und Physik haben. Bisher 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 Items, je eins pro Fach und Facette.

Die Auswertung des bisher gewonnenen 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 allerdings in Einklang mit weiteren rein schriftlichen Testinstrumenten für ähnliche Konstrukte (vgl. Neumann, 2011: 0.55; Terzer, 2013: 0.46; Wellnitz, 2012: 0.59).

2. Adaptive und lineare Testverfahren

Eine Möglichkeit, um Messinstrumente effizienter und messgenauer 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

Meist erfolgt die Messung persönlicher Merkmale und Fähigkeiten mit Tests in Papierform (Magis, Yan & Davier, 2017; Moosbrugger & Kelava, 2012). Diese werden aus einer festen Zusammenstellung von Items konstruiert und auch in dieser exakten Zusammenstellung an alle befragten Personen verteilt. Im englischsprachigen Raum finden sich hierfür unter anderem die Bezeichnungen als *Fixed-Item-Tests* (FIT) (Ling, Attali, Finn & Stone, 2017) oder *Linear Tests* (Magis et al., 2017).

Dass diese linearen Tests in ihrer Form starr sind, hat direkte Folgen für die Menge an benötigten Items. Für jeden zu messenden Merkmals- oder Fähigkeitsbereich müssen schließlich ausreichend Items im Test enthalten sein, um diesen ausreichend genau zu erfassen. Insbesondere bei der Erfassung von Kompetenzen, für die oft schon theoretisch mehrere Niveaustufen erwartet werden, sind die nötigen Messinstrumente entsprechend umfangreich.

Besonders wird dieser Umstand bei der Auswertung von Testdaten mittels der Item-Response-Theory (IRT) deutlich, die Itemschwierigkeiten und Personenfähigkeiten als zwei getrennte Merkmale modelliert. Im Itempool zeigt sich eine Spanne von unterschiedlichen Schwierigkeiten. Je nach Art und Konstruktion der Items 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 Items geben aber nur ein hohes Maß an Information über die bearbeitenden Personen, wenn Schwierigkeit und die Ausprägung der latenten Fähigkeit einander entsprechen. Um eine messgenaue Befragung zu ermöglichen, müssen dementsprechend allen Teilnehmer*innen wenigstens ein paar geeignete Items vorgelegt werden. Daraus folgt, dass in linearen Testformaten ausnahmslos alle Proband*innen eine ganze Reihe von Items bearbeiten, die für sie wenig geeignet sind und nur wenig Information liefern.

2.2 Adaptive Testverfahren

Das Problem der Passung zwischen Proband*in und Item soll und kann durch adaptive Testverfahren umgangen werden (Sari, Yahsi-Sari & Huggins-Manley, 2016).

Die Kernidee adaptiver Verfahren ist die individuelle Anpassung von Tests an das Verhalten verschiedener Personen (Frey, 2012). Es sollen dabei Items so ausgewählt werden, dass sie möglichst gut für die jeweiligen Proband*innen geeignet sind. Das bedeutet beispielsweise, keine unsinnig schweren Aufgaben an leistungsschwache Testsubjekte zu vergeben – das Ergebnis einer falschen Lösung ist dabei zu vorhersehbar und liefert keine sinnvollen, neuen Erkenntnisse.

Mit der Idee eines optimalen Tests für jede Person müssen solche Situationen also ausgeschlossen werden. Eine direkte Konsequenz davon ist die Notwendigkeit, allen Proband*innen unterschiedliche Items vorlegen zu können. Die Auswahl der geeigneten (beziehungsweise der Ausschluss der ungeeigneten) Items erfolgt als direkte Reaktion auf das Verhalten der Person im bisherigen Verlauf des Tests.

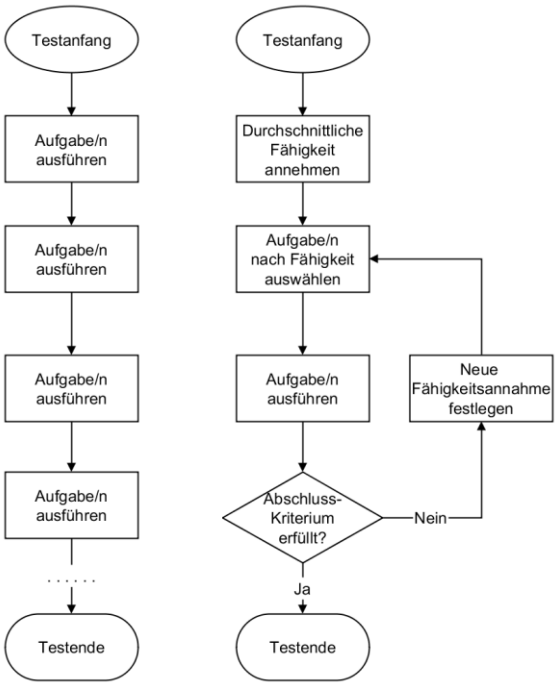

Abb. 1: Beispielalgorithmen klassischer/linearer (links) und adaptiver (rechts) Testverfahren

Nach der Bearbeitung des ersten Items wird durch einen vorgegebenen Algorithmus dasjenige Item aus dem Pool gesucht, das am besten für die weitere Messung der Testperson geeignet ist. Es wird aus dem Pool der übrigen Items entfernt und der Testperson vorgelegt. Dieser Prozess wird so häufig wiederholt, bis ein vorgegebenes Abschlusskriterium erfüllt wurde und die Messung beendet wird (vgl. Abb. 1).

Wie häufig adaptive Tests Fähigkeitsschätzungen durchführen, ist von Instrument zu Instrument unterschiedlich. ‚Echte' adaptive Tests, so wie sie ursprünglich entwickelt wurden, führen nach jedem einzelnen Item eine Berechnung durch. Sie werden praktisch immer computergestützt durchgeführt und als Computeradaptive Tests (CAT) bezeichnet. Daneben gibt es aber auch Multistage-Tests oder kurz MSTs (Hendrickson, 2007).

2.3 Multistage-Tests

Solche Tests bestehen aus einer Reihe von ‚Modulen'. Jedes Modul besteht aus mehreren Items in einem bestimmten Schwierigkeitsbereich und ist in sich selbst ein linearer Test. Für jeden Bereich gibt es mindestens ein Modul, sodass diese als verschieden schwere Elemente des gesamten Instruments agieren und zusammen alle möglichen Fähigkeitsbereiche abdecken. Die Fähigkeitsschätzung wird in

solchen Tests nach dem Absolvieren eines einzelnen Moduls durchgeführt, um danach zum nächsten weiterzuleiten.

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

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

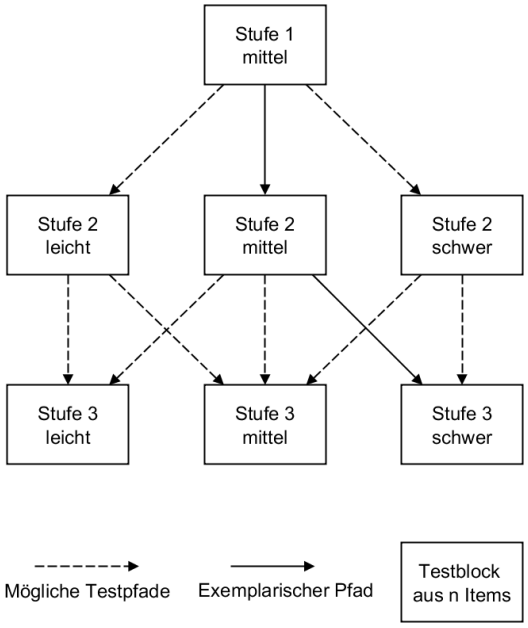

Abb. 2: Multistage-Tests im 1-3-3 Design mit 3 Stufen und 3 Schwierigkeitsbereichen

2.4 Vergleich der Testverfahren

Im direkten Vergleich der verschiedenen Verfahren zeigt sich ein eindeutiger Vorteil von beiden adaptiven Formaten gegenüber dem Einsatz von linearen Tests. CATs benötigen im Mittel nur halb so viele Items wie lineare Tests, um die gleiche Messgenauigkeit zu erreichen (Frey & Ehmke, 2008; Green, 2012). Auch gegenüber MSTs ist eine höhere Effizienz der CATs festzustellen. Dabei hängt es von der Struktur des konkreten MSTs ab, wie groß der Abstand zu den jeweils anderen Testformaten ausfällt. Patsula (1999) schließt aus ihren Daten, dass sich MSTs mit höherer Aufspaltung in Module nicht nur strukturell immer weiter einem CAT annähern, sondern gleichzeitig in ihrer Effizienz. Dabei waren auch die strukturell simpelsten MSTs ihrer Studie (1-3 Designs) noch um 10-40% effizienter als der entsprechende lineare Test, wenn auch signifikant ineffizienter als der CAT.

Der Nachteil von beiden computeradaptiven Verfahren gegenüber linearen besteht im Aufwand: Es ist notwendig, einen umfassenden Itempool zu generieren, der 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 einen nicht unerheblichen Aufwand dar. Neben den noch eher geringen Materialkosten sind besonders die Einrichtung und Wartung durch Fachpersonal kostspielig. Auch hier liegen MSTs im Bereich zwischen den anderen Formaten (Hendrickson, 2007): Es sind in der Messung weniger Berechnungen nötig als bei CATs, womit auch die entsprechende Rechenleistung geringer ausfällt. Dennoch ist die Implementierung komplizierter als bei linearen Tests, die schlicht ausgedruckt oder als interaktive html-Dokumente verlinkt werden können.

3. Simulationsstudie

Im Fall des Ko-WADiS-Tests wurde entschieden, die Effizienz des Instruments durch die Konstruktion eines MST zu verbessern. Hierfür gab es im Wesentlichen zwei praktische Gründe: Am Standort gab es zum einen keine bestehende Infrastruktur für den Einsatz adaptiver Formate. Diese einzurichten wurde als erheblicher Aufwand eingeschätzt und erschien im Falle eines im Vergleich zu CATs weniger Ressourcen kostenden MSTs realistischer. Zum anderen ermöglichte ein MST die Erstellung von Modulen, die jeweils verschiedene Fachgebiete und Facetten des Kompetenzkonstrukts abdecken. Aus mathematischer Sicht wäre das nicht notwendig, da die zu messene Kompetenz eindimensional modelliert wurde (Straube, 2016) und somit nicht mehrere Facetten zur Fähigkeitsschätzung notwendig wären. Es wurde aber dennoch als sinnvoll betrachtet, da die Einstufung von Studierenden auf der Basis von nur fachfremden Items (z. B. Biologiestudierende nur durch Physikitems, was in einem CAT auftreten kann) Zweifel an der Validität der Messung aufkommen ließe.

Für die Konstruktion von MSTs kommen allerdings verschiedene Strukturen in Betracht. Im Vorfeld der Testkonstruktion war nicht klar, in wie viele Schwierigkeitsbereiche und Module das Instrument gegliedert werden sollte. Bisherige Vergleiche zeigen keine eindeutig optimale Struktur für fixe Testlängen oder ähnliche allgemeingültige Kriterien (Armstrong, Jones, Koppel & Pashley, 2004; Patsula, 1999). Stattdessen scheint es notwendig zu sein, den Aufbau von MSTs in jedem individuellen Kontext einzeln zu entscheiden.

Um eine möglichst effiziente Struktur für den Ko-WADiS-MST zu finden, waren deshalb Vergleichsstudien zwischen den denkbaren Alternativen notwendig. Weil 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 Module erstellt. Anstelle von einer wirklichen Bearbeitung durch Proband*innen aus der Zielgruppe wurde 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 logistischen Modells imputiert.

3.2. Ergebnisse

Abbildung 3 zeigt die erreichten EAP/PV-Reliabilitäten verschiedener MST Versionen. Dabei wurden für jeden dieser Tests auch verschiedene Gesamtlängen geprüft.

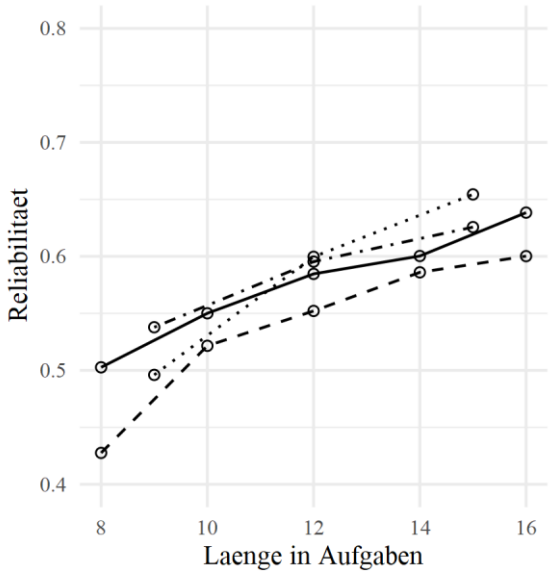

Abb. 3: Simulationsergebnisse; EAP/PV-Reliabilität verschiedener Multistage-Test-Strukturen für unterschiedliche Testlängen. Die Nummerierung steht für die Anzahl der differenzierten Schwierigkeitsbereiche in den aufeinanderfolgenden Stufen der Tests.

Um die Voraussagekraft der Ergebnisse einzuschätzen, wurden daneben auch 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 Befragungen und in den Simulationen gemessenen 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. Die Ergebnisse der Simulationen wurden insgesamt als eine sinnvolle Vorhersage für den Einsatz der verschiedenen Testversionen eingeschätzt.

Der MST im 1-2-2 Design mit 15 Items Gesamtlänge erreichte eine Messgenauigkeit von 0.66. Das stellt im Vergleich zur bisherigen Messgenauigkeit des linearen Tests im Gesamtdatensatz eine Verbesserung von 21% dar. Gleichzeitig wurde die Testlänge um 28% reduziert. Das Instrument wurde diesen Erwartungen gemäß umgesetzt.

4. Pilotierungsstudie

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

Die Zielgruppe für den Einsatz des Instruments sind eigentlich die Lehramtsstudierenden der drei naturwissenschaftlichen Fächer. Aufgrund kleiner Studierendenzahlen in diesen Studiengängen (insbesondere im Fach Physik) konnte aber keine ausreichend große Stichprobe für die Pilotierung gewonnen werden. Zudem gab es bei praktisch allen in Frage kommenden Proband*innen an den teilnehmenden Standorten das Problem, dass die Aufgaben des Instruments durch vorherige Befragungen in den Längsschnitterhebungen des Projekts bekannt waren.

Im Studiengang Sachunterricht im Grundschullehramt konnte allerdings eine sehr ähnliche Stichprobe gefunden werden. Die Studierenden arbeiten im Sachunterricht ebenfalls mit naturwissenschaftlichen Inhalten, sofern sie hier ihren Studienschwerpunkt setzen. Somit kommen sie auch für den Testeinsatz in Frage und wurden aus diesem Grund bereits in früheren Studien mit dem linearen Instrument untersucht (Straube, 2016). Durch das Ausweichen auf diese alternative Proband*innengruppe 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, da die neue Stichprobe noch vollständig ‚unbelastet' war.

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, Hartig & Moosbrugger, 2009).

4.1. Ergebnisse

Vor der Auswertung der Testdaten erfolgte eine Betrachtung der ebenfalls erhobenen Bearbeitungszeiten aller Items. Ziel war die Identifikation von solchen Items, die sich durch auffällig lange oder kurze Bearbeitung auszeichneten und von Personen, die systematisch signifikant schnellere Bearbeitungen 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 Items wurden nicht entdeckt.

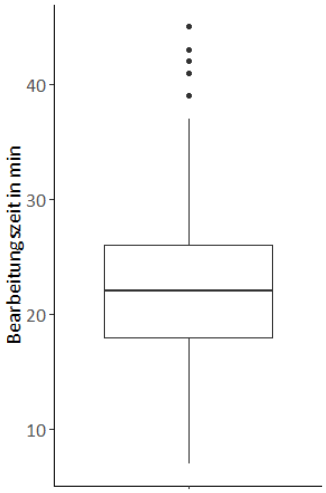

Abb. 4: Gemessene Gesamtbearbeitungszeiten des adaptiven Instruments in der Pilotierungsphase nach Datenbereinigung. Punkte stellen Ausreißer dar.

Die Bearbeitungszeit der Befragung lag im Mittel bei 22 Minuten (mit einer Standardabweichung von 6 Minuten, siehe Abb. 4). Für das papierbasierte Instrument werden Bearbeitungszeiten von 35 bis 45 Minuten angelegt. Diese Werte sich allerdings rein ‚anekdotisch' (Erfahrungsberichte der Testleiter*innen aus der Ko-WADiS-Phase, keine gezielten Beobachtungen) und damit möglicherweise verfälscht sowie gruppenbezogen. Sie entsprechen also dem Zeitpunkt, zu dem beim linearen Testeinsatz 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) mit der Zeitmarke im adaptiven Test abgeglichen, 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 Items von 21 im linearen 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.

Tabelle 1 gibt einen Überblick über die bisherige lineare Version sowie den simulierten und den tatsächlich pilotierten MST. Es zeigt sich eine Reduzierung der Testlänge um ~30% sowie eine gleichzeitige Erhöhung der Messgenauigkeit um ~13%. Der Informationsgewinn pro Item beziehungsweise die Effizienz des Instruments konnte also deutlich gesteigert werden.

	Testlänge	EAP/PV-Reliabilität
FIT, Ist-Stand	21	0.544
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.

4.2. Diskussion

Die Evaluation des neuen Instruments zeigt (erwartungskonform) positive Resultate (Reduzierung der Testlänge um ~30% mit gleichzeitiger Erhöhung der Messgenauigkeit um ~13%).

Die Ergebnisse der Pilotierung sind vermutlich aufgrund der Auswahl der Stichprobe aber 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.

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.

5. Literaturverzeichnis

Armstrong, R. D., Jones, D. H., Koppel, N. B. & Pashley, P. J. (2004). Computerized Adaptive Testing With Multiple-Form Structures. *Applied Psychological Measurement*, *28*(3), 147–164. https://doi.org/10.1177/0146621604263652

Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (Springer-Lehrbuch, 2., aktualisierte und überarbeitete Auflage, S. 275–293). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Frey, A. & Ehmke, T. (2008). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. In M. Prenzel, I. Gogolin & H.-H. Krüger (Hrsg.), *Kompetenzdiagnostik* (Bd. 34, S. 169–184). Wiesbaden: VS Verlag für Sozialwissenschaften. https://doi.org/10.1007/978-3-531-90865-6_10

Frey, A., Hartig, J. & Moosbrugger, H. (2009). Effekte des adaptiven Testens auf die Motivation zur Testbearbeitung am Beispiel des Frankfurter Adaptiven Konzentrationsleistungs-Tests. *Diagnostica*, *55*(1), 20–28. https://doi.org/10.1026/0012-1924.55.1.20

Green, B. F. (2012). The Promise of Tailored Tests. *Principals of Modern Psychological Measurement: A Festschrift for Frederic M. Lord*, 69.

Hartmann, S., Mathesius, S., Stiller, J., Straube, P., Krüger, D. & Upmeier zu Belzen, A. (2015). Kompetenzen der naturwissenschaftlichen Erkenntnisgewinnung als Teil des Professionswissens zukünftiger Lehrkräfte: Das Projekt Ko-WADiS. In B. Koch-Priewe, A. Köker, J. Seifried & E. Wuttke (Hrsg.), *Kompetenzerwerb an Hochschulen: Modellierung und Messung. Zur Professionalisierung angehender Lehrerinnen und Lehrer sowie frühpädagogischer Fachkräfte* (S. 39–58). Bad Heilbrunn: Verlag Julius Klinkhardt.

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

Ling, G., Attali, Y., Finn, B. & Stone, E. A. (2017). Is a Computerized Adaptive Test More Motivating Than a Fixed-Item Test? *Applied Psychological Measurement*, *41*(7), 495–511. https://doi.org/10.1177/0146621617707556

Magis, D., Yan, D. & Davier, A. A. von. (2017). *Computerized Adaptive and Multistage Testing with R*. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-69218-0

Mayer, J. (2007). Erkenntnisgewinnung als wissenschaftliches Problemlösen. In D. Krüger & H. Vogt (Hrsg.), *Theorien in der biologiedidaktischen Forschung. Ein Handbuch für Lehramtsstudenten und Doktoranden* (Springer-Lehrbuch, 1st ed., S. 177–187). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.

Moosbrugger, H. & Kelava, A. (Hrsg.). (2012). *Testtheorie und Fragebogenkonstruktion* (Springer-Lehrbuch, 2., aktualisierte und überarbeitete Auflage). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-20072-4

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

Patsula, L. N. (1999). *A comparison of computerized adaptive testing and multi-stage testing.* Dissertation. University of Massachusetts Amherst.

Sari, H. I., Yahsi-Sari, H. & Huggins-Manley, A. C. (2016). Computer Adaptive Multistage Testing: Practical Issues, Challenges and Principles. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, 388. https://doi.org/10.21031/epod.280183

Straube, P. (2016). *Modellierung und Erfassung von Kompetenzen naturwissenschaftlicher Erkenntnisgewinnung bei (Lehramts-) Studierenden im Fach Physik*. Dissertation. Freie Universität Berlin, Berlin.

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

Wellnitz, Nicole (2012). Kompetenzstruktur und -niveaus von Methoden der naturwissenschaftlichen Erkenntnisgewinnung. Berlin, Logos.

Zheng, Y. & Chang, h.-h. (2014). Multistage testing, on-the-fly multistage testing, and beyond. In Y. Cheng & h.-h. Chang (Hrsg.), *Advancing methodologies to support both summative and formative assessments* (Chinese American Educational Research and Development Association book series).