S. Franziska C. Wenzel[1], Claudia Krille[1], Sabine Fabriz[1] & Holger Horz[1]

Adaptive formative E-Assessments in der Lehrer*innenbildung

Zusammenfassung
Im Folgenden wird die Entwicklung eines computerbasierten adaptiven Tests (CAT) als formatives E-Assessment vorgestellt, das in bildungswissenschaftlichen Lehrveranstaltungen im Lehramtsstudium eingesetzt werden kann. In einem aufwändigen Entwicklungsprozess wurden Aufgaben zusammengestellt und ein adaptives Feedbackkonzept entwickelt. Begleitend zu einer Einführungsvorlesung wurde der CAT eingesetzt und evaluiert. Der Beitrag beschreibt die Entwicklungsschritte und erste Evaluationsergebnisse.
Schlagworte: Formatives Assessment, Bildungswissenschaften, adaptives Testen, Feedback

1. Einleitung

Digitale Lehr- und Lernmedien sind aus der Hochschule nicht mehr wegzudenken. Unter anderem erlauben Lernmanagementsysteme (LMS) den Austausch von Lehr-Lernmaterialien (z. B. Vorlesungsfolien, Literatur, Studierendenprodukte) und die Kommunikation zwischen Lehrenden und Studierenden. Darüber hinaus erlaubt die Infrastruktur der meisten Hochschulen, dass zunehmend elektronische Prüfungsformate eingesetzt werden (Friedrich, 2015). Vorteile solcher computerbasierten Prüfungen sind zum Beispiel die automatisierte und sofortige Auswertung von Studierendenantworten sowie die Möglichkeit von Rückmeldungen an Klausurteilnehmende (Jurecka & Hartig, 2007; Kröhne & Martens, 2011) oder die Möglichkeit zu objektiveren Auswertungen, da nicht leistungsrelevante Merkmale wie beispielsweise das Schriftbild die Bewertung nicht beeinflussen. Insgesamt stellt die Erfassung akademischer Leistung eines der Kernelemente von Hochschullehre dar (siehe auch Astin & Antonio, 2012), wobei eine möglichst präzise Messung und valide Interpretation von Testergebnissen von großer Bedeutung ist. Dies trifft nicht nur auf summative Bewertungen („assessment of learning", National Research Council, 1998) im Rahmen von Klausuren zu. Auch bei formativen Leistungsüberprüfungen soll angestrebt werden, Studierenden möglichst genaue Informationen über den aktuellen Lernstand und den Lehr-Lernprozess zur Verfügung zu stellen („assessment for learning", Brown, 2004).

Um diesem Anspruch gerecht zu werden, bieten sich adaptive Tests an. Dabei wird jedem Prüfling ein individueller Test präsentiert, der auf die Fähigkeiten der Person zugeschnitten und abhängig vom zuvor gezeigten Antwortverhalten ist (Cella et al., 2007). Damit wird ein Vorlegen von zu leichten oder zu schwierigen Aufgaben vermieden. Vorteil eines solchen Vorgehens ist unter anderem, dass weniger Aufgaben und somit weniger Testzeit benötigt werden, um ein genaues diagnostisches Ergebnis und damit eine Grundlage für eine präzise individuelle Rückmeldung zu erzielen. Besonders geeignet für solche computerbasierten adaptiven Tests (CATs) sind Aufgabenpools,

1 Arbeitsbereich Pädagogische Psychologie, Goethe-Universität Frankfurt, Deutschland

die unter Nutzung von Modellen der Item-Response-Theory (IRT, z. B. van der Linden & Hambleton, 1997) skaliert wurden, da die Nutzung einer gemeinsamen Metrik die Vergleichbarkeit von Testergebnissen bei unterschiedlichen adaptiv zusammengestellten Testformen ermöglicht. Die computerbasierte Umsetzung ermöglicht außerdem einen zeitlich und örtlich flexiblen Zugriff auf das E-Assessment. Zudem gibt es Hinweise darauf, dass CATs, verglichen mit traditionellen Paper-pencil-Prüfungen, die Motivation von Studierenden steigern können (Maravić Čisar et al., 2016) und als fair wahrgenommen werden (McCoubrie, 2004).

Ein wichtiger Aspekt für gelingende formative Assessments ist das Feedback an die Lernenden (z. B. Bimba et al., 2017). Es stellt ein zentrales Instrument dar, um Lernende nach der Bearbeitung von Aufgaben über deren Leistung zu informieren (Yorke, 2003). Formatives Feedback kann aber auch Informationen zur Unterstützung des weiteren Lernprozesses und Motivationssteigerung enthalten (Narciss, 2006). Damit kann selbstreguliertes Lernen angeregt und gefördert werden (z. B. Clark, 2012). Dafür müssen die im Feedback enthaltenen Informationen über eine bloße Leistungseinschätzung hinausgehen (z. B. Clark, 2012), zum Beispiel durch Angaben zu den Aufgabenanforderungen, inhaltlichen Konzepten oder Anregungen dazu, wie ähnliche Aufgaben zukünftig besser gelöst werden können (z. B. Narciss, 2006).

Der vorliegende Beitrag präsentiert die Konzeption eines CAT als formatives E-Assessment mit individueller Rückmeldung für Lehramtsstudierende im Rahmen einer Einführungsvorlesung in den Bildungswissenschaften. Bildungswissenschaftliches Wissen (BWW) stellt einen wesentlichen Aspekt der professionellen Lehrkräftekompetenz dar (KMK, 2004; Kunter et al., 2017). Es gibt Hinweise darauf, dass BWW negative Effekte wie Erschöpfung abmildern sowie die Reflexion des eigenen Unterrichts positiv beeinflussen kann (Überblick in Kunter et al., 2017). Trotzdem wird BWW bisher – im Vergleich zu Fachwissen und fachdidaktischem Wissen – häufig bei didaktischen Überlegungen und empirischen Studien nicht berücksichtigt (z. B. Kunter et al., 2017).

Im Rahmen des BMBF-geförderten Projekts „Computerbasiertes adaptives Testen im Studium“ (CaTS) wurde deshalb ein formativer CAT entwickelt, der im Rahmen von Lehrveranstaltungen zum BWW im Lehramtsstudium auch in Lehrveranstaltungen mit großen Studierendengruppen eingesetzt werden kann, um Lernprozesse individuell zu unterstützen. Im Folgenden wird aufgezeigt, welche Schritte für die Entwicklung eines adaptiven formativen E-Assessments unternommen wurden. Aufbauend auf den Ergebnissen einer begleitenden Evaluation werden die Herausforderungen für die Umsetzung eines solchen Systems diskutiert.

2. Konzeption eines adaptiven formativen E-Assessments

In Anlehnung an die von Thompson und Weiss (2011) vorgeschlagenen Schritte zur Entwicklung eines CAT wurden im Projekt CaTS fünf Schritte unternommen, wobei neben technischen und methodischen Aspekten zur Adaptivität auch inhaltliche Aspekte zum formativen Charakter des E-Assessments aufgegriffen wurden.

1. *Pilotierungsstudie*. Zunächst wurden 40 bereits existierende Aufgaben im Single- oder Multiple-Choice-Format zu BWW auf ihre Eignung für den Einsatz im Rahmen eines CAT geprüft. Die Aufgaben wurden von 673 Studierenden bearbeitet, Rasch-skaliert (Rasch, 1960/1980) und Aufgaben mit ungünstigen Eigenschaften eli-

miniert oder zur Überarbeitung identifiziert. Darüber hinaus wurde geprüft, inwiefern ein Wechsel der Darbietungsform (Papier zu Computer) einen Einfluss auf die Aufgabenschwierigkeiten hat (sog. *mode effects*; Kröhne & Martens, 2011). 29 eher leichte Aufgaben konnten für eine weitere Verwendung im Rahmen des CAT ausgewählt werden (Schwierigkeitsverteilung im Mittel -1.02 Logits, *SD* = 1.18). Es wurde ein minimaler, nicht signifikanter Vorteil für die Papierversion gefunden (Unterschied von 0.05 Logits, $\chi^2 = 1.78$, $df = 1$). Damit liegen günstige Voraussetzungen für die Weiterentwicklung zu einem formativen CAT vor.

2. *Entwicklung einer Aufgabendatenbank.* Um den Inhaltsbereich der Bildungswissenschaften zu repräsentieren und zu strukturieren, wurde basierend auf den KMK-Standards der Lehrkräftebildung (KMK, 2004) eine Ontologie entwickelt. Der resultierende Strukturbaum diente als Grundlage für die Aufgabenentwicklung und eine technisch umgesetzte Aufgabendatenbank. Im Rahmen eines mehrstufigen Entwicklungs- und Revisionsprozesses wurden weitere Aufgaben erstellt, durch Experten geprüft und gegebenenfalls überarbeitet. Der resultierende Pool von 133 Aufgaben im Single- oder Multiple-Choice-Format (Beispiel siehe Abbildung 1) repräsentierte die verschiedenen Themenbereiche sowie ein möglichst breites Schwierigkeitsspektrum des zu vermittelnden BWW.
3. *Kalibrierungsstudie*. Zur Bestimmung der für die CAT-Zusammenstellung relevanten Aufgabenmerkmale wurden die Aufgaben 264 Lehramtsstudierenden vorgelegt. Durch die Nutzung eines balancierten unvollständigen Testheftdesigns wurden jedem Studierenden etwa 60 Aufgaben in einer Testzeit von 90 Minuten vorgelegt. Auf Grundlage der Studierendenantworten wurden die Aufgaben Rasch-skaliert (Rasch, 1960/1980) und Aufgabenschwierigkeiten geschätzt. 30 Aufgaben wurden aufgrund schlechter Fit-Werte oder niedriger Korrelation mit der Gesamtskala aus dem Aufgabenpool eliminiert, sodass 103 Aufgaben für den CAT zur Verfügung standen (Schwierigkeitsverteilung im Mittel -0.24 Logits, *SD* = 1.23; MLE-Reliabilität bei .76).
4. *Spezifizierung des adaptiven formativen E-Assessments*. Im Projekt CaTS wurde eine Schnittstelle für Dozierende entwickelt, die es ihnen ermöglicht, auch ohne umfangreiche technische oder statistische Kenntnisse mit dem zuvor erarbeiteten Aufgabenpool ein Assessment zusammenzustellen. Das in CaTS eingesetzte adaptive formative E-Assessment für die Nutzung in Vorlesungen zu BWW setzte sich aus jeweils 15 Aufgaben zusammen und konnte während des Semesters durch Studierende beliebig oft genutzt werden. Es wurden insgesamt drei Tests über das Semester verteilt freigeschaltet, die den bis dahin behandelten Stoff abdeckten. Zudem wurde ein Rückmeldungskonzept mit Informationen zur Leistung und Prüfungsvorbereitung implementiert (vgl. Narciss, 2006). Dieses elaborierte Feedback enthielt die Schwierigkeit der gezeigten Aufgaben und die Anzahl richtig gelöster Aufgaben, die individuelle Leistung des Studierenden sowie eine Prognose hinsichtlich der Klausurleistung. Zur weiteren Klausurvorbereitung wurden Literaturempfehlungen passend zum diagnostizierten Lernstand gegeben. Die Verständlichkeit und wahrgenommene Nützlichkeit des Feedbacks wurde im Vorfeld in einer Lautdenken-Studie geprüft.
5. *Implementierung und Evaluation*. Das so konzipierte E-Assessment wurde im Rahmen zweier paralleler Einführungsvorlesungen zu BWW eingesetzt, indem es in das bestehende LMS (OLAT) eingebettet und mithilfe von freiwilligen Begleiterhebun-

gen evaluiert wurde. Das Vorgehen sowie die Ergebnisse werden im Folgenden präsentiert.

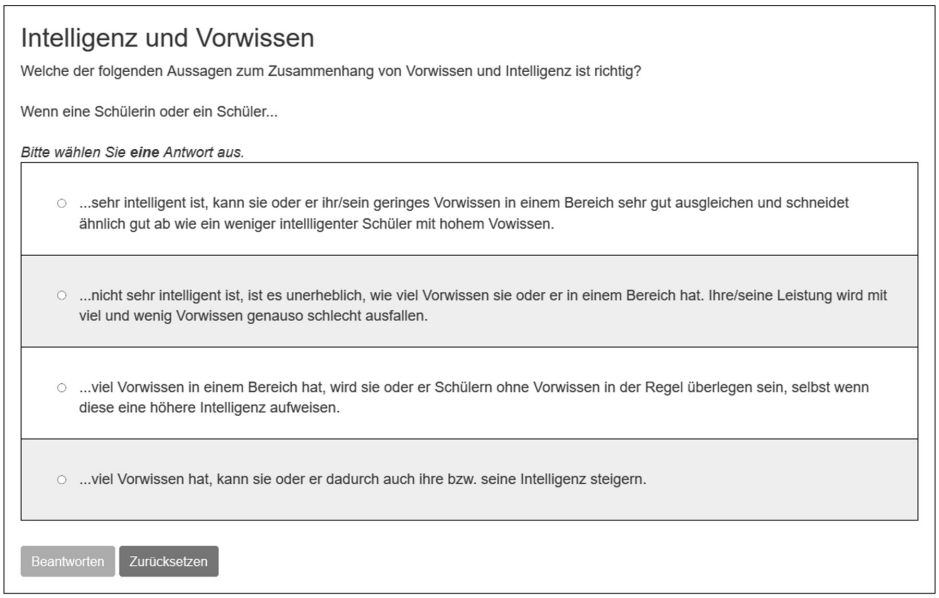

Abbildung 1: Beispielaufgabe aus dem CAT.

3. Methode

Drei Mal im Semester wurde ein neues E-Assessment zur Verfügung gestellt, in dem die bis dato in der Vorlesung behandelten Themen abgebildet waren. Die Studierenden konnten diese freiwillig beliebig oft nutzen. 512 Lehramtsstudierende nahmen an der Evaluationsbefragung und/oder mindestens einem E-Assessment teil. Im Rahmen der Evaluation wurden die Studierenden gebeten, die Aufgaben und das Gesamtassessment sowie das Feedback hinsichtlich ihrer Passung zur Vorlesung und Nützlichkeit zu bewerten (21 Items: offen oder 5-stufige Likertskala mit 1: stimme gar nicht zu bis 5: stimme voll zu; in Anlehnung an Strijbos et al., 2010).

4. Ergebnisse

Im Hinblick auf die studentische Nutzung des adaptiven formativen E-Assessments zeigte sich, dass lediglich 65 Studierende jedes der drei E-Assessments mindestens einmal nutzten. Generell wurden die drei Assessments über das Semester hinweg in ansteigender Häufigkeit genutzt (A1: $M = 0.99$, $SD = 2.77$, $n = 111$; A2: $M = 1.46$, $SD = 3.47$, $n = 134$; A3: $M = 5.26$, $SD = 6.99$, $n = 303$) und die Leistung nahm stetig zu (A1: $M = 0.24$, $SD = 0.68$; A2: $M = 0.37$, $SD = 0.74$; A3: $M = 0.79$, $SD = 0.80$). Allerdings lassen die Ergebnisse keinen Rückschluss zu, ob die Leistung einzelner Studieren-

der gestiegen ist oder leistungsfähigere Studierende erst gegen Ende des Semesters das Assessment nutzten.

Insgesamt schätzten die Studierenden die Nützlichkeit des E-Assessments auf einem mittleren Niveau ein (M = 3.20, SD = 0.60). Dies gilt auch für die darin enthaltenen Aufgaben (M = 3.61, SD = 0.55) und das individuelle Feedback (M = 3.24, SD = 0.62). Studierende empfanden es als positiv, dass sie sich mit den Vorlesungsinhalten noch einmal auseinandersetzen oder diese wiederholen konnten (n = 7), sich die Aufgabenschwierigkeit an ihre Fähigkeit anpasste (n = 4), es Rückmeldung zur Leistung gab und Erfahrungen mit Fragen zur Vorlesung gesammelt werden konnten (n = 3). Negativ merkten Studierende vor allem an, dass das Feedback nicht Aufgaben-bezogen war (n = 13) und einige sahen keinen Bezug zwischen den Aufgaben und Vorlesungsinhalten (n = 3).

5. Diskussion

Es ist gelungen, ein adaptives formatives E-Assessment für BWW im Lehramtsstudium zu entwickeln und für zwei Veranstaltungen zu implementieren. Über das verwendete LMS konnten die einzelnen Assessments einfach für Studierende zur Verfügung gestellt werden und orts- und zeitunabhängig genutzt werden. Das Angebot ermöglicht es den Studierenden, sich über die Vorlesung hinaus mit dem Stoff auseinanderzusetzen, was positiv wahrgenommen wurde. Allerdings ist der Aufwand zur Entwicklung eines solchen Instruments sehr hoch. Ein solcher Aufwand lohnt sich insbesondere für regelmäßig angebotene Veranstaltungen mit hohen Studierendenzahlen, bei denen sich die zu vermittelnden Inhalte nicht grundlegend ändern. Durch die Nutzung der KMK-Standards als Basis sollte das Assessment auch auf andere Veranstaltungen und Universitäten übertragbar sein, was bereits erprobt wird. Der bisher entwickelte Aufgabenpool stellt aus unserer Sicht eine gute Grundlage dar, um ihn zukünftig durch weitere Aufgaben zu ergänzen und für weitere Veranstaltungen zu spezifizieren oder in anderen Phasen der Lehrerbildung zu nutzen. Durch die entwickelte Dozierendenschnittstelle können auch ohne tiefere IRT-Kenntnisse adaptive E-Assessments gestaltet und Studierenden zur Verfügung gestellt werden.

Bei der Analyse des Nutzungsverhaltens zeigte sich, dass sich Studierende vor allem kurz vor der Prüfung vorbereiten und auch zusätzliche Lerngelegenheiten nutzen. Die Verfügbarkeit des formativen E-Assessments im Semesterverlauf nutzten immerhin einige Studierende, aber bei weitem nicht alle. In weiteren Untersuchungen soll geprüft werden, welche Maßnahmen Studierende dabei unterstützen können, sich kontinuierlich mit den Fachinhalten auseinanderzusetzen und das Assessment für ein selbstreguliertes Lernen zu nutzen (Panadero et al., 2017). Ein möglicher Ansatzpunkt könnten die angesprochenen Verbesserungswünsche der Studierenden sein. Darüber hinaus sollte untersucht werden, welche Lernendenmerkmale die Nutzung eines solchen Assessments beeinflussen und inwiefern diese gefördert werden kann. Im Rahmen der vorgestellten Studie wurden leistungsrelevante Merkmale erhoben und sollen in zukünftigen Analysen mit einbezogen werden.

Das Projekt „Computerbasiertes Adaptives Testen im Studium“ (CaTS) wurde im Rahmen der Förderlinie „Forschung zur digitalen Hochschulbildung“ aus Mitteln des Bundesministeriums für Bildung und Forschung (BMBF) unter dem Kennzeichen 16DHL1008/16DHL1009 gefördert. Die Verantwortung für den Inhalt dieser Veröffentlichung liegt bei den Autor*innen.

Literatur

Astin, A. W. & Antonio, A. L. (2012). *Assessment for excellence: The philosophy and practice of assessment and evaluation in higher education*. Lanham, Maryland: Rowman.

Bimba, A. T., Idris, N., Al-Hunaiyyan, A., Mahmud, R. B. & Shuib, N. L. B. M. (2017). Adaptive feedback in computer-based learning environments: A review. *Adaptive Behavior*, *25*(5), 217–234. https://doi.org/10.1177/1059712317727590

Brown, S. (2004). Assessment for learning. *Learning and Teaching in Higher Education, 1,* 81–89.

Cella, D., Gershon, R., Lai, J. S. & Choi, S. (2007). The future of outcomes measurement: Item banking, tailored short-forms, and computerized adaptive assessment. *Quality of Life Research*, *16*, 133–141. https://doi.org/10.1007/s11136-007-9204-6

Clark, I. (2012). Formative assessment: Assessment is for self-regulated learning. *Educational Psychology Review*, *24*(1), 205–249. https://doi.org/10.1007/s10648-011-9191-6

Friedrich, J. D. (2015). *Hochschulforum Digitalisierung. E-Assessment als Herausforderung. Handlungsempfehlungen für die Hochschulpolitik*. Berlin: Hochschulforum Digitalisierung.

Jurecka, A. & Hartig, J. (2007). Anwendungsszenarien computer- und netzwerkbasierten Assessments. In Bundesministerium für Bildung und Forschung (Hrsg.), *Möglichkeiten und Voraussetzungen technologiebasierter Kompetenzdiagnostik* (S. 69–79).

Kröhne, U. & Martens, T. (2011). Computer-based competence tests in the national educational panel study: The challenge of mode effects. *Zeitschrift für Erziehungswissenschaft*, *14*(S2), 169–186. https://doi.org/10.1007/s11618-011-0185-4

Kultusministerkonferenz (KMK) (2004). *Standards für die Lehrerbildung: Bildungswissenschaften.* Abgerufen am 01.07.2020 von: https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2004/2004_12_16-Standards-Lehrerbildung-Bildungswissenschaften.pdf

Kunter, M., Kunina-Habenicht, O., Baumert, J., Dicke, T., Holzberger, D., Lohse-Bossenz, H., Leutner, D., Schulze-Stocker, F. & Terhart, E. (2017). Bildungswissenschaftliches Wissen und professionelle Kompetenz in der Lehramtsausbildung. In C. Gräsel & K. Trempler (Hrsg.), *Entwicklung von Professionalität pädagogischen Personals* (S. 37–54). Wiesbaden: Springer. https://doi.org/10.1007/978-3-658-07274-2_3

Maravić Čisar, S., Čisar, P. & Pinter, R. (2016). Evaluation of knowledge in object oriented programming course with computer adaptive tests. *Computers & Education*, *92-93*, 142–160. https://doi.org/10.1016/j.compedu.2015.10.016

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, *26*(8), 709–712. https://doi.org/10.1080/01421590400013495

Narciss, S. (2006). *Informatives tutorielles Feedback: Entwicklungs- und Evaluationsprinzipien auf der Basis instruktionspsychologischer Erkenntnisse*. *Pädagogische Psychologie und Entwicklungspsychologie: Bd. 56*. Münster: Waxmann.

National Research Council (1998). *High stakes: Testing for tracking, promotion, and graduation.* National Academies Press.

Panadero, E., Jonsson, A. & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. (Original work published 1960). Chicago: University of Chicago Press.

Strijbos, J. W., Narciss, S. & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction*, *20*(4), 291–303.

https://doi.org/10.1016/j.learninstruc.2009.08.008
Thompson, N. A. & Weiss, D. J. (2011). A framework for the development of computerized adaptive tests. *Practical Assessment, Research & Evaluation, 16*, Article 1.
Van der Linden, W. J. & Hambleton, R. K. (1997). Item response theory: Brief history, common models, and extensions. In W. J. van der Linden & R. K. Hambleton (Hrsg.), *Handbook of modern item response theory* (S. 1–28). New York: Springer. https://doi.org/10.1007/978-1-4757-2691-6
Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education, 45*, 477–501. https://doi.org/10.1023/A:1023967026413