Christian SPODEN[1] (Emden), Aron FINK (Frankfurt), Andreas FREY (Frankfurt), Hanna KÖHLER (Jena) & Patrick NAUMANN (Frankfurt)

Kompetenzorientierung und Fairness bei individualisierten E-Klausuren

Zusammenfassung

Die stärkere Individualisierung des Studiums erfordert adaptiv und individuell zugeschnittene Prüfungsformen, die Anforderungen der Kompetenzorientierung sowie der Vergleichbarkeit und Fairness von Prüfungsleistungen erfüllen. Kompetenzorientierte adaptive E-Klausuren stellen hier ein erfolgsversprechendes Konzept dar, welches im vorliegenden Beitrag erläutert wird. Es wird ferner eine Anwendung dargestellt, aus der sich Empfehlungen zur Umsetzung im regulären Klausurbetrieb ableiten lassen. Abschließend werden Erweiterungen und Herausforderungen der Konzeption diskutiert.

Schlüsselwörter

E-Klausuren, Kompetenzdiagnostik, Computerisiertes Adaptives Testen, Test Equating

1 E-Mail: christian.spoden@hs-emden-leer.de

DOI: 10.3217/zfhe-17-01/08

Competence orientation and fairness in individualised e-exams

Abstract

Individualised pathways in higher education studies require adaptive and customised exams, which meet the requirements of a genuine competence orientation and the comparability and fairness of exams. For this purpose, competence-oriented adaptive e-exams provide a promising concept, which is explained in this paper. In addition, the paper presents an trial implementation, from which recommendations for implementation in regular exam operations can be derived. Possible extensions and challenges of the concept are summarised in the discussion.

Keywords

e-exams; assessment of competencies; computerised adaptive testing; test equating

1 Die Individualisierung des Studiums als Herausforderung für das Prüfungswesen

Das Hochschulwesen in Deutschland ist seit einiger Zeit durch eine stärkere Ausdifferenzierung von Fachgebieten sowie der Aufspaltung der Studiengänge und Vergrößerung des Studienangebotes gekennzeichnet (HACHMEISTER & GREVERS, 2019). In vielen Fächern beschränkten sich Wahlmöglichkeiten lange auf den Studienort mit dem dort ansässigen Lehrpersonal aus einer Gruppe gleichartig konzipierter Diplom- oder Magister-Studiengänge an Universitäten oder Fachhochschulen. Mit dem Wechsel auf Bachelor-/Master-Studiengänge infolge der Bologna-Reform hat sich dies gewandelt. Heute existieren in vielen Fächern nicht nur Hochschul- und Fachhochschulstudiengänge, sondern Studiengänge mit unterschiedlicher Schwerpunktsetzung sowie Studiengänge in neuen Querschnittsbereichen. Auch hat sich die Durchlässigkeit zwischen Studiengängen infolge der Etablierung von Bachelor-/Master-Studiengängen erhöht. Als Resultat eines gestiegenen Bewusstseins um die Diversität der Studierendenschaft im Hinblick auf Lernvoraussetzungen

und -gelegenheiten im Studium entwickelte sich außerdem ein breiteres Angebot unterschiedlicher Studienformen (Präsenz- oder Fernstudium in Voll- oder Teilzeit, berufsbegleitendes Studium), in dem die Studierenden oft auf unterschiedliche Art studieren (zur Mediennutzung ZAWACKI-RICHTER, 2015). Schließlich hat sich etwa in Deutschland der Anteil ausländischer Studierender und damit die Diversität der Lernvoraussetzungen erhöht.

Für das Prüfungswesen ist diese Tendenz zu einer stärkeren Individualisierung des Studiums eine Herausforderung. Einerseits fehlen überzeugende Konzepte, um diesen vielfältigeren Studienwegen bei Prüfungen gerecht zu werden. Andererseits geht mit Individualisierung auch die Frage nach der Vergleichbarkeit und somit der Fairness von Prüfungen, der Aufrechterhaltung wissenschaftlicher Standards und der Umsetzung der geforderten kompetenzorientierten Prüfungen unter diesen Bedingungen einher. Studierende haben nicht nur ein Anrecht auf eine angemessene Bilanzierung der im Studium angeeigneten Kompetenzen, sondern erwarten, dass Prüfungen überzeugende Fairness-Konzepte enthalten (SAMBELL, McDOWELL & BROWN, 1997).

Die Herausforderung liegt nun darin, bei Hochschulprüfungen Individualisierung, Vergleichbarkeit und Fairness der Bewertung sowie Kompetenzorientierung gemeinsam zu realisieren. Ein möglicher Lösungsansatz für diese Problemstellung liegt in der konsequenten Anwendung etablierter und auf der *Item Response Theory* (IRT; van der LINDEN, 2016) basierender Methoden aus den Bereichen Educational Measurement und Psychometrie. Insbesondere bietet sich das computerisierte adaptive Testen (CAT; FREY, 2020) als individualisierter Ansatz für computerbasierte Test- und Prüfungsverfahren an. Mit dem vorliegenden Beitrag werden daher zwei Ziele verfolgt: Erstens wird ein neues Konzept individualisierter und gleichzeitig fairer, kompetenzorientierter Hochschulklausuren beschrieben. Zweitens wird die Umsetzung dieses Konzeptes für Studiengangsverantwortliche und Prüfende an einem illustrativen Beispiel in drei Entwicklungsphasen dargestellt. Abschließend werden in der Diskussion Voraussetzungen und Einschränkungen der Nutzbarkeit aufgeführt, aber auch mögliche Erweiterungen der Konzeption vorgeschlagen.

2 Konzeption adaptiver kompetenzorientierter E-Klausuren

Kompetenzorientierte adaptive E-Klausuren verknüpfen eine ernsthafte Kompetenzorientierung und eine über Studierendengruppen hinweg konstante und damit faire Berichtsmetrik für Klausurergebnisse in einem Lehrgebiet mit der Individualisierung der Prüfungen durch die computerisiert-adaptive Auswahl der Prüfungsaufgaben (im Folgenden: Items). Sie basieren auf vier Prinzipien: *Erstens* sind adaptive kompetenzorientierte E-Klausuren als Instrumente zum Messen des Umfangs des Kompetenzerwerbs konzipiert. Um den Messgegenstand zu spezifizieren und festzulegen, was in der Klausur geprüft werden soll, müssen die Lernziele einer Lehrveranstaltung anhand von Inhaltsbereichen und kognitiven Anforderungen explizit dargelegt werden. Dies ist notwendig, um das Erreichen kompetenzorientierter Lernziele zu beurteilen und kein reines Wissen abzufragen. Für die Abgrenzung unterschiedlicher kognitiver Anforderungsniveaus sind Lehrzieltaxonomien (z. B. BLOOM, 1956) als Orientierung geeignet. Für jede Kombination von Inhalt und kognitivem Anforderungsniveau sind dann Items zu konstruieren. Im Sinne des *Constructive Alignment* (BIGGS, 1996) ist diese Strukturierung dahingehend bereits nützlich, Lernaktivitäten zu planen und Lernziele, -methoden und Prüfung schon bei der Planung einer Lehrveranstaltung aufeinander zu beziehen. In aller Regel werden zudem kontextualisierte Klausuritems benötigt, die auf typische wissenschaftliche und praktische Tätigkeiten in der jeweiligen Disziplin Bezug nehmen.

Um Aussagen über das Erreichen der in den Modulkatalogen festgelegten Lernziele und Kompetenzanforderungen bei Klausuren ableiten zu können, ist *zweitens* ihre Konzeption als kriteriumsorientierte Tests (HERZBERG & FREY, 2011) und die Auswertung der Klausuren mithilfe der IRT zielführend. Die Modelle der IRT erlauben Aussagen dazu, mit welcher Wahrscheinlichkeit eine mit Lernzielen verknüpfte Aufgabe eines bestimmten Schwierigkeitsgrads (δ_i) von Studierenden mit einer aus den kodierten Itemlösungen (richtige Lösung: 1; falsche Lösung: 0) geschätzten Kompetenz θ_v (für Student v) bewältigt werden können. Sie sind damit als testtheoretisches Rahmenwerk für kompetenzorientierte Tests besonders geeignet. Im einparametrischen IRT-Modell wird diese Wahrscheinlichkeit bestimmt als:

$$P(x = 1 \mid \theta_v, \delta_i) = \frac{e^{(\theta_v - \delta_i)}}{1 + e^{(\theta_v - \delta_i)}}. \quad (1)$$

Die resultierenden Kompetenzschätzungen können in einem zweiten Schritt über vereinfachte Standard-Setting-Verfahren wie der Bookmark-Methode (MITZEL, LEWIS, PATZ & GREEN, 2001) in Notenstufen überführt werden. Die Grenzen zwischen Notenstufen werden mit Schlüsselanforderungen verknüpft, die in bestimmten Items abgebildet werden, sodass die Grenze zwischen Notenstufen an der Schwierigkeit dieser Items festgemacht wird. Mithilfe dieses Vorgehens werden die auf Kompetenzschätzungen basierenden Klausurergebnisse an die individuelle Bewältigung der Kompetenzanforderungen geknüpft. Zweifelhafte Bewertungspraktiken, nach denen das Bestehen der Klausur etwa an einem willkürlich festgesetzten Anteil richtiger Itemlösungen festgemacht wird, werden ausgeschlossen.

Drittens ist die Etablierung einer zeitlich konstanten Kompetenzskala und die kohortenübergreifende Verortung von Kompetenzen auf dieser Skala zu nennen. Mithilfe von *Equating*-Methoden (KOLEN & BRENNAN, 2014) ist es wiederum im Rahmenwerk der IRT möglich, den Bewertungsmaßstab einer Klausur auf die Klausurbewertungen in nachfolgenden Kohorten zu übertragen. Hierzu werden bei der Bestimmung der Studierendenkompetenz und, darauf aufbauend, der Noten die Unterschiede in der mittleren Schwierigkeit zwischen unterschiedlichen eingesetzten Itemstichproben berücksichtigt. Notwendig ist dafür, dass ein kleiner Teil der Items (Linkitems) aus vorangegangenen Klausuren erneut eingesetzt und im Rahmen einer *Fixed Parameter Calibration* wieder auf die gleichen Schwierigkeiten geschätzt werden kann. Auf Basis dieser konstanten Schätzer wird die statistische „Verankerung" der Klausuren über die Zeit vorgenommen. Die Schwierigkeiten werden dabei im ersten Schritt frei geschätzt. Dann werden die Schwierigkeiten über eine lineare Transformation um den geschätzten mittleren Kompetenzunterschied zwischen den Studierendenkohorten verschoben, wodurch für die Linkitems wieder die Schwierigkeiten aus vorherigen Klausuren erreicht werden sollten. Die in der Realität resultierenden Abweichungen werden statistisch auf *Item-Drift* (z. B. DeMARS, 2004) geprüft. In einem weiteren Schritt wird die Fixed Parameter Calibration mit fixierten Schwierigkeiten jener Linkitems, bei denen keine signifikanten Abweichungen im Drift-Test identifiziert wurden, vorgenommen. Abbildung 1 zeigt anhand sogenannter *Wright-Maps* beispielhaft, wie sich die Itemschwierigkeiten (rechts) und die

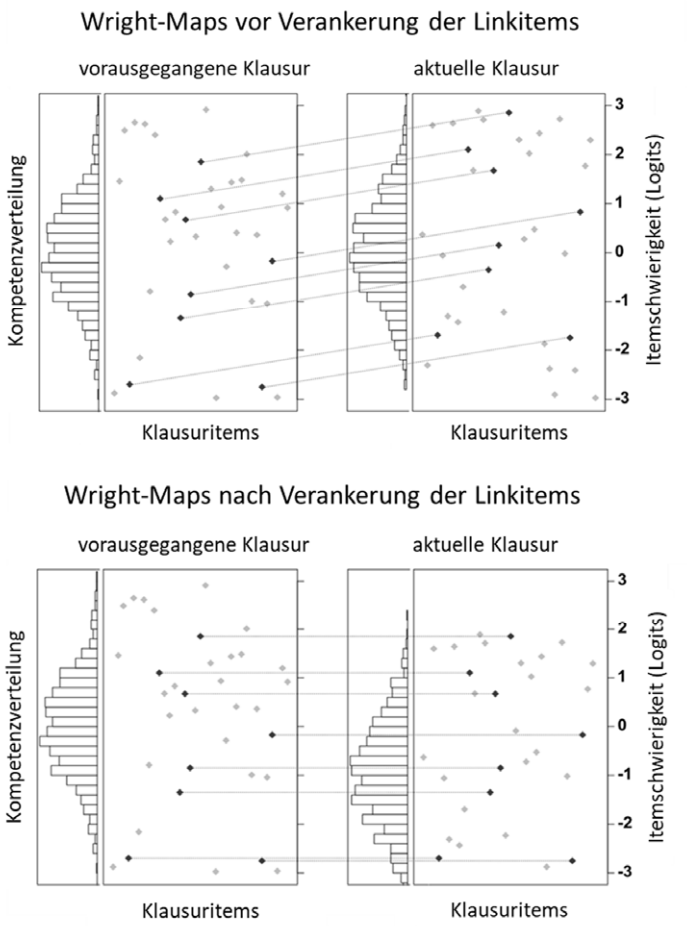

Abb. 1: Wright-Maps vor und nach Verankerung der Linkitems

Kompetenzverteilung (links) zwischen freier Skalierung und Fixed Item Parameter Calibration anhand dieser Verankerung verschieben. Wichtig für das Gelingen dieses Verfahrens ist die Vermeidung systematischer Verzerrungen der Schätzung der Itemschwierigkeiten, die durch gezieltes Vorbereiten der Itemlösung durch Studierende bei Bekanntwerden der Items resultieren könnte. Hierzu ist es notwendig, die Klausuritems geheim zu halten und nur Beispielitems freizugeben.

Das Konstanthalten des Bewertungsmaßstabs und die glaubhafte Vermittlung der Tatsache, dass bei einem Zweittermin oder einer Nachklausur keine bessere Note aufgrund leichterer Items zu erreichen ist (wie es bei Festlegung der Bestehensgrenze anhand eines festen Prozentsatzes richtiger Lösungen geschehen könnte), dürfte bei Studierenden zur Verringerung der Prokrastination beim Lernen beitragen. Bei gleichbleibenden Lernzielen und Anforderungen in Modulkatalogen leistet der konstante Bewertungsmaßstab unabhängig vom Kompetenzniveau der jeweils getesteten Kohorte insbesondere aber einen Beitrag zur Verbesserung der Fairness (FREY, SPODEN & BORN, 2020). Unter den Studierenden kursierende Itemsammlungen und „Studierendenfunk“ sorgen dafür, dass Studierende die Klausuranforderungen über Jahrgänge vergleichen. Das vorgeschlagene Vorgehen liefert Transparenz in der Vergleichbarkeit von Prüfungsergebnissen über Jahrgänge hinweg. Davon abgesehen sind Vergleiche der durchschnittlichen Kompetenzen über Studierendenkohorten geeignet, eine Form von Monitoring in der Lehre einzusetzen und gegebenenfalls Anpassungen vorzunehmen (siehe Beispiel unten).

Viertens ermöglichen adaptive E-Klausuren die individualisierte Itemauswahl durch den Computer. Die computerbasierte Durchführung ist bereits aufgrund der Reduktion des Prüfungsaufwands durch Automatisierung großer Teile der Klausuradministration, -auswertung und Ergebnisrückmeldung (Zeitgewinn ausführlich an einem Beispiel dargestellt in SPODEN, 2021) sowie der nachhaltigeren, weil papierlosen Klausurdurchführung vorteilhaft. Die Möglichkeit für eine adaptive Itemauswahl ergibt sich aus der Eigenschaft der IRT, dass Kompetenzen auf Basis unterschiedlicher Itemstichproben geschätzt und auf derselben Skala abgebildet werden können. Bei üblichen Klausuren wird allen Studierenden, unabhängig von Lernvoraussetzungen, individuellen Schwerpunktsetzungen im Studium oder Vorleistungen, der gleiche Itemsatz vorgelegt. Dies spiegelt nicht die Individualisierung des Studiums wider. Beim CAT werden die Klausuritems hingegen aus einem vorab zusammengestellten Itempool anhand bekannter Informationen wie der Item-

schwierigkeit vom Computer individuell ausgewählt. Meist zielt die Itemauswahl darauf ab, informative Items im Hinblick auf die Verortung der Studierenden auf der Kompetenzdimension vorzulegen und so im gesamten Kompetenzspektrum den Messfehler zu reduzieren. Im einparametrischen Modell bestimmt sich die Iteminformation als

$$I_i(\theta) = P_i(\theta)[1 - P_i(\theta)] \,. \tag{2}$$

Die maximale Information (Maximum Fisher Item Information Criterion) im Hinblick auf die Verortung der Kompetenz eines Prüflings besitzt ein Item, wenn die Lösungswahrscheinlichkeit gleich der Gegenwahrscheinlichkeit (bzw. die Itemschwierigkeit gleich der Kompetenzausprägung) ist. Der obere Teil von Abbildung 2 verdeutlicht diesen Itemauswahlprozess beim CAT in Form eines Flussdiagramms (nach FREY, 2020); unten ist in Form eines Entscheidungsbaums dargestellt, wie sich die Itemauswahl anhand der Schwierigkeit an der nach jedem bearbeiteten Item jeweils neu geschätzten Kompetenz orientiert. Die Reduktion des Messfehlers im gesamten Kompetenzspektrum durch CAT ist vorteilhaft, da dieser bei Tests ohne CAT-Einsatz im Vergleich zum mittleren Kompetenzbereich üblicherweise an den Rändern der Kompetenzverteilung höher ausfällt (FREY & EHMKE, 2007), sodass kritische diagnostische Entscheidungen (insb. Klausur bestanden / nicht bestanden im unteren Kompetenzbereich) mit einer hohen Fehlerwahrscheinlichkeit getroffen werden. Neben dem Kriterium der maximalen Information kann mit Verfahren des Content Management sichergestellt werden, dass die vorgegebenen Items weiteren Anforderungen genügen. Beispielsweise ist es möglich, Inhaltsbereiche entsprechend zu kodieren, bei der Itemauswahl diese Inhaltsbereiche systematisch abzudecken und so die Repräsentativität der Items in Bezug auf die Lehrinhalte bei adaptiver Itemvorgabe aufrechtzuerhalten. Denkbar ist auch, unterschiedliche Itemkontexte in Codes zu überführen, die vom Computer für eine noch stärker individualisierte Itemauswahl genutzt werden können. In Lehrveranstaltungen finden sich oft Studierende unterschiedlicher Studiengänge. Es wäre möglich, parallele Items zu entwickeln, bei denen der Kontext auf die unterschiedlichen Anforderungen in diesen Studiengängen ausgerichtet wird und der Computer studiengangsspezifisch passende Items auswählt. Als Beispiel zeigt Abbildung 3 ein Übungsitem zu der unten beschriebenen Klausur im Fach Forschungsmethoden der Erziehungswissenschaft mit einem wirtschaftspädagogischen Kontext (Kantine) sowie ein inhalt-

lich gleiches Parallelitem mit schulpädagogischer Kontextualisierung. Analog zum Itemkontext ist denkbar, dass Studierende bei der Klausurbearbeitung die in ihrem Arbeitsbereich jeweils üblichen Hilfsmittel wie fachspezifische Softwarepakete freigeschaltet bekommen.

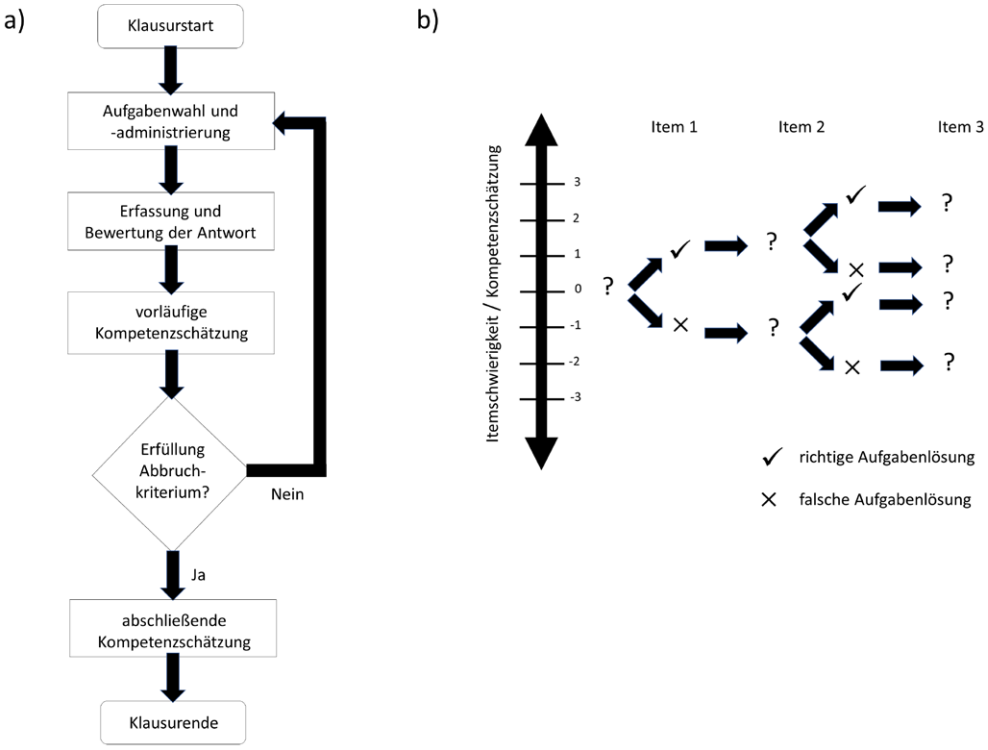

Abb. 2: a) Ablaufschema beim CAT. Angepasst aus „Computerisiertes adaptives Testen“, Frey, 2020, S. 505.

b) Entscheidungsbaum zum CAT

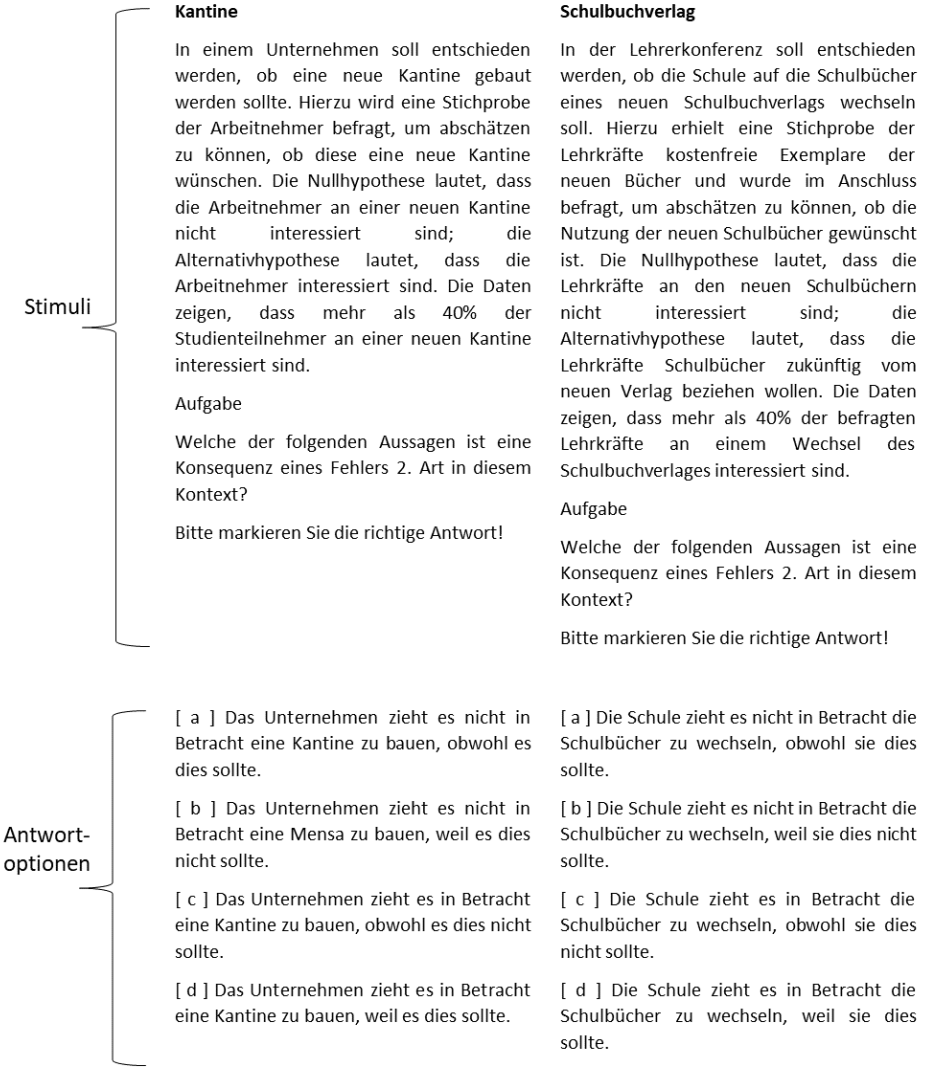

Abb. 3: Item aus einer Klausur „Forschungsmethoden der Erziehungswissenschaft“ in zwei kontextualisierten Fassungen

Einschränkend ist zu Möglichkeiten der Individualisierung zu bedenken, dass E-Klausuren im Hinblick auf die Authentizität, also die Überprüfbarkeit der Zuordnung einer Prüfungsleistung zu einem Prüfling, und die Integrität, die sich auf die Sicherstellung der Nicht-Veränderung der Daten nach der Prüfung bezieht, anderen Anforderungen als papierbasierte Klausuren genügen müssen. Einige Lösungen hierfür werden im Anwendungsbeispiel unten beschrieben. Grundsätzlich sind bei E-Klausuren außerdem Vorkehrungen zur sicheren und störungsfreien Durchführung zu treffen, etwa eine Absicherung gegen Stromausfall. Zumeist dürfte diese Absicherung in universitären Computerpools gut umsetzbar sein.

3 Anwendungsbeispiel: Entwicklung einer individualisierten Klausur zu empirischen Forschungsmethoden

Aufbauend auf dieser Kurzdarstellung des Konzeptes (Zielsetzung 1) wird nun die Umsetzung kompetenzorientierter adaptiver E-Klausuren an einem Anwendungsbeispiel illustriert (Zielsetzung 2). Die Implementation erfolgte in drei Phasen.

3.1 Phase 1: Kompetenzorientierte Klausuren als kriteriumsorientierter Test

Das Assessment Framework für sechs angeschlossene Klausurzyklen zu Forschungsmethoden der Erziehungswissenschaft bezog sich auf Designs, Datenerhebungen und Inhalte der einführenden Statistik für die Anwendung im Bereich Bildung und Erziehung. Für die Klausuren wurde ein Itempool generiert, der zwischen 2012 und 2017 von zunächst 49 auf 124 Items anwuchs. Die Items repräsentierten überwiegend die Stufen Wissen, Verständnis und Anwendung und wurden kontextualisiert mit Bezug zu Anwendungsbereichen der Erziehungswissenschaft konstruiert. Die Klausuren wurden wie folgt administriert (FREY, SPODEN & BORN, 2020): In jeder Klausur mit einer Prüfungszeit von 90 Minuten wurden 35 und 40 Items vorgegeben, von denen mindestens 15 Items als Linkitems aus vorausgegangenen Klausuren übernommen wurden. Die Linkitems wurden genutzt, um die Klausuren, wie

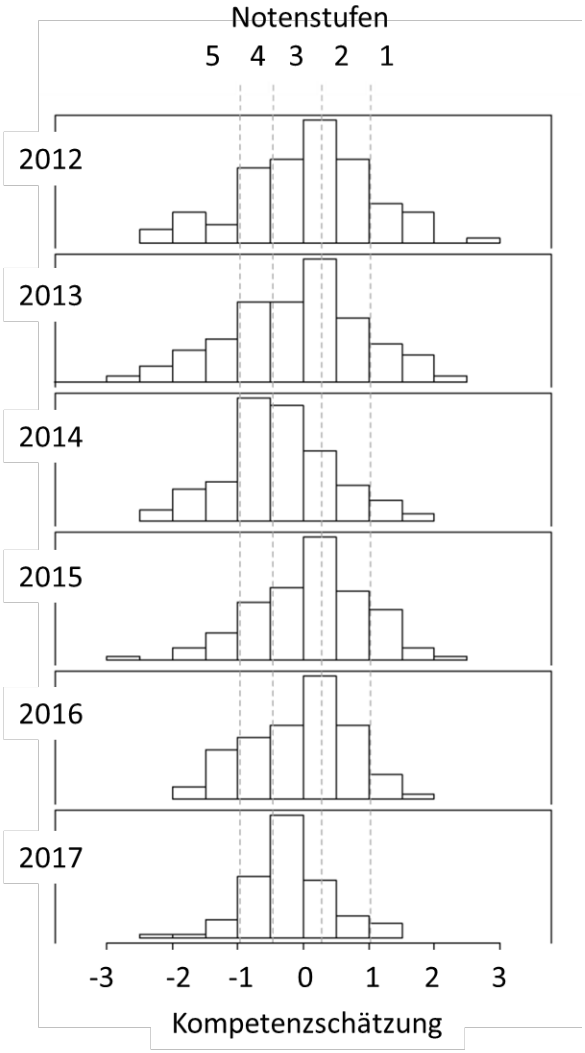

Abb. 4: Notenstufen und Kompetenzschätzungen der Klausur „Forschungsmethoden der Erziehungswissenschaft“ (2012–2017). Angepasst aus FREY, SPODEN & BORN, 2020, S. 482.

zuvor erläutert, über ein Equating-Verfahren statistisch zu verbinden. Die für diese Items geschätzten Schwierigkeiten wurden auf Item-Drift geprüft, wobei jeweils für mindestens 10 Items eine Verankerung auf die früheren Schwierigkeiten möglich war. Auf Basis eines *Common Item Nonequivalent Group*-Designs wurde so eine über sechs Jahrgänge reliable Skala etabliert. Für die Notenvergabe wurden auf der Kompetenzskala Grenzen zwischen Notenstufen anhand der Bookmark-Methoden eingezogen. Die Ergebnisse sind in Abbildung 4 ersichtlich. Durch den Vergleich der durchschnittlichen Kompetenzen der Studierendenkohorten auf einer gemeinsamen Kompetenzdimension wurden Unterschiede zwischen Studierendenkohorten offenbart, die mit Veränderungen in der Lehre in Verbindung standen. Der Kompetenzeinbruch 2014 kann auf eine Reduzierung der Anzahl begleitender Tutorien zurückgeführt werden. Die niedrigen Kompetenzausprägungen 2017 lassen sich durch eine Umstellung der Lehre in einem häufig gewählten Nebenfach auf neue, im Vergleich zur vorigen Veranstaltung nicht mehr mit Forschungsmethoden in Verbindung stehende Inhalten erklären. In der Zusammenschau zeigen die Ergebnisse, dass die Etablierung einer längsschnittlich angelegten, reliablen Kompetenzskala im Klausurbetrieb möglich ist und sich interessante Einblicke zur Erklärung beobachteter Klausurergebnisse und Schlussfolgerungen zur Anpassung der Lehre ableiten lassen.

3.2 Phase 2: Kompetenzorientierte E-Klausuren als kriteriumsorientierter computerbasierter Test

Um die individualisierte Itemauswahl beim CAT zu realisieren, müssen Charakteristika der Items inklusive der Schwierigkeit bekannt sein. Letztere muss aus empirischen Daten bei einer computerbasierten Testadministration geschätzt werden (Testkalibrierung). Unter der Bedingung kleiner Stichproben, wie sie im Hochschulwesen häufig anzutreffen sind, existieren zwei Möglichkeiten: Zum einen kann die kontinuierliche Kalibrierungsstrategie (FINK et al., 2018) genutzt werden. Zum anderen ist eine gemeinsame Kalibrierung über verschiedene Studienorte möglich, bei der oft mehr Items kalibriert werden können. Die Itemschwierigkeiten wurden hier in Vorbereitung eines adaptiven Tests zunächst auf Basis einer solchen kooperativen und computerbasierten Datenerhebung geschätzt. Der Test wurde an drei Standorten unter leicht unterschiedlichen Durchführungsbedingungen administriert: In Jena als Präsenzklausur mit 69 Studierenden der Erziehungswissenschaft; in Bo-

chum und Ulm als Online-Übungsklausur mit 27 beziehungsweise 43 Psychologiestudierenden. An jedem Standort wurden 30 Items eingesetzt, 15 Items (50 %) wurden gemeinsam an allen Standorten administriert, um die Skalierung über drei Standorte vorzunehmen. Somit existierten vier Cluster von Items (gemeinsame Linkitems sowie spezifische Items an den drei Standorten). Die Klausuren wurden mithilfe der KAT-HS-App (FINK et al., 2021) vorgegeben. Bei der Klausur in Jena wurde die Kommunikation nach außen durch Nutzung des Safe Exam Browsers (HALBHERR et al., 2016) mit eigeschränkten Nutzungsrechten unterbunden. Die Schätzung der Itemschwierigkeiten erfolgte mithilfe des Mehrfacetten-Rasch-Modells mit den Facetten Itemschwierigkeit und Universitätsort. Gemeinsame Schwierigkeiten konnten letztlich nur für sieben Linkitems bestimmt werden, bei denen keine Unterschiede in der Lösungswahrscheinlichkeit zwischen den Standorten unter Kontrolle der durchschnittlichen Leistungen der Studierendengruppen gefunden wurden, also kein *Differential Item Functioning* (DIF; z. B. OSTERLIND & EVERSON, 2009) vorlag. Mithilfe dieser 7 Linkitems war es möglich, Itemschwierigkeiten für den kompletten Itempool zwischen -1.508 und 4.058 zu schätzen. Die mittlere Kompetenz der Studierenden wurde in Jena auf 0.097, in Bochum auf 0.393 und in Ulm auf -0.489 geschätzt (negative Werte entsprechen hier besseren Kompetenzschätzungen). Diese deutlichen Unterschiede sowie der hohe Anteil von Items, die bei der standortübergreifenden Skalierung DIF aufwiesen, legen nahe, dass die kontinuierliche Item-Kalibrierung für die Vorbereitung eines adaptiven Tests zu bevorzugen ist.

3.3 Phase 3: Kompetenzorientierte E-Klausuren als kriteriumsorientierter, computerbasierter adaptiver Test

In Phase 3 wurde eine computerbasierte und teilweise adaptive E-Klausur mit 10 Testversionen à 25 Items aus fünf Item-Clustern realisiert. Der CAT-Algorithmus realisierte die adaptive Itemauswahl im Safe Exam Browser mit eigeschränkten Nutzungsrechten und der KAT-HS-App entsprechend dem Maximum Fisher Item Information Criterion. Zur Identitätsfeststellung wurden Studierendenausweis und Matrikelnummer genutzt. Die Integrität der Prüfungsleistung und die Funktionsfähigkeit des Prüfungssystems wurden anhand von Log-Daten sichergestellt. Die Prüfungs- und Log-Datei der Studierenden wurde auf einem sicheren Universitäts-

server mit einem Hashwert, der Informationen zur Prüfung in Ganzzahlen abbildet, hinterlegt.

Die Klausur wurde von $N = 84$ Erstsemesterstudierenden (87 % weiblich, 13 % männlich) mit Hauptfach Erziehungswissenschaft (Bachelor) in Jena bearbeitet, denen vorab die Möglichkeit eingeräumt wurde, das elektronische Klausursystem zu erproben. Die mittlere geschätzte Kompetenzausprägung lag bei $M = 0.03$ ($SD = 0.79$), die Reliabilität der Klausur lag bei $rel_{WLE} = .722$. Diese Ergebnisse verdeutlichen, dass kompetenzorientierte adaptive Hochschulklausuren sicher und reliabel durchführbar sind. Da frühere Studien auch negative Effekte von CAT beim emotionalen Erleben aufgezeigt haben (z. B. PITKIN & VISPOEL, 2001), beantworteten die Studierenden einige Minuten vor der Klausurdurchführung einen Fragebogen zum emotionalen Erleben, um konventionelle Klausuren zu bewerten (Pretest), und nochmals unmittelbar nach der Klausur zur Bewertung des neuartigen Klausurkonzepts (Posttest). Der Vergleich zeigte, dass sowohl positive (Pretest: $M = 2.77$, $SD = 0.59$; Posttest: $M = 2.65$, $SD = 0.69$; $t(80) = 2.21$, $p = .03$), als auch negative Emotionen (Pretest: $M = 2.31$, $SD = 0.69$; Posttest: $M = 2.10$, $SD = 0.70$; $t(81) = 4.71$, $p < .01$) bei der neu konzipierten Klausur niedriger ausfielen, also ein gewisses Maß an Indifferenz nach der kognitiv anspruchsvollen Klausur ausgedrückt wurde. Diese Ergebnisse legen nahe, dass die Vorteile einer adaptiven Klausur nicht durch Nachteile beim emotionalen Erleben der Studierenden aufgewogen werden.

4 Diskussion

Der vorliegende Beitrag war mit den Zielen verknüpft, (1) adaptive kompetenzorientierte E-Klausuren als einen Ansatz für individualisierte und gleichzeitig fair und vergleichbar konzipierte Klausuren vor- und (2) die Umsetzung in drei Entwicklungsphasen am Beispiel einer Klausur zu Forschungsmethoden der Erziehungswissenschaft darzustellen. Bei adaptiven kompetenzorientierten E-Klausuren werden etablierte Methoden aus den Bereichen Educational Measurement und Psychometrie basierend auf der IRT an die spezifischen Herausforderungen von Hochschulprüfungen (z. B. der kleineren Stichprobengröße) angepasst. Die bisherigen Implementationsergebnisse (siehe auch unten) erlauben eine optimistische Einschätzung zukünftiger Nutzungsmöglichkeiten des Gesamtkonzeptes.

Der Einsatz des hier vorgestellten Konzeptes ist dabei an Voraussetzungen gebunden. Die Verlinkung von Klausuren über die Zeit und die adaptive Administration bedürfen, wie zuvor erwähnt, eines geheim gehaltenen Itempools. Das Konzept ist somit für den Einsatz bei den während der Covid-19-Pandemie eingesetzten, online über das Internet administrierten Klausuren kaum geeignet, da hier die Gefahr des Item-Diebstahls mithilfe von Screenshots hoch ist. Neben dem denkbaren Einsatz anderer Cheating-Indikatoren sollte insbesondere die Schwierigkeit wiederholt genutzter Items regelmäßig auf Item-Drift geprüft werden, um systematisches Vorbereiten auf möglicherweise publik gewordene Items aufzudecken. Ferner sollte berücksichtigt werden, dass der effiziente Einsatz von E-Klausuren eine entsprechende Infrastruktur wie bestenfalls ein E-Testzentrum notwendig macht. Interessanterweise haben sich die Infrastrukturbedingungen in einer Befragung von Prüfenden an 74 Hochschulen in Deutschland (SPODEN et al., 2020) nicht als relevanter Prädiktor der Intention zur Nutzung von E-Klausuren herausgestellt. Dies könnte damit zusammenhängen, dass die digitale Infrastruktur mit Computerpools tatsächlich besser ausgebaut ist, als zuweilen angenommen. Schließlich sollte angemerkt werden, dass die Umsetzung anspruchsvoller Teile des Konzeptes bei vielen Prüfenden Fortbildungsmöglichkeiten in der Hochschuldidaktik (und bei der Erstellung adaptiver Prüfungen in Testsoftware gegebenenfalls „handwerklicher" Unterstützung durch die für E-Klausuren verantwortlichen Personen bedarf, die inzwischen in vielen hochschuldidaktischen Abteilungen angestellt sind). Um auf diese Herausforderung zu reagieren, wurde eine Fortbildungsveranstaltung zum Konzept erstellt, durchgeführt und evaluiert (Ergebnisse nicht publiziert). Obwohl die Anforderungen an Wissen zur IRT als anspruchsvoll eingeschätzt wurden, zeigten die Ergebnisse der Evaluation bei den Teilnehmenden deutliche Wissenszuwächse bezüglich der Konstruktion von E-Klausuren nach dem Workshop. Auch beinhalten die Ergebnisse eine positive Einschätzung der Nützlichkeit der Veranstaltung sowie eine klar geäußerte Nutzungsabsicht bezüglich des Klausurkonzeptes.

Gleichzeitig ist hervorzuheben, dass das Konzept auf alternative Prüfungsformen übertragen oder auf formative Assessments zur Lernbegleitung erweitert werden kann. Trotz der hohen Relevanz, die Modulabschlussklausuren im Prüfungswesen besitzen, haben Befragungen gezeigt, dass Studierende eine kontinuierliche Erfassung der Lernleistungen und -fortschritte als Alternative zu einer einzigen, das „Bulimielernen" provozierenden Prüfung am Semesterende präferieren (z. B. SAMBELL, McDOWELL, & BROWN, 1997). Auch wünschen sie sich individuelles

Feedback zu ihren Ergebnissen. Das vorliegende Konzept kann hierzu mit lernförderlichen Konzepten der formativen Assessments oder des E-Learnings verknüpft werden. So könnten an adaptive Kurztests im Semesterverlauf Tutor- oder Empfehlungssysteme angebunden werden, die Hilfestellungen oder individuelle Hinweise zur Weiterarbeit bereitstellen. Die Vorteile der hier vorgestellten Prinzipien sind dabei, dass aufgrund des adaptiven Modus bei kurzen Tests bereits reliable Informationen zum Lernstand auf der Berichtsmetrik der späteren Prüfungen im Sinne einer Diskrepanz zu Zielstellungen (wie erwarteten Kompetenzen an einer bestimmten Stelle im Studium) bereitgestellt werden. Auch semesterbegleitende Teilklausuren lassen sich so zeitökonomisch durchführen und die Ergebnisse auf der etablierten Berichtsmetrik verorten. Obwohl adaptive kompetenzorientierte E-Klausuren also an bestimmte Anwendungsvoraussetzungen gebunden sind, beinhalten sie doch gleichzeitig zahlreiche Möglichkeiten, um deutliche Entwicklungssprünge bei der Konstruktion von Hochschulklausuren zu erzielen.

5 Literaturverzeichnis

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347–364. https://doi.org/10.1007/bf00138871

Bloom, B. (Hrsg.) (1956). Taxonomy of Educational Objectives, the classification of educational goals – Handbook I: Cognitive Domain. New York: McKay.

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, *17*, 265–300. https://doi.org/10.1207/s15324818ame1703_3

Fink, A., Born, S., Spoden, C., & Frey, A. (2018). A continuous calibration strategy for computerized adaptive testing. *Psychological Test and Assessment Modeling*, *60*, 327–346.

Fink, A., Spoden, C., Frey, A., & Naumann, P. (2021). Kriteriumsorientiertes adaptives Testen mit der KAT-HS-App. *Diagnostica*, *67*, 110–114. https://doi.org/10.1026/0012-1924/a000268

Frey, A. (2020). Computerisiertes adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), Testtheorie und Fragebogenkonstruktion (3. Aufl., S. 501–524). Berlin: Springer.

Frey, A., & Ehmke, T. (2007). Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. *Zeitschrift für Erziehungswissenschaft, Sonderheft 8*, 169–184. https://doi.org/10.1007/978-3-531-90865-6_10

Frey, A., Spoden, C., & Born, S. (2020). Construction of psychometrically sound written university exams. *Psychological Test and Assessment Modeling*, *65*(4), 472–486.

Hachmeister, C.-D., & Grevers, J. (2019). *Im Blickpunkt: Die Vielfalt der Studiengänge 2019. Entwicklung des Studienangebotes in Deutschland zwischen 2014 und 2019*. Gütersloh: CHE.

Halbherr, T., Dittmann-Domenichini, N., Piendl, T., & Schlienger, C. (2016). Authentische, kompetenzorientierte Online-Prüfungen an der ETH Zürich. *Zeitschrift für Hochschulentwicklung*, *11*(2), 247–269.

Herzberg, P. Y., & Frey, A. (2011). Kriteriumsorientierte Diagnostik. In L. F. Hornke, M. Amelang & M. Kersting (Hrsg.), *Methoden der psychologischen Diagnostik. Enzyklopädie der Psychologie*, B/II/2 (S. 281–324). Göttingen: Hogrefe.

Kolen, M. J., & Brennan, R. L. (2014). *Test Equating, Scaling and Linking: Methods and Practices* (3rd ed.). New York, NY: Springer.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark method: Psychological perspectives. In G. J. Cizek (Hrsg.), *Setting performance standards: Concepts, methods, and perspectives* (S. 249–281). Mahwah, NJ: Erlbaum.

Osterlind, S. J., & Everson, H. T. (2009). *Differential item functioning.* Thousand Oaks, CA: Sage Publishing.

Pitkin, A. K., & Vispoel, W. P. (2001). Differences between self-adapted and computerized adaptive tests: A meta-analysis. *Journal of Educational Measurement*, *38*(3), 235–247. https://doi.org/10.1111/j.1745-3984.2001.tb01125.x

Sambell, K., McDowell, L., & Brown, S. (1997). "But is it fair?": An exploratory study of student perceptions of the consequential validity of assessment. *Studies in Educational Evaluation*, *23*, 349–371. https://doi.org/10.1016/S0191-491X(97)86215-3

Spoden, C. (2021). Ressourceneinsatz bei der Erstellung psychometrisch fundierter Hochschulklausuren. In C. Spoden & A. Frey (Hrsg.), *Psychometrisch fundierte E-Klausuren für die Hochschule* (S. 106–111). Lengerich: Pabst.

Spoden, C., Frey, A., Fink, A., & Naumann, P. (2020). Kompetenzorientierte elektronische Hochschulklausuren im Studium des Lehramts. In K. Kaspar, M. Becker-Mrotzek, S. Hofhues, J. König & D. Schmeinck (Hrsg.), *Bildung, Schule und Digitalisierung* (S. 184–189). Münster: Waxmann.

van der Linden, W. J. (Hrsg.). (2016). *Handbook of item response theory. Volume one: Models*. London: Chapman & Hall/CRC. https://doi.org/10.1201/9781315374512

Zawacki-Richter, O. (2015). Zur Mediennutzung im Studium – unter besonderer Berücksichtigung heterogener Studierender. *Zeitschrift für Erziehungswissenschaft*, *18*, 527–549. https://doi.org/10.1007/s11618-015-0618-6

Autor*in*en

Prof. Dr. Christian SPODEN || Hochschule Emden/Leer, Fachbereich Wirtschaft || Constantiaplatz 4, D-26723 Emden

www.hs-emden-leer.de

christian.spoden@hs-emden-leer.de

Aron FINK || Goethe-Universität Frankfurt, Arbeitsbereich Pädagogische Psychologie || Theodor-W.-Adorno-Platz 6, D-60629 Frankfurt am Main

https://www.psychologie.uni-frankfurt.de/73548927/Aron_Fink

a.fink@psych.uni-frankfurt.de

Prof. Dr. Andreas FREY || Goethe-Universität Frankfurt, Arbeitsbereich Pädagogische Psychologie || Theodor-W.-Adorno-Platz 6, D-60629 Frankfurt am Main

https://www.psychologie.uni-frankfurt.de/73548872/Prof__Dr__Andreas_Frey

frey@psych.uni-frankfurt.de

Hanna KÖHLER || Friedrich-Schiller-Universität Jena, Institut für Psychologie || Am Steiger 3, D-07743 Jena

https://www.klipsy.uni-jena.de/team/hanna-koehler-m-sc

hanna.koehler@uni-jena.de

Patrick NAUMANN || Goethe-Universität Frankfurt, Arbeitsbereich Pädagogische Psychologie || Theodor-W.-Adorno-Platz 6, D-60629 Frankfurt am Main

https://www.psychologie.uni-frankfurt.de/77209044/Patrick_Naumann

naumann@psych.uni-frankfurt.de