Issue 13
2020

Kompetenzorientierte individualisierte Hochschulklausuren und deren prüfungsrechtliche Einordnung

Andreas Frey
frey@psych.uni-frankfurt.de

Goethe-Universität Frankfurt
Theodor-W.-Adorno-Platz 6
60323 Frankfurt

University of Oslo
Centre for Educational Measurement (CEMO)
Postboks 1161 Blindern
0318 Oslo, Norway

Christian Spoden

Deutsches Institut für Erwachsenenbildung, Leibniz-Zentrum für Lebenslanges Lernen e.V.
Bonn

Aron Fink
a.fink@psych.uni-frankfurt.de

Goethe-Universität Frankfurt
Theodor-W.-Adorno-Platz 6
60323 Frankfurt

Sebastian Born

Universitätsklinikum Jena

urn:nbn:de:0009-5-51197

Zusammenfassung

Schriftliche Klausuren sind für Studierende von hoher Relevanz, da ein erfolgreicher Hochschulabschluss und der damit verbundene Zugang zum Arbeitsmarkt von den Klausurergebnissen abhängen. Dieser hohen Relevanz werden typische Hochschulklausuren aus wissenschaftlicher Sicht jedoch nicht gerecht. Es werden vier zentrale Herausforderungen bei der Weiterentwicklung von Hochschulklausuren beschrieben und spezifiziert, wie diese auf Basis einer zielgerichteten neuartigen Kombination von Methoden, Qualitätsstandards und Ansätze der Psychometrie und der pädagogisch-psychologischen Diagnostik zu überwinden sind. Das vorgestellte Konzept zur Konstruktion, Administration und Auswertung von Hochschulklausuren ist hochschul- und fächerübergreifend nutzbar. Es wird prüfungsrechtlich umfassend eingeordnet und Rahmenbedingungen seiner erfolgreichen Implementation im hochschulischen Regelbetrieb diskutiert.

Stichwörter: e-learning, Messen, Testen, digitale Hochschulbildung, Kompetenzdiagnostik, Item Response Theory

Abstract

Written exams in higher education are highly relevant for students, because success in their studies and access to the labour market depend on the exam results. But typical written exams do not do justice to this high importance from an academic point of view. This paper describes four major challenges when it comes to the development of high-quality exams in higher education and specifies how to meet them on the basis of a novel combination of methods, quality standards and approaches stemming from psychometrics and educational and psychological measurement. The suggested concept for the construction, administration and scoring of exams in higher education can be used across universities and across disciplines. Its feasibility is classified according to examination law and the general conditions for its successful implementation in the regular operation of higher education institutions are discussed.

Keywords: e-learning, measurement, testing, digitalization in higher education, competence measurement, item response theory

---

Spätestens seit Inkrafttreten der Bologna-Reform (Nickel, 2011) wird von den Hochschulen in der Europäischen Union gefordert, die Konzeption von Studiengängen von den Lernzielen her zu planen und die Module auf den Erwerb von Kompetenzen (z. B. Klieme & Leutner, 2006 für eine psychologisch orientierte Kompetenzdefinition) auszurichten. Damit werden konkrete Erwartungen formuliert, was eine Studentin oder ein Student auf Basis fachlicher Kenntnisse in definierten Anwendungsbereichen wissen und können soll, um eine Lehrveranstaltung, ein Modul oder einen Studiengang mit Erfolg abzuschließen. Mit der Kompetenzorientierung ist verbunden, dass die Studentinnen und Studenten mit dem Besuch verschiedener Lehrveranstaltungen vor allem wirksame Handlungsweisen zur Lösung von relevanten Problemen in berufs- und wissenschaftsbezogenen Kontexten erwerben sollen, um somit einen erfolgreichen Übergang ins Berufsleben zu ermöglichen.

Das Ausmaß, mit dem Studierende die kompetenzorientierten Lernziele ihres Hochschulstudiums erreichen, wird am Abschneiden bei Prüfungen festgemacht. Im zeitlichen Verlauf einer Lehrveranstaltung bilden Prüfungen zwar das letzte Element, konzeptionell gesehen sollte allerdings schon in der Planungsphase mitgedacht werden, welche Lerngelegenheiten zum Kompetenzerwerb bereitgestellt werden und wie der Kompetenzerwerb letztendlich geprüft wird. Diesen Ansatz konkretisierend empfiehlt Biggs (1996) in seinem Constructive Alignment-Ansatz die zentralen Elemente der Lehr-, Lern- und Prüfungsgestaltung stringent auf die intendierten Lernziele zu beziehen. Er entwickelte diesen Ansatz vor dem Hintergrund, dass Hochschullehrende sich bei der Planung und Durchführung von Lehrveranstaltungen typischerweise an anderen Aspekten orientieren als Studentinnen und Studenten. Während für die Lehrenden vor allem die Konzeption und Umsetzung wirkungsvoller Lerngelegenheiten für die Studentinnen und Studenten im Fokus stehen, orientieren sich die Lernhandlungen der Lernenden häufig an den abschließenden Prüfungen einer Lehrveranstaltung zur Erlangung von Leistungsnachweisen (Schaper & Hilkenmeier, 2013). Dozentinnen und Dozenten können sich dies zu Nutze machen, indem sie bei der Planung ihrer Lehrveranstaltung Lernziele, Methoden zur Bereitstellung lernzielkonformer Lerngelegenheiten und die abschließende Prüfung des Lernerfolgs aufeinander abstimmen. Wenn dies konsequent geschieht und in den Lehrveranstaltungen transparent kommuniziert wird, ergibt sich die Möglichkeit das Lernverhalten von Studentinnen und Studenten sehr wirkungsvoll zu leiten, zu unterstützen

und den Lernerfolg zielsicher zu überprüfen. Entsprechend sollte eine gute Passung von Lernzielen, Lehre und Prüfungen bei der Gestaltung hochschulischer Lehre stets angestrebt werden (Biggs & Tang, 2011).

Für die Studierenden sind insbesondere die abschließenden Prüfungen als Manifestation des individuellen Studienerfolgs von hoher persönlicher Relevanz. Ohne erfolgreich absolvierte Prüfungen kann ein Studienabschluss nicht erlangt werden. Vor allem schriftliche Prüfungen in Form von Klausuren sind vor dem Hintergrund wachsender Studierendenzahlen in den letzten Jahren zunehmend in den Vordergrund gerückt. Es ist deshalb wichtig, dass diese Klausuren auch zielsicher, präzise und fair Kompetenzen prüfen. Nimmt man den aktuellen Forschungsstand in der Psychometrie und der pädagogisch-psychologischen Diagnostik zum Maßstab, so ist dies aktuell jedoch typischerweise nicht der Fall. Vor allem digitale Technologien eröffnen hier noch nicht systematisch genutzte Möglichkeiten, Hochschulklausuren gemäß dem aktuellen wissenschaftlichen Kenntnisstand zu optimieren.

Im Folgenden wird skizziert, welche zentralen Herausforderungen sich bei der Konstruktion wissenschaftlich fundierter Hochschulklausuren zur Überprüfung von Kompetenzständen ergeben und wie diesen effektiv begegnet werden kann. Es handelt sich dabei um die Kurzfassung eines umfangreichen neuartigen Konzeptes für Hochschulklausuren (Spoden & Frey, in Vorb.; www.kat-hs.uni-frankfurt.de). Darauf aufbauend wird erörtert, wie das vorgeschlagene Klausurkonzepts prüfungsrechtlich zu bewerten ist und Rahmenbedingungen für dessen Nutzung im Regelbetrieb von Hochschulen diskutiert.

Herausforderungen bei der Konstruktion wissenschaftlich fundierter Hochschulklausuren

Betrachtet man typische Hochschulklausuren vor dem Hintergrund aktueller wissenschaftlichen Qualitätsstandards der Psychometrie und der pädagogisch-psychologischen Diagnostik, zeigen sich vier zentrale Herausforderungen:

1. Die im Modulkatalog verankerten Lernziele werden durch die genutzten Klausuraufgaben oft nicht angemessen operationalisiert. Häufig erfolgt eine Fokussierung auf Wissensabfragen, obgleich die Modulbeschreibung auch auf höhere kognitive Prozesse abzielt. Darüber hinaus wird der jeweilige Inhaltsbereich oft nicht systematisch durch die Klausuraufgaben abgedeckt. Dies kann dazu führen, dass einige Inhalte der Lehrveranstaltung gar nicht und andere sehr intensiv geprüft werden. Die Interpretation von Klausurergebnissen ermöglicht dann keine angemessenen Schlüsse über das Erreichen der kompetenzorientierten Lernziele. Hier ist also ein Mangel an Validität (z. B. Hartig, Frey & Jude, 2020) zu verzeichnen. Zur Überwindung dieses Problems ist es wichtig, kontextspezifisch und in engem Bezug zu den tatsächlichen Lernzielen alle relevanten Inhalte und kognitiven Anforderungen mit geeigneten Klausuraufgaben zu prüfen.
2. Bei der Bewertung von Klausuren wird häufig der prozentuale Anteil korrekt beantworteter Aufgaben genutzt. So werden beispielsweise Klausuren teilweise dann als bestanden bewertet, wenn mehr als 50 % der Aufgaben korrekt beantwortet wurden. Dieses Vorgehen ist problematisch. Es wäre nur dann gerechtfertigt, wenn (a) die Aufgaben die kompetenzorientierten Lernziele

angemessen operationalisieren und (b) alle Aufgaben einer Klausur in gleichem Umfang und gegeneinander austauschbar den gleichen Anteil an der Lernzielerreichung repräsentieren. Wie bei Herausforderung 1 geschildert ist die Voraussetzung (a) bei typischen Hochschulklausuren regelmäßig als nicht gegeben anzusehen. Die Voraussetzung (b) ist nur bei eindeutig zählbaren und äquivalenter Einzelleistungen (z. B. Anzahl produzierter identischer Werkstücke; Sand schippen usw.) angemessen. Bei der Messung von Kompetenzen ist es aufgrund der Binnenstruktur des Messgegenstandes indes nicht möglich, Aufgaben zu generieren, die jede für sich den gesamten zu prüfenden Kompetenzbereich abdecken und gegenseitig austauschbar sind. Vielmehr ist die Vorgehensweise der Kompetenzdiagnostik (Frey & Hartig, 2019) zu nutzen, um Klausuren im Sinne kriteriumsorientierter Tests (Herzberg & Frey, 2011) zu konstruieren. Mit diesen können dann die benötigten Aussagen bezüglich des Erreichens von kompetenzorientierten Lernzielen valide abgeleitet werden.

3. Klausuren sind zumeist zwischen verschiedenen Testzeitpunkten (z. B. verschiedenen Semestern) nicht statistisch miteinander verbunden. Zusammen mit den aus Sicherheitsgründen zumindest teilweise über Testzeitpunkte variierenden Aufgaben hat dies zur Folge, dass bei jeder Kohorte der Maßstab zum Erreichen der einzelnen Notenstufen neu festzusetzen ist. Derartige Klausuren sind über die Jahre hinweg nicht vergleichbar, da Klausurergebnisse bei konstanter Kompetenz zwischen Testzeitpunkten variieren können. Beispielsweise könnte eine Studentin mit sehr hoher Ausprägung der qua Modulbeschreibung geforderten Kompetenzen in einem Semester die Note 1,0 bekommen und bei der Klausur im kommenden Semester die Note 1,7 - und dies bei exakt gleicher individueller Kompetenzausprägung. Diese Probleme können vermieden werden, indem mit Modellen der Item Response Theory (IRT; z. B. van der Linden, 2016) Klausuren über Testzeitpunkte verbunden werden. Damit können individuelle Klausurergebnisse stets auf derselben Metrik verortet und über Klausurzeitpunkte hinweg verglichen werden.

4. Übliche Klausuren messen im mittleren Bereich der Kompetenzverteilung am präzisesten, an ihren Rändern hingegen fällt die Messpräzision deutlich niedrigerer aus (Dolan & Burling, 2017). Das bedeutet, dass die Genauigkeit der Bewertung der Antworten bei einer Klausur (z. B. die Notengebung), und damit die Beurteilung inwieweit die im Modulkatalog verankerten Kompetenzen erlangt wurden, in Abhängigkeit der individuellen Kompetenzausprägung unterschiedlich hoch ist. Vor allem für Studentinnen und Studenten mit sehr hoher beziehungsweise sehr niedriger Kompetenz liefern die Klausuren also häufig kein verlässliches Resultat. Eine Angleichung der Messpräzision über den gesamten Merkmalsbereich kann mit computerisiertem adaptiven Testen (CAT; z. B. Frey, 2020) erreicht werden.

Vermutlich existieren zahlreiche Gründe dafür, dass diesen Herausforderungen bislang kaum adäquat begegnet wurde. Anzunehmen ist, dass (a) ein relativ hoher Initialaufwand, wenn alle diese Elemente von einer einzelnen Lehrperson oder universitären Arbeitsgruppe umgesetzt werden, (b) ein nicht vorhandenes Know-how bezüglich Testkonstruktion und IRT-Skalierung beziehungsweise fehlende Unterstützung durch Expertinnen und Experten und/oder (c) ein hoher Bedarf an Testzeit, wenn Kompetenzen kriteriumsorientiert mit

herkömmlichen papierbasierten Klausuren gemessen werden, zur gegenwärtigen, eher spärlichen Verwendung wissenschaftlich fundierter Methoden bei Hochschulklausuren beigetragen haben.

Lösungsansätze

Für die dargelegten Herausforderungen steht eine Reihe wissenschaftlich fundierter Methoden zur Verfügung, die sich in der Praxis in Bereichen außerhalb der Hochschulbildung etabliert haben und regelmäßig genutzt werden. Im Rahmen von Hochschulklausuren finden diese Methoden jedoch bisher kaum Anwendung. Nachfolgend wird skizziert, wie diese Methoden zur Überwindung der genannten Herausforderungen nutzbar gemacht werden können. Die Detailliertere Ausführungen und die formale Dokumentation der einzelnen Methoden sind dem Buch von Spoden und Fink (in Vorb.) zu entnehmen.

Der ersten Herausforderung der angemessenen Abbildung der Lernziele einer Lehrveranstaltung in einer Klausur kann durch die Definition des Messgegenstandes als Kombination von kognitivem Prozess (z. B. unter Verwendung der Taxonomie von Bloom, Engelhart, Furst, Hill & Krathwohl, 1956) und Inhaltsbereich des interessierenden Themenfeldes und der Operationalisierung anhand geeigneter Klausuraufgaben entlang etablierter Richtlinien zur Aufgabenkonstruktion (z. B. Haladyna & Rodriguez, 2013; sowie Rodriguez & Albano, 2017 mit direktem Bezug auf Hochschulklausuren) begegnet werden.

Für das Bewältigen der zweiten Herausforderung der Generierung von Klausurergebnissen, die valide Aussagen über die individuelle Lernzielerreichung ermöglichen, kann kriteriumsorientiertes Testen eingesetzt werden. Hierbei ist für die Interpretation des Klausurergebnisses nur das Ausmaß, mit dem die kompetenzorientierten Lernziele erreicht werden, relevant und nicht etwa das durchschnittliche Kompetenzniveau der jeweiligen Studierendenkohorte oder die Schwierigkeit der vorgegebenen Klausur. Standardsetzungsverfahren (bspw. die häufig genutzte Bookmark-Methode in einer vereinfachten Form; Lewis, Mitzel, Mercado & Schulz, 2012), welche der Festlegung von Grenzwerten zwischen Bewertungskategorien (Notenstufen oder bestanden/nicht bestanden) dienen, ermöglichen es die Klausurergebnisse direkt auf das Ausmaß der Lernzielerreichung zu beziehen.

Um der dritten Herausforderung eines über verschiedene Klausurzeitpunkte konstanten Bewertungsmaßstabs gerecht zu werden, können Equating-Methoden (z. B. Kolen & Brennan, 2014) für den statistisch angemessenen Umgang mit Schwierigkeitsunterschieden zwischen Klausuren genutzt werden. Hierbei ist es nützlich bei zwei benachbarten Klausurzeitpunkten eine Gruppe identischer Aufgaben (sog. Ankeraufgaben) einzusetzen. Unter Hinzuziehung eines IRT-Modells kann dann sichergestellt werden, dass beide Klausuren das identische Merkmal (=Ausmaß der Lernzielerreichung) auf der identischen Skala messen. Damit können die Ergebnisse verschiedener Klausurzeitpunkte direkt miteinander verglichen werden. Prüflinge müssen somit immer über die gleichen Kompetenzen verfügen, um den einzelnen Bewertungskategorien zugeordnet zu werden.

Die letzte Herausforderung einer vergleichbaren Messgenauigkeit für alle Studentinnen und Studenten unabhängig von ihrem Kompetenzniveau ist erstrebenswert, da ansonsten das Problem besteht, dass die Zuverlässigkeit der aus der Klausurbearbeitung abgeleiteten Bewertung (Noten oder bestanden/nicht-bestanden) je nach Kompetenzniveau unterschiedlich ausfällt. Eine Antwort auf diese Herausforderung liefert CAT, das eine Angleichung der Messgenauigkeit über das gesamte Kompetenzspektrum ermöglicht. CAT wird bislang bei Hochschulklausuren noch nicht im Regelbetrieb eingesetzt. Aus diesem Grund wird es nachfolgend genauer beschrieben und die durch die Methode realisierte Individualisierung prüfungsrechtlich diskutiert.

Computerisiertes adaptives Testen

Bei den meisten Testverfahren wird allen Testpersonen eine vorab zusammengestellte Menge von Aufgaben in einer bestimmten Reihenfolge präsentiert. Beim CAT werden einer Testperson nur solche Aufgaben vorgelegt, bei denen aufgrund des bislang gezeigten Antwortverhaltens der Testperson davon ausgegangen werden kann, dass sie besonders viel diagnostische Information über die interessierende Merkmalsausprägung liefern. In der Praxis läuft dies verkürzt darauf hinaus, dass die Schwierigkeit der vorgegebenen Aufgaben an die Leistungsfähigkeit der getesteten Personen angepasst wird.

Das Grundprinzip des CAT ist in Abbildung 1 als Flussdiagramm dargestellt. Der Test startet mit der Auswahl und Vorgabe einer ersten Aufgabe (Schritt 1). Die Antwort der Testperson auf diese Aufgabe wird vom Computer registriert und mit der richtigen Lösung verglichen (Schritt 2). Damit die Antworten vom Computer bewertet werden können, sind bei einem adaptiven Test automatisch kodierbare Aufgabenformate zu verwenden. Mit dem Ergebnis der Prüfung liegt eine erste empirisch gewonnene Information vor, mit der die individuelle Merkmalsausprägung der Testperson grob eingeschätzt werden kann (Schritt 3). Danach wird geprüft, ob ein oder mehrere vorab fixierte Kriterien zur Beendigung des Tests erfüllt sind (z. B. maximale Testzeit erreicht, maximale Anzahl von Aufgaben bearbeitet usw.).

Nach der ersten Aufgabe wird dies in der Regel noch nicht der Fall sein, so dass erneut die Schritte 1 bis 3 durchlaufen werden. Dies geschieht so lange, bis alle Kriterien zur Beendigung des Tests erfüllt sind. Im Rahmen der adaptiven Aufgabenauswahl können neben der zu erzielenden diagnostischen Information weitere nicht-statistische Einschränkungen Berücksichtigung finden. Diese bestehen beispielsweise darin, dass einzelne Aufgaben nicht zu vielen Personen vorgegeben werden oder dass für jede getestete Person die gleichmäßige Abdeckung aller Bereiche des Messgegenstandes (z. B. verschiedene Inhaltsbereiche) sichergestellt wird. Nachdem alle Abbruchkriterien erfüllt sind, wird der computerisierte adaptive Test mit der endgültigen Schätzung der individuellen Merkmalsausprägung beendet (Schritt 4).

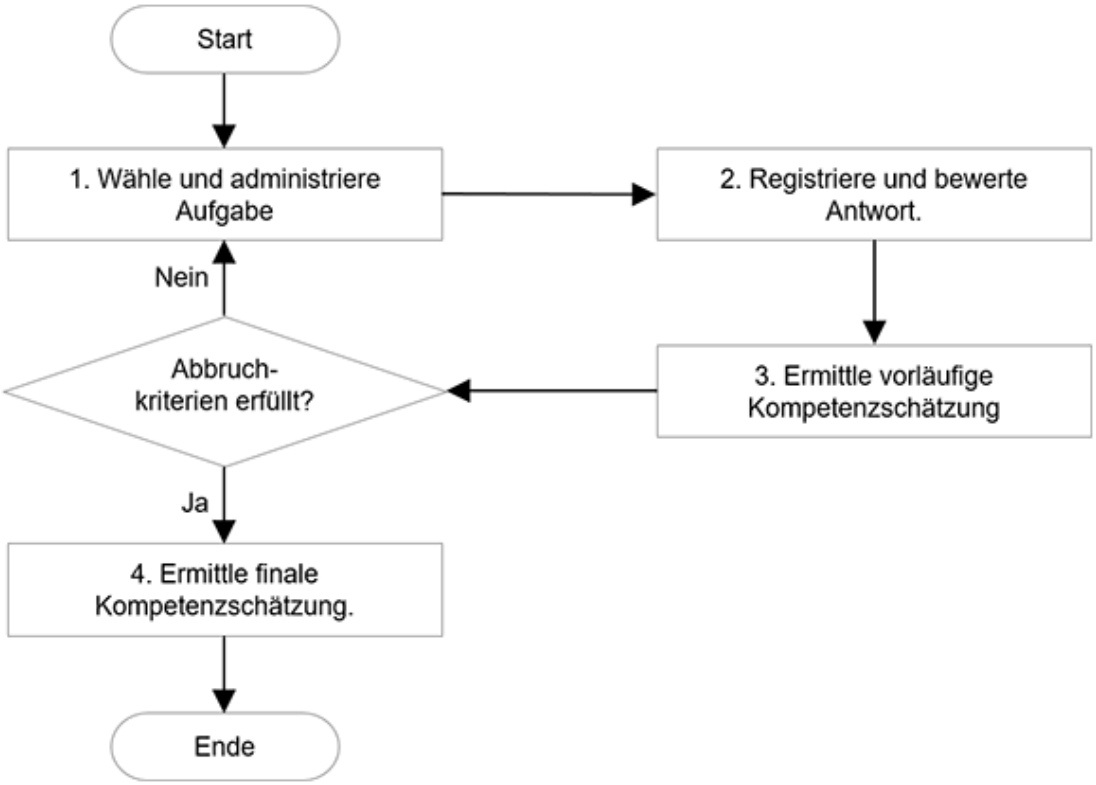

Abbildung 1. Flussdiagramm zum typischen Ablauf von computerisierten adaptiven Tests (modifiziert aus Frey, 2020).

Da Testpersonen mit hoher Merkmalsausprägung beim CAT schwierigere Aufgaben zu bearbeiten haben als Testpersonen mit niedrigerer Merkmalsausprägung, kann ein fairer interindividueller Vergleich nicht anhand der Anzahl korrekter Antworten erfolgen. Vielmehr ist es notwendig, bei der Bestimmung des Testergebnisses die Schwierigkeit der bearbeiteten Aufgaben zu berücksichtigen: Zehn korrekte Antworten auf Aufgaben mit hoher Schwierigkeit sprechen für eine höhere Merkmalsausprägung als zehn korrekte Antworten auf sehr leichte Aufgaben. Zur Berücksichtigung der Schwierigkeiten der bearbeiteten Aufgaben werden beim CAT IRT-Modelle eingesetzt. Da bei IRT-Modellen die Aufgabencharakteristika inklusive deren Schwierigkeit unabhängig von der individuellen Merkmalsausprägung (z. B. Kompetenzausprägung de/r/s Student/in) bestimmt werden, können Testergebnisse für interindividuelle Vergleiche genutzt werden, auch wenn die Testpersonen jeweils andere Aufgaben bearbeitet haben. CAT ist weltweit in zahlreichen Anwendungsbereichen etabliert. Diese umfassen auch individualdiagnostische Tests mit hoher Bedeutung für die getesteten Individuen.

Die Verwendung von CAT ermöglicht psychometrisch belastbare, individualisierte Hochschulklausuren. Durch die Computernutzung und insbesondere die Individualisierung von Klausuren werden aber auch prüfungsrechtliche Fragen tangiert, die im folgenden Abschnitt besprochen werden.

Individualisierte Klausuren im Rahmen des Prüfungsrechts

Aufgrund der vorstehenden Ausführungen kann konstatiert werden, dass die derzeit an Hochschulen in Deutschland üblichen Klausuren den Zweck, des Erreichens kompetenzorientierter Lernziele zu bestimmen, bestenfalls begrenzt erfüllen und diese Klausuren hinter etablierten, erprobten und wissenschaftlich abgesicherten Methoden deutlich zurückbleiben. Hinzu kommt, dass mittelfristig davon auszugehen ist, dass der Großteil der Hochschulklausuren in digitalisierter Form zu bearbeiten sein wird. In einzelnen deutschen Hochschulen werden digitale Klausuren bereits in großem Stil eingesetzt. Die nun in der Breite anstehende Umstellung hin zu digitalen Klausuren bietet die Möglichkeit, den aus fachlicher Sicht nicht angemessenen Stand von Hochschulklausuren auf eine wissenschaftlich abgesicherte Basis zu stellen. Dies erfordert auch einige prüfungsrechtliche Erwägungen, da bestehende rechtliche Grundsätze mit den technischen Entwicklungen bei modernen, computerbasierten Testverfahren in Einklang zu bringen sind. Betrachtet man die Anforderungen an elektronische Prüfungssysteme für Deutschland, welche sich aus dem Prüfungsrecht ergeben, so ist neben datenschutzrechtlichen Bestimmungen und den Forderungen nach Authentizität und Integrität der Prüfungsleistung, die Gewährleistung von Chancengleichheit für Prüflinge (abgeleitet aus Art. 3. Abs. 1 des Grundgesetzes der Bundesrepublik Deutschland) ein essentieller Punkt. Durch die Nutzung elektronischer Prüfungssysteme ergibt sich auch die Möglichkeit des Einsatzes adaptiver und damit individualisierter Klausuren. Neben didaktischen (Prüfen auf individuell angemessenem Kompetenzniveau) und psychometrischen Vorteilen (Erhöhung der Messeffizienz, Angleichung der Messpräzision über Merkmalsbereich) gelten individualisierte Klausuren als sehr sicher, da eine individualisierte Aufgabenzusammenstellung die Möglichkeit des Abschreibens von Aufgabenlösungen reduziert beziehungsweise ein Erschleichen von Vorteilen aufgrund vorheriger Kenntnis von Prüfungsaufgaben (z. B. aus Gedächtnisprotokollen von Studentinnen und Studenten höherer Semester) deutlich erschwert. Indes haben bei individualisierten Prüfungen die einzelnen Prüflinge nicht mehr die gleiche Klausur vorliegen. Dies könnte man als Verletzung der Chancengleichheit ansehen. Hierbei ist jedoch zu beachten, dass der Grundsatz der Chancengleichheit keine absolute Gleichbehandlung gewährt, sondern nur vergleichbare äußere Modalitäten, um gewährleisten zu können, dass die gemessenen individuellen Kompetenzausprägungen der Prüflinge von äußeren Bedingungen unbeeinflusst sind (Niehues, Fischer & Jeremias, 2018). So liegt nach aktueller Rechtsauffassung dann eine Ungleichbehandlung vor, wenn das Prüfungssystem beziehungsweise die Prüferin oder der Prüfer wahllos Fragen aus einem vorhandenen Fragenkatalog auswählt, ohne beispielsweise deren Themenzugehörigkeit, Schwierigkeitsgrad oder Bearbeitungsdauer zu berücksichtigen. Solch willkürlicher Auswahl kann man durch die Nutzung standardisierter Fragenkataloge, die hinsichtlich dieser Eigenschaften zusammengestellt werden, oder die Kontrolle nicht-statistischer Einschränkungen im Rahmen des CAT entgegenwirken und damit Vergleichbarkeit herstellen. Individualisierte Klausuren sind daher nicht per se ein Verstoß gegen den Gleichheitsgrundsatz.

Der Gleichbehandlungsgrundsatz im Prüfungsrecht führt zudem nicht dazu, dass die individuellen Kompetenzausprägungen der Teilnehmenden im Prüfungsverlauf nicht berücksichtigt werden dürfen. Wie bereits dargelegt, ist der individuelle Kompetenzerwerb das zentrale Ziel eines jeden Hochschulstudiums. Die Feststellung dieses individuellen

Kompetenzniveaus sollte dadurch ermöglicht werden, dass in Prüfungen auf das individuelle Kompetenzniveau eingegangen wird (Schaper & Hilkenmeyer, 2013). In mündlichen Prüfungen ist dies gängige Praxis. Hier ist die Möglichkeit der Prüferin oder des Prüfers auf das individuelle Kompetenzniveau des Prüflings einzugehen integraler Bestandteil der Prüfpraxis und wird im Allgemeinen nicht als Verletzung der Gleichbehandlung angesehen. Eine analoge Sichtweise kann man auch auf eine adaptive Klausur anwenden. Gleichheit wird hier dadurch gewährleistet, dass alle Teilnehmenden durch dasselbe System und nach denselben Regeln bewertet werden. Den individuellen Verlauf der Prüfung bestimmt für jeden Prüfling derselbe Algorithmus. Die Rahmenbedingungen der adaptiven Klausur sind für alle Prüflinge die gleichen, das System passt sich nur an das individuelle Kompetenzniveau an, ähnlich wie die Prüferin oder der Prüfer in der mündlichen Prüfung. Im Vergleich zu einer Prüferin oder einem Prüfer, bei welchen nicht auszuschließen ist, dass beispielsweise subjektive Wahrnehmung, inkonsistente Gewichtung einzelner Fragen oder die unbewusste Beeinflussung durch Äußerlichkeiten der geprüften Person oder andere Kontextfaktoren zu einer Bewertungsverzerrung führen können (z. B. Schaper & Hilkenmeyer, 2013; Niehues, Fischer & Jeremias, 2018), ist außerdem die hohe Objektivität des Algorithmus hervorzuheben. Wichtig dabei ist, dass der Schwierigkeitsgrad der bei in einem CAT-System genutzten Aufgaben vorher bekannt ist. Warum die Kenntnis der Aufgabenschwierigkeiten von zentraler Wichtigkeit ist, wird nachfolgend am Beispiel der Massenbestimmung eines Objektes mit einer Balkenwaage verdeutlicht.

Angenommen man hätte eine Balkenwaage mit 100 Gewichtsstücken, deren Masse unbekannt ist und es bestünde das Ziel Messobjekte hinsichtlich ihrer Masse zu vergleichen. Ohne das Wissen über die genaue Masse der einzelnen Gewichtsstücke und somit auch über die Rangreihe der Gewichtsstücke, könnte man die Masse eines spezifischen Messobjektes einzeln mit jedem der 100 Gewichtsstücke vergleichen und zählen wie viele Gewichtsstücke leichter als das Messobjekt sind. Die Anzahl der leichteren Gewichtsstücke wäre das Ergebnis der Messung. Die beschriebene Regel „Zähle die Anzahl der leichteren Gewichtsstücke“ würde in diesem Fall eine Skala etablieren, welche einen Wertebereich von 0 bis 100 aufweist. Auf dieser Skala könnte man jedes beliebige Messobjekt lokalisieren und verschiedene Messobjekte in Bezug auf ihre Masse miteinander vergleichen. Voraussetzung für den Vergleich wäre, dass alle Objekte „gleich behandelt“ werden. Das heißt, dass alle Objekte mit dem gleichen Satz an Gewichtsstücken gemessen werden müssen.

Die zweite Art der Messung geht davon aus, dass die Massen der 100 Gewichtsstücke bekannt sind. Unter dieser Voraussetzung ist es möglich die 100 Gewichtsstücke entsprechend ihrer Masse in eine Rangreihe zu bringen. Um nun die Masse eines Messobjektes zu bestimmen, könnte man zuerst ein Gewichtsstück mit mittlerer Masse auf die Balkenwaage legen. Erweist sich das Messobjekt im Vergleich zum Gewichtsstück als leichter beziehungsweise schwerer, kann man für den weiteren Verlauf der Messung alle schwereren beziehungsweise alle leichteren Gewichtsstücke ausschließen. Ein Vergleich mit diesen würde keine zusätzliche Information über die Masse des Messobjektes liefern. Stattdessen würde man von den verbleibenden Gewichtsstücken eines wählen (also im ersten Fall ein leichteres und im zweiten Fall ein schwereres) und dieses erneut mit dem Messobjekt vergleichen. Diesen Vorgang würde man wiederholen, solange es noch Gewichtsstücke gibt, die zusätzliche Information über die Masse des Messobjektes liefern können. Am Ende der Messung ist es möglich, das Messobjekt in der Rangreihe der 100 Gewichtsstücke zu verorten. Das Messergebnis wäre entweder die Masse eines

Gewichtsstückes, bei welchem die Balkenwaage im Gleichgewicht steht oder der Mittelwert der Massen aus den zwei direkt benachbarten Gewichtsstücken der Rangreihe, in welcher das Messobjekt verortet wurde. Auf Basis dieses Vorgehens würde man ebenfalls jedes beliebige Messobjekt auf einer etablierten Skala verorten. Die Grundvoraussetzung dieser Art der Messung ist, dass die Masse aller Gewichtsstücke vor Beginn der Messung bekannt ist. Im Gegensatz zur ersten beschriebenen Art der Messung, ist die Zweite deutlich effizienter, da nicht jedes Messobjekt mit allen 100 Gewichtsstücken verglichen werden muss. Bei Gleichverteilung der Gewichtsstücke über die Skala resultiert bei jedem Messvorgang, dass die Hälfte der noch nicht vorgegebenen Gewichtsstücke keine zusätzliche Information bringt und nicht mehr vorgegeben werden muss. Man braucht also nur in etwa halb so viele Messungen durchzuführen, um zum bestmöglichen Resultat zu kommen. Es wird deutlich, dass keine „Gleichbehandlung“ im Sinne der Verwendung des Gesamtsatzes an Gewichtsstücken für eine Messung nach dem zweiten Ansatz notwendig ist. Vielmehr erfolgt die Gleichbehandlung dadurch, dass alle Messgegenstände mit Gewichtsstücken verglichen werden, deren Masse (und damit deren Verortung auf der Gewichtsskala) bekannt ist.

Obwohl die Messung von Kompetenzen etwas anspruchsvoller ist, da sich diese beispielsweise zu verschiedenen Inhaltsbereichen und unterschiedlichen kognitiven Anforderungen facettieren, lassen sich die oben beschriebenen Arten der Messung direkt auf Hochschulklausuren übertragen. So stellt die erste Art der Messung das momentan übliche Vorgehen bei Hochschulklausuren dar, bei welchen allen Studentinnen und Studenten die gleichen Aufgaben vorgelegt werden. Die Anzahl der richtig gelösten Aufgaben (Punkte) entspräche dem Ergebnis der Klausur. Hierbei ist in der Regel nicht bekannt wie schwer die einzelnen Aufgaben sind, so dass eine Ungleichbehandlung über verschiedene Klausurzeitpunkte (mit verschiedenen Aufgaben) resultieren kann. Die zweite Art der Messung entspricht dem Vorgehen eines computerisierten adaptiven Tests, bei dem die Auswahl der vorgegebenen Aufgaben in Abhängigkeit der im Testverlauf gegebenen Antworten erfolgt. Bei der ersten Art der Messung kann Gleichbehandlung (an einem Testzeitpunkt) durch die Vorgabe identischer Aufgaben sichergestellt werden. Bei der zweiten Art der Messung kann Gleichbehandlung (an einem Testzeitpunkt und über mehrere hinweg) dadurch sichergestellt werden, dass die Schwierigkeiten der Aufgaben bekannt sind. Dies ist bei computerisierten adaptiven Tests standardmäßig der Fall, bei dem die Schwierigkeiten vorab in sogenannten Kalibrierungsstudien (Thompson & Weiss, 2011) oder im laufenden Betrieb durch Nutzung von Online-Kalibrierungsmethoden (z. B. Fink, Born, Spoden & Frey, 2018) bestimmt werden. Da menschliches Antwortverhalten nicht in gleichem Maße deterministisch ist wie der Vergleich von Gewichten (d. h. ein Prüfling kann eine individuell leichte Aufgabe auch einmal nicht lösen), wird beim CAT der Zusammenhang zwischen individueller Merkmalsausprägung und Antwortverhalten in Wahrscheinlichkeiten ausgedrückt. Durch die Verwendung probabilistischer IRT-Modelle wird dem Sachverhalt Rechnung getragen, dass Menschen sich zwar systematisch aber nicht vollständig deterministisch verhalten.

Neben dem Zustandekommen der Bewertung stellt auch die Zuverlässigkeit der aus der Klausurbearbeitung abgeleiteten Bewertung (in Form von Noten oder als bestanden/nicht bestanden) eine wesentliche Voraussetzung für einen Vergleich von Prüfungsergebnissen dar. Typische Klausuren, bei denen alle Personen unabhängig von ihrer Kompetenz Aufgaben vergleichbarer Schwierigkeit vorgelegt bekommen, liefern vor allem für Studentinnen und Studenten mit sehr hoher beziehungsweise sehr niedriger Kompetenz häufig kein zuverlässiges Resultat (Dolan & Burling, 2017). Gerade bei den genannten

Studierendengruppen sind aber die vergebenen Bewertungen oft besonders relevant. Am unteren Ende des Kompetenzspektrums geht es darum, ob eine Klausur noch als bestanden gewertet werden kann, was sich nachhaltig auf das Vorankommen im Studium auswirken kann. Am oberen Ende des Kompetenzspektrums könnte von der Note die Zulassung zu einem zulassungsbeschränkten weiterführenden Studiengang, der Erhalt eines Stipendiums oder die Aufnahme in ein Promotionsprogramm abhängen. Aufgrund der geringen Messgenauigkeit an den Rändern der Kompetenzverteilung sind aber genau solche Entscheidungen mit einer besonders hohen Fehlerwahrscheinlichkeit verbunden. CAT ist auch in dieser Hinsicht nützlich, da die oben beschriebene Vorgehensweise zu einer Angleichung der Messgenauigkeit über das gesamte Kompetenzspektrum genutzt werden kann (z. B. Frey & Ehmke, 2007).

Folgt man der dargelegten Argumentation, so sind individualisierte CAT-Klausuren in Analogie zu mündlichen Prüfungen ein gangbarer Weg. Sie stellen keine Ungleichbehandlung dar, weil die Aufgabenschwierigkeiten vorab bekannt sind und damit alle Personen auf der gleichen Metrik verortet werden können. Die ineffiziente bisherige Lösung Gleichbehandlung durch die Vorgabe der identischen Klausur an alle Prüflinge zu gewährleisten ist als Anachronismus einzustufen, der zu Zeiten etabliert wurde, als die technischen Möglichkeiten für angemessenere Vorgehensweisen noch nicht gegeben waren.

Da es sich bei dem vorgestellten Klausurkonzept um eine neue Entwicklung handelt, wurden adaptive Hochschulklausuren noch nicht vor Gericht behandelt. Eine abschließende rechtliche Einschätzung steht somit noch aus. Aufgrund der besprochenen Argumente sind wir aber zuversichtlich, dass adaptive Hochschulklausuren als prüfungsrechtlich zulässig eingestuft werden.

Fazit

Den vier zentralen Herausforderungen bei der Gestaltung von Hochschulklausuren kann mit etablierten und vielfach erprobten Methoden aus Psychometrie und pädagogisch-psychologischer Diagnostik begegnet werden. Das hier vorgestellte Klausurkonzept widerspricht keinen prüfungsrechtlichen Grundsätzen. Wissenschaftliche Basis und prüfungsrechtliche Einschätzung sind jedoch noch kein Garant dafür, dass das Klausurkonzept auch breite Anwendung findet. Immerhin erhebt es den weitreichenden Anspruch hochschul- und fächerübergreifender Einsetzbarkeit. Um eine Anwendung in der Breite realistisch zu machen sind neben einem leistungsfähigen Konzept (a) eine geeignete Software und (b) Implementationsforschung notwendig.

Software

Eine in der freien Statistik-Software R (R Core Team, 2020) geschriebene Software mit dem Namen KAT-HS-App liegt bereits in einer ersten Version vor (https://kat-hs.uni-frankfurt.de/materialien/software/). Mit dieser Software können kompetenzorientierte adaptive oder nicht-adaptive Hochschulklausuren zusammengestellt, vorgegeben und IRT-basiert ausgewertet werden. Die KAT-HS-App verfügt über eine grafische Benutzeroberfläche und ist so gestaltet, dass sie auch von Lehrenden ohne jegliche

Kenntnisse in empirischer Forschung und Statistik genutzt werden kann. Die Software wurde bereits mit Erfolg zur Abnahme von Klausuren im regulären Studienbetrieb eingesetzt. Es ist geplant, sie nach abschließender Fertigstellung unter der offenen GNU gpl-3-Lizenz auf GitHub zur Verfügung zu stellen. Mit der Bereitstellung der Software für nicht-kommerzielle Zwecke ist die Erwartung verbunden, dass sich ein universitätsübergreifendes Netz von Entwicklerinnen und Entwicklern sowie Nutzerinnen und Nutzern des KAT-HS-Klausurkonzepts etabliert. Eine solche dezentrale Struktur unter Einbezug engagierter Kolleginnen und Kollegen erachten wir bei dem sich schnell wandelnden Bereich technologiebasierter Ansätze in der Hochschulbildung als zeitgemäß und deutlich zielführender als die vertragliche Bindung an einen kommerziellen Softwareanbieter. Hierdurch kann eine kontinuierliche wissenschaftsbasierte Weiterentwicklung von Hochschulklausuren als Grundlage für eine nachhaltig qualitativ hochwertige und anpassungsfähige Lösung von Hochschulen für Hochschulen geschaffen werden.

Bedingungsfaktoren erfolgreicher Implementationen

Weiterhin ist passende Implementationsforschung (z. B. Petermann, 2014; Schrader, Hasselhorn, Hetfleisch & Goeze, 2020) notwendig, um zu eruieren, unter welchen Rahmenbedingungen eine Implementation im hochschulischen Regelbetrieb gelingen wird. Entsprechende Studien zum vorgestellten Klausurkonzept wurden bereits durchgeführt. Bei mehreren zum Teil deutschlandweiten, Studien wurden die Rahmenbedingungen für die erfolgreiche Implementation des KAT-HS-Konzepts aus Sicht von Studierenden (Esmaeili-Bijarsari, Frey, Spoden, Born & Fink, 2019), Lehrenden (Fink, Spoden, Born & Frey, 2019), IT-Verantwortlichen und Studiendekaninnen und Studiendekanen (Fink, Siegwart, Hericks, Spoden & Frey, 2020) kürzlich umfassend auf Basis von Technologieakzeptanzmodellen systematisch untersucht. Die zugehörigen Publikationen werden voraussichtlich zur Drucklegung des vorliegenden Artikels verfügbar sein, so dass mögliche Stolpersteine bei der Implementation kompetenzorientierter adaptiver Klausuren zielgerichtet aus dem Weg geräumt werden können.

Entsprechend möchten wir mit diesem Artikel einerseits Kolleginnen und Kollegen für die Defizite aktueller Hochschulklausuren sensibilisieren sowie zur Nutzung und bestenfalls Mitarbeit bei der Optimierung zeitgemäßer Hochschulklausuren zu stimulieren. Mit dem vorgestellten Klausurkonzept liegt hierfür ein niedrigschwelliges, wissenschaftlich fundiertes und direkt in der Breite einsetzbares Konzept inklusive offener Software von Hochschule für Hochschule vor.

Literatur

Biggs, J.: Enhancing teaching through constructive alignment. In: Higher Education, 32, 1996, pp. 347-364. https://doi.org/10.1007/BF00138871 (last check 2020-10-28)

Biggs, J.; Tang, C.: Teaching for quality learning at university. Open University Press, Buckingham, UK, 2011.

Bloom, B. S.; Engelhart, M. D.; Furst, E. J.; Hill, W. H.; Krathwohl, D. R.: Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain. Longmans, Green, New York, Toronto, 1956.

Dolan, R. P.; Burling, K. S.: Computer-based testing in higher education. In: Secolsky, C.; Denison, D. B. (Eds.): Handbook on measurement, assessment, and evaluation in higher education, pp. 370-384. Routledge, New York, NY, 2017.

Esmaeili-Bijarsari, S.; Frey, A.; Spoden, C.; Born, S.; Fink, A.: Emotionale Effekte von Itemreview in Hochschulklausuren. Vortrag auf der 7. Tagung der Gesellschaft für empirische Bildungsforschung (GEBF), Köln, 2019, Februar.

Fink, A.; Born, S.; Spoden, C.; Frey, A.: A continuous calibration strategy for computerized adaptive testing. In: Psychological Test and Assessment Modeling, 60, 2018, pp. 327– 346.

Fink, A.; Spoden, C.; Born, S.; Frey, A.: Testing an explanatory model for the intention to use e-exams by the university teaching staff. Paper presented at the 18th Conference of the European Association for Research on Learning and Instruction, Aachen, 2019, August.

Fink, A.; Siegwart, M.; Hericks, N.; Spoden, C.; Frey, A.: Hinderungsgründe von IT-Verantwortlichen und Studiendekaninnen und Studiendekanen bei der Nutzung des KAT-HS-Konzept (Projektbericht). Frankfurt a. M.: Goethe-Universität Frankfurt, 2020.

Frey A.: Computerisiertes adaptives Testen. In: Moosbrugger, H.; Kelava, A. (Hrsg.): Testtheorie und Fragebogenkonstruktion, 3. Aufl., pp. 501-524. Springer, 2020. https://doi.org/10.1007/978-3-662-61532-4_20 (last check 2020-10-28)

Frey, A.; Ehmke, T.: Hypothetischer Einsatz adaptiven Testens bei der Überprüfung von Bildungsstandards. In: Zeitschrift für Erziehungswissenschaft, Sonderheft 8, 2007, pp. 169-184. https://doi.org/10.1007/978-3-531-90865-6_10 (last check 2020-10-28)

Frey, A.; Hartig, J.: Kompetenzdiagnostik. In: Harring, M.; Rohlfs, M., C.; Gläser-Zikuda, M. (Hrsg.): Handbuch Schulpädagogik, pp. 849-858, Waxmann, Münster, New York, 2019.

Grundgesetz für die Bundesrepublik Deutschland (GG) in der im Bundesgesetzblatt Teil III, Gliederungsnummer 100-1, veröffentlichten bereinigten Fassung, zuletzt geändert durch Artikel 1 des Gesetzes vom 23. Dezember 2014 (BGBl. I S. 2438).

Haladyna, T. M.; Rodriguez, M. C.: Developing and validating test items. Taylor & Francis, New York, NY, 2013. https://doi.org/10.4324/9780203850381 (last check 2020-10-28)

Hartig, J.; Frey, A.; Jude, N.: Validität von Testwertinterpretationen. In: Moosbrugger, H.; Kelava, A. (Hrsg.): Testtheorie und Fragebogenkonstruktion. 3. Auflage, pp. 529-545. Springer, Berlin, Heidelberg, 2020. http://doi-org-443.webvpn.fjmu.edu.cn/10.1007/978-3-662-61532-4_21 (last check 2020-10-28)

Herzberg, P. Y.; Frey, A.: Kriteriumsorientierte Diagnostik. In: Hornke, L. F.; Amelang, M.; Kersting, M. (Hrsg.): Enzyklopädie der Psychologie: Methoden der Psychologische Diagnostik: Serie2/ Bd. 2, pp. 281-324. Hogrefe, Göttingen, 2011.

Klieme, E.; Leutner, D.: Kompetenzmodelle zur Erfassung individueller Lernergebnisse und zur Bilanzierung von Bildungsprozessen. Beschreibung eines neu eingerichteten Schwerpunktprogramms der DFG. In: Zeitschrift für Pädagogik, 52, 2006, pp. 876-903.

Kolen, M. J.; Brennan, R. L.: Test equating, scaling, and linking: Methods and practices, 3rd ed. New York, 2014. NY: Springer. https://doi.org/10.1007/978-1-4939-0317-7 (last check 2020-10-28)

Lewis, D. M.; Mitzel, H. C.; Mercado, L. R.; Schulz, E. M.: The bookmark standard setting procedure. In: Cizek, G. J. (Ed.): Setting performance standards: Foundations, methods, and innovations, pp. 225–254. New York, NY, Routledge, 2012.

Nickel, S. (Hrsg.): Der Bologna-Prozess aus Sicht der Hochschulforschung. Analysen, und Impulse für die Praxis. CHE gemeinnütziges Cent-rum für Hochschulentwicklung, Gütersloh, 2011. https://www.che.de/wp-con-tent/uploads/upload/CHE_AP_148_Bologna_Prozess_aus_Sicht_der_Hochschulforschung.pdf (last check 2020-10-28)

Niehues, N., Fischer, E. & Jeremias, C.: Prüfungsrecht. 7. Auflage, Beck, München, 2018.

Petermann, F.: Implementationsforschung: Grundbegriffe und Konzepte. In: Psychologische Rundschau, 65, 2014, pp. 122-128. https://doi.org/10.1026/0033-3042/a000214 (last check 2020-10-28)

Rodriguez, M.; Albano, A.: The college instructor's guide to writing test items. Measuring student learning. Routledge, New York, NY, 2017. https://doi.org/10.4324/9781315714776 (last check 2020-10-28)

Schaper, N.; Hilkenmeier, R.: Umsetzungshilfen für kompetenzorientiertes Prüfen. Fachgutachten für die Hochschulrektorenkonferenz. HRK, Bonn, 2013.

Schrader, J.; Hasselhorn, M.; Hetfleisch, P.; Goeze, A.: Stichwortbeitrag Implementationsforschung: Wie Wissenschaft zu Verbesserungen im Bildungssystem beitragen kann. In: Zeitschrift für Erziehungswissenschaft, 23, 2020, pp. 9–59. https://doi.org/10.1007/s11618-020-00927-z (last check 2020-10-28)

Spoden, C.; Frey, A. (Hrsg.) (in Vorb.): Psychometrisch fundierte E-Klausuren für die Hochschule. Manuskript in Vorbereitung.

Thompson, N. A.; Weiss, D. J.: A framework for the development of computerized adaptive tests. In: Practical Assessment, Research & Evaluation, 16, 2011, 1. http://assess.com/docs/v16n1.pdf (last check 2020-10-28)

van der Linden, W. J. (Ed.): Handbook of item response theory. Volume one: Models. Chapman & Hall/CRC, Boca Raton, 2016. https://doi.org/10.1201/9781315374512 (last check 2020-10-28)