Studentische Bewertung und Präferenz von Hochschulprüfungen mit Aufgaben im offenen und geschlossenen Antwortformat

Marlit A. Lindner¹, Sonja M. Mayntz¹ und Johannes Schult²

¹Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN), Kiel, Deutschland

²Landesinstitut für Schulentwicklung, Stuttgart, Deutschland

Zusammenfassung: Multiple-Choice-Aufgaben (MCA) werden gegenüber Freitextaufgaben (Constructed-Response-Aufgaben [CRA]) in Hochschulprüfungen zunehmend eingesetzt. Diese Studie betrachtet formatspezifische Einstellungen und Präferenzen von Studierenden, die relevant für selbstreguliertes Lernverhalten sein könnten (vgl. *Assessment Drives Learning*). Dreihundertfünfzig Studierende bewerteten MCA und CRA auf verschiedenen Dimensionen separat, aber mit identischen Frageformulierungen. Wie angenommen finden sich signifikante ($p \leq .001$) Bewertungsunterschiede: Die Erwartung eines (geringeren) Lernaufwandes ($d = 1.22$), die Erfolgserwartung von Testbearbeitungsstrategien ($d = 0.70$) sowie die wahrgenommene Bewertungsobjektivität ($d = 1.47$) und Fairness ($d = 0.27$) fielen höher für MCA gegenüber CRA aus. CRA schnitten dagegen hinsichtlich des wahrgenommenen Potenzials Leistung zu zeigen deutlich stärker ab ($d = -2.53$). Allgemein zeigten sich eher individuelle Präferenzen zugunsten eines der Formate, die teilweise durch die Formatbewertungen erklärbar waren. Implikationen für die Prüfungspraxis an Hochschulen werden diskutiert.

Schlüsselwörter: Multiple-Choice-Aufgaben, Testaufgaben, Aufgabenformate, Prüfungen an Hochschulen, selbstreguliertes Lernen

Students' appraisal of open and closed response formats in university exams

Abstract: Compared to constructed response (CR) items, multiple-choice (MC) items play an increasing role in university exams. This study investigates format-specific expectations and preferences of students which may influence their self-regulated learning behavior (*assessment drives learning*). Three-hundred-fifty students rated MC and CR exams on multiple dimensions. As hypothesized, there were significant ($p \leq .001$) differences: Students expected a lower workload ($d = 1.22$), higher success of test-wiseness strategies ($d = 0.70$), a higher objectivity of test results ($d = 1.47$) and more fairness ($d = 0.27$) in MC exams, whereas they perceived a greater potential to showcase their knowledge in CR exams ($d = -2.53$). We found no overarching global preference, but individual preferences for each format; these could be partially explained by format-specific expectations. Implications for higher education exams are discussed.

Keywords: Multiple-choice items, test items, test item formats, higher education exams, self-regulated learning

Prüfungen sind wie die Lehre selbst schon immer unverzichtbarer Bestandteil des Hochschulalltags. Einerseits dienen Prüfungen im Einklang mit dem Prinzip des *Constructive Alignments* (Biggs, 1996) der Diagnostik beruflich erforderlicher Kompetenzen (vgl. Huber, 2008). Andererseits sollen Studierende durch Prüfungen zu angemessenem Lernverhalten angeregt werden. Dies ist besonders bedeutsam, da Lehrende auf das selbstregulierte, eigenständige Lernen an Hochschulen (z.B. Landmann, Perels, Otto, Schnick-Vollmer & Schmitz, 2015) nur indirekt Ein-

fluss haben, nämlich durch die vorab vermittelten Erwartungen an die Inhalte und die Schwierigkeit summativer Prüfungen (vgl. *Assessment Drives Learning*; vgl. z.B. Thiede, 1996). Beide Funktionen bedingen hohe Anforderungen an die Qualität gestellter Prüfungsaufgaben.

Infolge der Bologna-Reform und der damit verbundenen Einführung von Bachelor-/Master-Studiengängen hat sich die Prüfungskultur in den letzten Jahren auch an deutschen Hochschulen systematisch gewandelt (vgl. z.B. Huber, 2008; Müller & Schmidt, 2009). Durch die studi-

enbegleitende Kreditierung der Leistungen von Studierenden in allen Leistungsmodulen jeden Semesters ist ein deutlich erhöhtes Prüfungsaufkommen zu verzeichnen (Winkel, 2010). Dies hat substanziellen Einfluss auf das Arbeitspensum der Lehrenden bei der Prüfungsabnahme, die somit an Prüfungsformen mit ökonomischen Vorteilen bei der Leistungsermittlung interessiert sind. Schriftliche Prüfungen sind daher gegenüber mündlichen Prüfungen attraktiver geworden und scheinbar ist zudem das Interesse an geschlossenen Mehrfach-Wahl-Aufgaben (Multiple-Choice-Aufgaben [MCA]) gegenüber Freitextaufgaben mit offenem Antwortformat (Constructed-Response-Aufgaben [CRA]) gestiegen. Dies geht wohl vorrangig auf die hohe Effizienz von MCA bei der Auswertung und Rückmeldung von Ergebnissen zurück. Gleichwohl der Einsatz von MCA vor allem ökonomische Vorteile bietet, lassen sich auf Seiten von Lehrenden und Studierenden auch Nachteile gegenüber CRA antizipieren (vgl. z. B. Lindner, Strobel & Köller, 2015).

In diesem Spannungsfeld konzentriert sich die vorliegende Studie auf die Perspektive der Studierenden. Es wird untersucht, inwiefern offene (CR) und geschlossene (MC) Aufgabenformate unterschiedlich bewertet werden und ob sich verschiedene, aus der Literatur ableitbare Annahmen in einer aktuellen deutschen Hochschulstichprobe bestätigen. Wissen über derartige Zusammenhänge ist relevant, da sich spezifische Erwartungen an Prüfungsanforderungen als bedeutsamer Faktor für das (selbstregulierte) Lernverhalten von Studierenden erwiesen haben (z. B. Kellas & Butterfield, 1971; Scouller, 1998; Thiede, 1996).

Offene und geschlossene Aufgabenformate

Das geschlossene MC-Format zeichnet sich durch die Vorgabe eines Aufgabenstamms mit mehreren Antwortoptionen aus, von denen im einfachsten Fall *eine* richtige Lösung unter mehreren Distraktoren identifiziert werden muss (Single-Choice). Allerdings lassen sich viele weitere komplexere Varianten von MCA unterscheiden (vgl. z. B. Lindner et al., 2015; Waugh & Gronlund, 2013). Dagegen erfordern offene CRA die eigenständige Konstruktion einer schriftlichen Antwort bezogen auf eine Frage oder ein Problem im Aufgabenstamm. Die Formatunterschiede beziehen sich also vor allem auf die unterschiedlichen Handlungsoptionen bei der Beantwortung der Fragen (d. h. *Antwortauswahl* vs. *Antwortkonstruktion*). Damit geht einher, dass Leistungen in MCA gegenüber CRA nicht durch die Handschrift oder den verbalen Ausdruck verzerrt werden können und MCA mehr diagnostische Information pro Zeiteinheit liefern (Wan & Henly, 2012). Aufgrund der Standardisierung und darauf basierenden Vorteilen gelten Objektivität und Reliabilität von MCA gegenüber CRA als

klar überlegen (z. B. Lukhele, Thissen & Wainer, 1994). Es gibt jedoch eine lange Debatte, inwiefern offene und geschlossene Formate gleichermaßen zur Erfassung von Wissen und Kompetenzen höherer Ordnung geeignet sind (vgl. z. B. Kubinger, 2014). Zuletzt deutet sich aber ein zunehmender Konsens an, der MCA unter Voraussetzung einer sorgfältigen Konstruktion das Potenzial zuschreibt, auch höhere Lernziele geeignet zu erfassen, sofern es sich nicht um kreativ-schöpferische Leistungen handelt (Haladyna & Rodriguez, 2013; Lindner et al., 2015; Lukas, Melzer & Much, 2017; Martinez, 1999; Rodriguez, 2003; Schult & Sparfeldt, 2018). Allerdings sind MCA oft einfacher lösbar als CRA (Kastner & Stangl, 2011) und bleiben aufgrund einer schwierigeren Konstruktion bei mangelnden zeitlichen Ressourcen für die Testkonstruktion hinter ihren Möglichkeiten zurück und erfüllen das anzustrebende Qualitätsniveau vielfach nicht (vgl. Haladyna, Downing & Rodriguez, 2002). Zudem differenzieren MCA in Randbereichen der Leistung schlechter (Liu, Lee & Linn, 2011). CRA sind dagegen einfacher zu erstellen und werden von Lehrenden häufiger eingesetzt, um höhere Lernziele (vgl. z. B. Bloom et al., 1956) zu erfassen. Diese anekdotisch vielfach berichtete unterschiedliche Nutzungspraxis von MCA (z. B. für basale Wissensabfragen) und CRA (z. B. für analytische Aufgaben) an Hochschulen könnte vor allem bei gut vorbereiteten Studierenden die Wahrnehmung begünstigen, Gelerntes in MC-Prüfungen nicht angemessen zum Ausdruck bringen zu können, wenn das Anforderungsniveau eingesetzter MCA dies aufgrund einer Konfundierung von Format und Anspruchsniveau nicht erlaubt.

Die Darbietung von Antwortoptionen in MCA bietet zudem die Möglichkeit zu taktischem Abwägen und Raten (z. B. Haladyna, 2004). Dieses Problem besteht vor allem, wenn Fehler in der Konstruktion von MCA den einfachen logischen Ausschluss bestimmter Optionen ermöglichen (vgl. Gierl, Bulut, Guo & Zhang, 2017; Haladyna et al., 2002). Dies erhöht systematisch die Wahrscheinlichkeit, die richtige Lösung auch ohne substanzielles Wissen zu identifizieren, und verbessert Prüfungsergebnisse in unangemessener Art und Weise (vgl. z. B. Sparfeldt, Kimmel, Löwenkamp, Steingraber & Rost, 2012). Metakognitive Strategien Studierender zur Nutzung von Konstruktionsfehlern in MCA werden unter dem Begriff *Testwiseness* zusammengefasst (vgl. Dodeen, 2008). Gleichmaßen gibt es auch für CRA Bearbeitungsstrategien, die Studierenden helfen können, ihre Prüfungsergebnisse zu optimieren (z. B. Waugh & Gronlund, 2013). Doch trotz einer weit zurückreichenden Etablierung des Testwiseness-Begriffs von Millman, Bishop und Ebel (1965) ist bisher kaum bekannt, ob Studierende sich solche Strategien vorsätzlich aneignen und ob ihr antizipierter Erfolg die Bewertung der Aufgabenformate beeinflusst.

Einstellungen gegenüber offenen und geschlossenen Prüfungsformaten

Basierend auf den in der Praxis typischerweise unterschiedlichen Anforderungen von MCA und CRA ist es naheliegend, dass sich stereotype Erwartungen bezogen auf die Prüfungsformate ausbilden. So schreibt Kubinger (2009; S.367 – 368) beispielsweise:

„Informellen Befragungen an Studierenden zufolge, ziehen diese bei universitären Prüfungen dort das Multiple-Choice-Format vor, wo es um auswendig zu lernende Wissensinhalte geht, die für die weitere Ausbildung und den angestrebten Beruf als wenig relevant eingeschätzt werden; sie ziehen demgegenüber das freie Antwortformat dort vor, wo es um Verständnis des vermittelten Stoffs geht und die Befürchtung besteht, das eigene Können im Multiple-Choice-Format nicht adäquat vermitteln zu können.“

Bereits derart anekdotische Evidenz indiziert, dass Studierende einschlägige Einstellungen gegenüber Prüfungsformaten haben, wenngleich diese im deutschen Sprachraum bisher nicht systematisch erfasst und analysiert wurden.

Die Wahrnehmung der Prüfungsanforderungen und die damit verknüpfte Erwartungshaltung – beispielsweise an die Schwierigkeit der Aufgaben – kann das Lernverhalten Studierender systematisch beeinflussen (z. B. McCoubrie, 2004; McDaniel, Blischak & Challis, 1994; Scouller, 1998; Scouller & Prosser, 1994; Thiede, 1996). In diesem Zusammenhang kommt der Erwartung einer guten Prüfungsleistung auch bei nur geringem Lernaufwand eine große Bedeutung zu, da davon auszugehen ist, dass sich eine derartige Einstellung (z. B. mit Blick auf ein bestimmtes Prüfungsformat) in ein oberflächlicheres Lernverhalten übersetzt (McCoubrie, 2004; Thiede, 1996).

Frühere Studien liefern bereits erste Informationen zu subjektiven Einstellungen, die zu einem besseren Verständnis potenzieller Wirkungen der Aufgabenformate auf das Verhalten Lernender beitragen. Konkret wurden empirische Vergleiche der Bewertung offener und geschlossener Aufgabenformate hinsichtlich verschiedener Dimensionen auf Seiten von Schülerinnen und Schülern sowie von Studierenden angestellt (vgl. z. B. Scouller, 1998; Struyven, Dochy, & Janssens, 2005; Zeidner, 1987). Insbesondere zeigte Zeidner (1987) mit zwei schulischen Stichproben zunächst mit (a) einpoligen Likert-Skalen [absolute Bewertung] und anschließend mit (b) bipolaren Ratingskalen [relative Bewertung], dass MCA unter anderem tendenziell als leichter, klarer, weniger komplex, weniger angestaussend, erfolgsversprechender und angenehmer gegenüber CRA bewertet werden. Die Befragten gaben jedoch auch an, in CRA mehr Wissen zum Ausdruck bringen zu können

als in MCA. Hinsichtlich der Fairness zeigte sich ein heterogenes Bild: In einer ersten Teilstudie ($N = 176$) zeigte sich, dass MCA gegenüber CRA fairer wahrgenommen wurden, dieser Befund konnte aber in einer zweiten Teilstudie ($N = 101$) nicht signifikant repliziert werden.

Auch Ratings hinsichtlich der Valenz der Formate im Rahmen einer Präferenzentscheidung geben einen Einblick in die globale Bewertung von MCA und CRA (vgl. z. B. Struyven et al., 2005; Zeidner, 1987). Unter Präferenz versteht man im psychologischen Sinne eine subjektive Abwägung der relativen Nützlichkeit bzw. Vorliebe einer Entität, die sich aus der Bewertung von verschiedenen Facetten dieser Entität ergibt (vgl. z. B. Pfister, Jungermann & Fischer, 2017). Besonders interessant ist es zu untersuchen, durch welche Facetten der Bewertung sich die Valenz bzw. eine mögliche Präferenz für eines der Formate erklären lässt und inwiefern diese generisch, also im Mittel über Personen homogen oder heterogen ausfällt. Birenbaum und Feldman (1998) zeigten beispielsweise, dass Schülerinnen und Schüler mit guten Lernfähigkeiten und hohem Vertrauen in ihre akademischen Fähigkeiten CRA gegenüber MCA bevorzugten.

Die wenigen einschlägigen Studien zu subjektiven Aufgabenformat-Bewertungen liegen meist viele Jahrzehnte zurück, stützen sich hauptsächlich auf ausländische (Schülerinnen und Schüler-)Stichproben und können somit nicht als repräsentativ für die aktuelle Lage an deutschen Hochschulen gelten. Die Befunde sollten somit auf den Prüfstand gestellt und systematisch erweitert werden.

Forschungsfragen

Die vorliegende Studie untersucht subjektive studentische Einschätzungen und Bewertungen offener und geschlossener Aufgabenformate im Hochschulkontext. Wir haben zwei Fragestellungen und folgende Annahmen aus der Literatur abgeleitet:

Fragestellung 1: Gibt es Unterschiede in der Bewertung von MC- und CR-Klausuren hinsichtlich des wahrgenommenen (a) Lernaufwands, (b) Potenzials Leistung zu zeigen, (c) erwarteten Erfolgs von Testwissenness-Strategien sowie der wahrgenommenen (d) Objektivität der Bewertung und (e) Fairness? Zudem untersuchen wir, inwiefern sich die wahrgenommene Fairness der Formate durch die Faktoren (a) bis (d) aufklären lässt.

Wir erwarten für MCA einen geringeren wahrgenommenen Lernaufwand, eine bessere erwartete Anwendbarkeit von Testwissenness-Strategien und eine höhere wahrgenommene Objektivität und Fairness (vgl. Zeidner, 1987). Hingegen sollte das Potenzial Leistung zu zeigen für CRA höher eingeschätzt werden. Hinsichtlich der Fairnessbewertung erwarten wir, dass die wahrgenom-

mene Objektivität der Bewertung und das eingeschätzte Potenzial Leistung zu zeigen in beiden Formaten Varianz aufklären.

Fragestellung 2: Zeigt sich über Personen hinweg eine übergreifende Präferenz oder zeigen sich individuelle Präferenzen zugunsten eines der Aufgabenformate? Lassen sich die Präferenzen durch die Bewertungsfaktoren (a) bis (d) aufklären?

Wir erwarten, dass eher individuelle statt generische Präferenzen für offene und geschlossene Aufgabenformate vorliegen, sich also über Studierende hinweg keine eindeutige Richtung der Präferenz abzeichnet. Dabei sollte vor allem der wahrgenommene Lernaufwand, das eingeschätzte Potenzial Leistung zu zeigen, die Objektivität und wahrgenommene Fairness eine bedeutsame Rolle bei der Bewertung der Aufgabenformate spielen.

Methoden

Stichprobe und Design

Unsere erste Stichprobe umfasst $n = 130$ Psychologiestudierende einer norddeutschen Universität (86.9 % weiblich; $M_{\text{Alter}} = 24.1$; $SD_{\text{Alter}} = 4.1$ Jahre) im $M = 4.3$ Fachsemester ($SD = 2.2$; Range = 2–12) und $M = 5.4$ Hochschulsesemester ($SD = 3.2$; Range = 2–15). Eine zweite Stichprobe wurde auf Wunsch der Begutachtenden im Rahmen der Manuskriptrevision zur Sicherstellung der Repräsentativität der Befunde erhoben. Für unsere Online-Erhebung (Software Unipark) wurden Psychologiestudierende in ganz Deutschland über fachspezifische Facebook-Gruppen, Lehrpersonen an Universitäten sowie Fachschaften auf die Studie aufmerksam gemacht. Der Studienlink wurde insgesamt 637-mal aufgerufen; die Beendigungsquote nach Einstieg in die Studie betrug 89.4 %. Zur Qualitätssicherung wurden nur vollständige Online-Fragebögen von Personen analysiert, die in einem Freitextfeld zum Studienfach (sinngemäß) „Psychologie“ angaben. Somit lagen weitere Datensätze von $n = 220$ Studierenden (84.5 % weiblich; $M_{\text{Alter}} = 23.5$; $SD_{\text{Alter}} = 3.8$ Jahre) aus 27 Studienstandorten in ganz Deutschland sowie einem Studienstandort in Österreich vor, die sich im $M = 5.4$ Fachsemester ($SD = 2.6$; Range = 2–16) und $M = 7.4$ Hochschulsesemester ($SD = 3.9$; Range = 2–21) fanden. Aufgrund der vergleichbaren inhaltlichen Befunde in beiden Stichproben berichten wir ausschließlich Ergebnisse der Gesamtstichprobe ($N = 350$).

Diese fachbezogen homogenen Stichproben wurden gewählt, um individuelle Unterschiede beim Vergleich der Formatbewertungen besser betrachten zu können (Minimierung externer Varianz auf Gruppenebene). Beispielsweise könnten unterschiedliche Bewertungen offener und geschlossener Formate auf stark variierende Erfahrungen mit MC- und CR-Prüfungen in verschiedenen Fachbereichen zurückgehen und die Formatwahrnehmung beeinflussen. Bei der Auswahl des Studienfaches wurde berücksichtigt, dass sowohl Prüfungen im offenen als auch im geschlossen Format zur üblichen Praxis gehören und die Studierenden daher vermutlich mit beiden Formaten wiederholt konfrontiert werden. Der Studiengang Psychologie bot sich zudem an, da es sich um ein Fach handelt, in dem sowohl naturwissenschaftlich als auch sozialwissenschaftlich geprägte Lehrveranstaltungen abgehalten werden. Wir haben ausschließlich Studierende ab dem 2. Fachsemester zur Teilnahme zugelassen, um ein Mindestmaß eigener Erfahrungen mit Hochschulprüfungen zu gewährleisten. Alle Studierenden nahmen freiwillig teil und gaben ihr Einverständnis zur vollständig anonymisierten wissenschaftlichen Nutzung ihrer Daten. Die Papier-Bleistift-Fragebögen wurden unter Anleitung einer geschulten Studienleitung bearbeitet. Die Online-Studie erlaubte eine räumlich und zeitlich unabhängige Durchführung, wobei ein Überspringen von Fragen zur Formatbewertung nicht möglich war.

Instrumente¹

Um eine vergleichende Einschätzung von MC- und CR-Prüfungsformaten vornehmen zu können, entwickelten wir Items, die mit einem formatspezifischen Aufgabenstamm beginnen („Klausuren im offenen / geschlossenen Aufgabenformat [MC / CR] ...“) und dann jeweils verschiedene Aspekte durch sprachlich parallele Aussagesätze für MC- und CR-Prüfungen auf räumlich separierten Skalen abfragen. So erfassten wir sieben Bewertungsdimensionen für das MC-Format und das CR-Format getrennt voneinander, aber inhaltlich vergleichbar. Alle Skalen und Formulierungen finden sich in Tabelle 1. Die Items wurden jeweils auf einer vierstufigen Likert-Skala (von 1 = *trifft gar nicht zu* bis 4 = *trifft voll und ganz zu*) bewertet. Abschließend erfragten wir die allgemeine Präferenzrichtung (1 Item; „Alles in allem bevorzuge ich Prüfungsaufgaben im ... MC-Format / CR-Format“) in Form einer zweipoligen, siebenstufigen Skala (von –3 = *äußerst* [MC] über 0 = *keine Präferenz* bis +3 = *äußerst* [CR]; Abbildung 1). Als Kontextvariablen wurden *Geschlecht*, *Alter*, *Fach-* und

¹ Weitere Fragebögen (z. B. zu spezifischen Testaufgabenvarianten, Persönlichkeitsfaktoren, Prüfungsangst, Lern- und Testverhalten im Studium) wurden der Stichprobe 1 vorgelegt; teils aus Platzgründen und teils auf Wunsch im Rahmen der Begutachtung beschreiben wir hier nur die für die Fragestellungen relevanten Skalen und Ergebnisse.

Tabelle 1. Items der Bewertungsskalen mit Mittelwerten (*M*), Standardabweichungen (*SD*) und interner Konsistenz (Cronbachs α), getrennt für die Multiple-Choice (MC)- und Constructed-Response (CR)-bezogen formulierte Skala.

Allgemeiner Stamm: Klausuren im offenen [CR]/geschlossenen ¹ [MC] Aufgabenformat ...	<i>M</i> _{MC}	<i>SD</i> _{MC}	α _{MC}	<i>M</i> _{CR}	<i>SD</i> _{CR}	α _{CR}
Geringer Lernaufwand						
...halte ich für geeignet, um auch bei geringem Lernaufwand ein gutes Ergebnis zu erzielen.	2.49	0.84	.72	1.88	0.86	.72
...sind für mich mit geringem Lernaufwand verbunden.	2.21	0.80		1.49	0.62	
...sind für mich in der Vorbereitung besonders aufwendig. (R)	2.72	0.74		1.89	0.80	
...erlauben mir, beim Lernen „Mut zur Lücke“ zu haben.	2.59	0.87		1.87	0.83	
Potenzial Leistung zu zeigen						
...geben mir die Gelegenheit zu zeigen, dass ich den Lernstoff sehr gut verstanden habe.	2.01	0.86	.71	3.53	0.62	.69
...geben mir die Gelegenheit zu zeigen, dass ich mehr weiß als andere Studierende.	1.66	0.83		3.20	0.85	
...erlauben mir, mein Wissen präzise zum Ausdruck zu bringen.	1.73	0.82		3.51	0.61	
...sind ein angemessenes Prüfungsformat für wichtige Klausuren.	2.33	0.87		3.16	0.68	
Erfolg von Testwissenness-Strategien²						
Goldene Regeln zum Lösen von [CR/MC]-Aufgaben können selbst bei völliger Ahnungslosigkeit hilfreich sein.	2.22	0.86	.75	1.65	0.71	.65
Wer ein paar Tipps zum Lösen von [CR/MC]-Aufgaben beherrscht, kann in Klausuren deutlich besser abschneiden.	2.57	0.82		2.15	0.81	
Objektivität						
...werden sehr objektiv bewertet.	3.32	0.84	–	2.19	0.70	–
Fairness						
...sind fair.	2.75	0.85	–	2.54	0.72	–
Vertrautheit						
...sind mir vertraut.	3.73	0.63	.75	3.51	0.72	.71
...wurden schon in meinen Klausuren gestellt.	3.78	0.64		3.51	0.87	
...waren mir bisher unbekannt. (R)	3.92	0.38		3.89	0.42	
Valenz						
...gefallen mir gut.	2.50	0.91	.86	2.63	0.87	.83
...sollte es in meinen Prüfungen häufiger geben.	2.71	0.81		2.68	0.76	

Anmerkungen: ¹ Anker wurde je Skala nur bezogen auf eines der Formate vorgegeben; ² für die Erfassung der Testwissenness gab es keinen allgemeinen Anker; (R) = Rohwerte des Items wurden gemäß der Richtung der Gesamtskala rekodiert.

Hochschulsemester, Studienort, sowie die Abiturabschlussnote der Teilnehmenden erfasst.

Statistische Analysen

Wir berichten zunächst deskriptive Parameter zur Beschreibung der Stichproben. Da alle Aussagen zur Bewertung von MC- und CR-Aufgaben innerhalb Studierender mit parallelen Items gleichen Wortlauts formuliert waren, ziehen wir für den Vergleich der Bewertungen der Formate *t*-Tests für abhängige Stichproben heran. Die kompetitive Präferenzbewertung (MCA vs. CRA) wurde mit einem *t*-Test gegen den festen Wert 0 (= keine Präferenz) getestet.

Zur Vorhersage der wahrgenommenen Fairness und Valenz von MC- und CR-Formaten, die wir mit den Bewertungen der Formate auf anderen Dimensionen in Relation setzen, berichten wir multiple Regressionsmodelle. Die Analysen wurden mit IBM SPSS® 23 ausgeführt.

Ergebnisse

Deskriptive Analysen

Die Psychologiestudierenden hatten viel Erfahrung mit MC- und CR-Prüfungen, was sich in der hohen Ausprä-

Tabelle 2. T-Tests für verbundene Stichproben zur Bewertung von Multiple-Choice (MC)- und Constructed-Response (CR)-Aufgaben.

Fragestellung 1	M	SD	Korrelationen		Mittelwertvergleiche		
			<i>r</i>	<i>p</i> _(<i>r</i>)	<i>t</i> ₍₃₄₉₎	<i>p</i> _(<i>t</i>)	<i>d</i> _(<i>t</i>) ^b
Geringer Lernaufwand MC	2.50	0.60	-.13	.016	15.14	≤ .001	1.22
Geringer Lernaufwand CR	1.78	0.58					
Potenzial Leistung zu zeigen MC	1.93	0.62	-.29	≤ .001	-29.51	≤ .001	-2.53
Potenzial Leistung zu zeigen CR	3.35	0.50					
Testwissenness MC	2.40	0.75	.31	≤ .001	11.19	≤ .001	0.70
Testwissenness CR	1.90	0.66					
Objektivität MC	3.32	0.84	-.22	≤ .001	17.58	≤ .001	1.47
Objektivität CR	2.19	0.70					
Fairness MC	2.75	0.85	-.05	.335	3.47	.001	0.27
Fairness CR	2.54	0.72					
Fragestellung 2	M	SD	r	p _(<i>r</i>)	t ₍₃₄₉₎	p _(<i>t</i>)	d _(<i>t</i>)
Valenz MC	2.60	0.81	-.58	≤ .001	-0.67	.502	-0.06
Valenz CR	2.65	0.76					
Präferenz CR vs. MC ^a	-0.01	1.66	–	–	-0.16	.871	-0.01

Anmerkungen: ^a Einstichproben t-Test gegen den festen Wert 0 = keine Präferenz für eines der Formate (positive Werte = CR Präferenz); * $p < .05$; ** $p < .01$;

^b Effektstärkenberechnung erfolgte gemäß der Formel von Dunlap, Cortina, Vaslow & Burke (1996, S. 171) für abhängige Stichproben.

gung der Vertrautheitsratings für beide Formate im Hochschulkontext widerspiegelt ($M_{MC} = 3.81$; $SD_{MC} = 0.46$; $M_{CR} = 3.64$; $SD_{CR} = 0.55$). Die Studierenden gaben dabei eine leicht höhere Vertrautheit mit MC- gegenüber CR-Aufgaben an, $t(349) = 4.96$; $p < .001$; $d = 0.35$, was die relative Bedeutung des MC-Formats im modularisierten Studiensystem unterstreicht. Die mittlere Abiturnote ($M = 1.67$; $SD = 0.62$) indiziert, dass es sich um eine allgemein eher leistungsstarke Stichprobe handelt.

Bewertungen der Prüfungsformate (Fragestellung 1)

Hinsichtlich der Dimensionen (geringer) Lernaufwand ($p \leq .001$; $d = 1.22$), Erfolg von Testwissenness-Strategien ($p \leq .001$; $d = 0.70$), Objektivität ($p \leq .001$; $d = 1.47$) und Fairness ($p = .001$; $d = 0.27$) wurden MCA signifikant „positiver“ (mit höheren Werten) eingeschätzt als CRA (s. Tabelle 2). Mit großer Effektstärke zeigte sich dagegen ein Unterschied zugunsten von CRA in dem wahrgenommenen Potenzial Leistung zu zeigen ($p \leq .001$; $d = -2.53$). Zur Identifikation von Faktoren, die mit dem Eindruck der Formatfairness assoziiert sind, führten wir Regressionsanalysen durch (Tabelle 3). Es zeigt sich, dass vor allem das wahrgenommene Potenzial Leistung zu zeigen ($p \leq .001$) und die wahrgenommene Objektivität ($p \leq .001$) signifikant zur Varianzaufklärung der Fairnessbewertung

von MCA und CRA beitrugen. Für CRA ergab sich zudem, dass ein hoher wahrgenommener Lernaufwand ($p = .002$) und die Erfolgserwartung von Testwissenness-Strategien ($p = .035$) mit einer höher wahrgenommenen Fairness assoziiert waren. Dagegen wurden MCA tendenziell fairer bewertet, wenn ein geringer Lernaufwand für dieses Format erwartet wird ($p \leq .001$).

Präferenzen der Prüfungsformate (Fragestellung 2)

Im Mittel waren die Valenzratings für MCA und CRA ähnlich stark ausgeprägt ($p = .502$; $d = -0.06$). Eine positive Bewertung des einen Formats ging allerdings tendenziell mit einer negativen Bewertung des anderen Formats einher ($r = -.581$; $p \leq .001$). Auch hinsichtlich des kompetitiven Präferenzratings zeigt der nicht signifikante Unterschied des Skalenmittels von dem Wert „0“, dass in der Stichprobe keine allgemeine Tendenz zugunsten eines bestimmten Formats vorlag ($p = .871$; $d = -0.01$), sondern eher individuelle Präferenzen zugunsten des einen oder des anderen Formats sowie ein gewisser Anteil unentschlossener Studierender (vgl. Abbildung 1, Tabelle 2). Die Valenzbewertung von MCA und CRA konnte jeweils bedeutsam durch die Bewertung des Potenzials Leistung zu zeigen ($p \leq .001$), einem (geringeren) wahrgenommenen Lernaufwand ($p \leq .001$) sowie die wahrgenommene Fairness ($p \leq .001$) des Formats

aufgeklärt werden. Die wahrgenommene Objektivität der Bewertung sowie der antizipierte Erfolg von Testwissenness-Strategien klärten hingegen keine Varianz der Präferenzratings auf. Die Regressionskoeffizienten können Tabelle 3 entnommen werden. Die Abiturabschlussnote korrelierte nicht signifikant mit den Präferenzratings ($|r| \leq .09$).

Diskussion

Ziel der vorliegenden Studie war es, Erkenntnisse über wahrgenommene Unterschiede von offenen (CR) und geschlossenen (MC) Aufgabenformaten durch Studierende sowie ihre subjektive Präferenz der Formate in Hochschulprüfungen zu gewinnen, um Rückschlüsse auf potenziell günstige oder ungünstige Einflüsse subjektiver Einstellungen auf das Lernverhalten in Vorbereitung auf verschiedene Prüfungsformate zu ziehen.

Bewertungen der Prüfungsformate

Die erste Fragestellung zielte auf die Bewertung verschiedener Facetten offener und geschlossener Prüfungsformate ab. Dabei zeigte sich hypothesenkonform, dass MCA mit der Erwartung eines geringeren Lernaufwands verbunden waren, CRA jedoch ein höheres Potenzial Leistung zu zeigen zugesprochen wurde. Diese Befunde sind konsistent mit früheren Studien (z.B. Zeidner, 1987), was beispielsweise gegen spezifische Bologna-Reform-geprägte Format-einschätzungen spricht. Auch die vorliegenden Ergebnisse weisen in die Richtung, dass die typische Verwendung von MCA für die Erfassung leicht reproduzierbaren Wissens zu einer nachlässigeren Prüfungsvorbereitung führen könnte, da dieses Format gegenüber CRA als deutlich weniger anspruchsvoll hinsichtlich der Lernanforderungen und sehr viel weniger potent hinsichtlich der Möglichkeit Leistung zu zeigen wahrgenommen wird. Gleichzeitig schnitten MCA hinsichtlich der Erfolgserwartung von Testwissenness-Strate-

Tabelle 3. Regression zur Aufklärung der (a) wahrgenommenen Fairness und der (b) Valenz von Multiple-Choice (MC)- und Constructed-Response (CR)-Aufgaben durch verschiedene Facetten der jeweiligen Formatbewertung.

Fairness MC ($R^2 = .25$)	B	SE	β	t	p
Konstante	0.32	0.25		1.28	.203
(Geringer) Lernaufwand _{MC}	0.26	0.07	.19	3.76	<.001
Potenzial Leistung zu zeigen _{MC}	0.47	0.07	.34	7.21	<.001
Erfolg von Testwissenness-Strategien _{MC}	0.02	0.06	.02	0.41	.679
Objektivität _{MC}	0.25	0.05	.25	5.19	<.001
Fairness CR ($R^2 = .18$)	B	SE	β	t	p
Konstante	0.65	0.27		2.40	.017
(Geringer) Lernaufwand _{CR}	-0.20	0.07	-.16	-3.06	.002
Potenzial Leistung zu zeigen _{CR}	0.47	0.07	.32	6.47	<.001
Erfolg von Testwissenness-Strategien _{CR}	0.12	0.06	.11	2.12	.035
Objektivität _{CR}	0.22	0.06	.21	3.93	<.001
Valenz MC ($R^2 = .42$)	B	SE	β	t	p
Konstante	0.24	0.21		1.12	.263
(Geringer) Lernaufwand _{MC}	0.25	0.06	.19	4.25	<.001
Potenzial Leistung zu zeigen _{MC}	0.53	0.06	.40	9.04	<.001
Erfolg von Testwissenness-Strategien _{MC}	0.003	0.05	.003	0.06	.952
Objektivität _{MC}	-0.03	0.04	.03	-0.76	.450
Fairness _{MC}	0.30	0.05	.31	6.59	<.001
Valenz CR ($R^2 = .37$)	B	SE	β	t	p
Konstante	-0.64	0.25		-2.52	.012
(Geringer) Lernaufwand _{CR}	0.41	0.06	.32	6.63	<.001
Potenzial Leistung zu zeigen _{CR}	0.49	0.07	.32	6.93	<.001
Erfolg von Testwissenness-Strategien _{CR}	0.06	0.05	.05	1.09	.277
Objektivität _{CR}	0.10	0.05	.09	1.88	.061
Fairness _{CR}	0.24	0.05	.23	4.77	<.001

gien stärker gegenüber CRA ab und wurden – ebenso erwartungskonform – als objektiver hinsichtlich der Leistungsbewertung eingeschätzt. Vergleichsweise geringere (aber signifikante) Unterschiede zeigten sich bei der Wahrnehmung der Fairness zugunsten von MCA gegenüber CRA, was dafür spricht, dass das subjektive Gefühl der Fairness nicht allein durch die Standardisierung bzw. Objektivität des MC-Formats geprägt zu sein scheint. Bereits bei Zeidner (1987) fiel die Bewertung der Fairness der Formate über zwei Stichproben nicht stabil aus. Unsere Regressionsanalysen weisen vielmehr darauf hin, dass das wahrgenommene Potenzial Leistung zu zeigen den größten positiven Zusammenhang mit der Fairnessbewertung beider Formate aufweist. Zukünftige Forschungsarbeiten sollten vor dem Hintergrund der Befundlage daher weitergehend untersuchen, ob sich insbesondere die Fairnesswahrnehmung von MCA weiter verbessern ließe, wenn gezielt hochwertige, komplexe MCA in Prüfungen eingesetzt werden.

Präferenzen der Prüfungsformate

Die zweite Fragestellung bezog sich auf mögliche übergreifende oder individuell geprägte Präferenzen zugunsten von MCA oder CRA. In unserer Stichprobe konnten wir im Gegensatz zur schulischen Stichprobe von Zeidner (1987) keine Anhaltspunkte für eine generische Präferenz zugunsten eines der Formate finden. So zeigte sich beim konkurrierenden Vergleich der Formate auf einer bipolaren Skala eine symmetrische, zweigipflige Verteilung (s. Abbildung 1). Studierende bevorzugten also in ähnlicher Art und Weise jeweils eines der Formate „eher“, wohingegen nur wenige eine starke Präferenz zugunsten eines der Formate berichteten und ein geringerer Anteil sich als unentschieden („weder noch“) äußerte. Diese Befunde belegen, dass Lehrende es offenbar mit heterogenen Präferenzen Studierender zu tun haben, die sich nicht geschlossen ein bestimmtes Antwortformat wünschen. Allerdings erwiesen sich das wahrgenommene Potenzial Leistung zu zeigen sowie die wahrgenommene Fairness formatübergreifend als substanzielle Prädiktoren der Valenzbewertung. Dies indiziert, dass zumindest leistungsstarke Studierende (wie die hier untersuchten Psychologiestudierenden mit vorwiegend sehr guten Abiturabschlüssen) anspruchsvolle Prüfungsaufgaben bevorzugen, in denen sie ein (hohes) erreichtes Kompetenzniveau zum Ausdruck bringen können.

Fazit und Limitationen

Diese Studie konnte verschiedene Ausprägungen der Bewertung von MCA und CRA aufzeigen, jedoch keine grundlegend ablehnende oder befürwortende Haltung ge-

Abbildung 1. Verteilung der individuellen Präferenz zugunsten von Multiple-Choice-Aufgaben (Kategorie -3 bis -1), keinem der Formate (Kategorie 0) und zugunsten von Constructed-Response-Aufgaben (Kategorie 1 bis 3) auf einer bipolaren Skala.

genüber einem der Formate. Studierende scheinen demnach mit beiden Formaten vertraut zu sein und auch jeweils verschiedene Vor- und Nachteile wahrzunehmen. Zudem konnten wir keine Evidenz für die Annahme finden, dass leistungsstärkere Studierende CRA gegenüber MCA bevorzugen (vgl. Birenbaum & Feldman, 1998). Dies könnte aber auch an der recht homogen leistungsstarken Stichprobe und einer damit verbundenen Varianzeinschränkung liegen. Prädiktiv für die Präferenz war jedoch die Einschätzung der Formate hinsichtlich verschiedener Facetten (Potenzial Leistung zu zeigen, wahrgenommener Lernaufwand, wahrgenommene Fairness). Insgesamt wäre auch denkbar, dass weitere, hier nicht erfasste situative Faktoren der Prüfungsgestaltung (z.B. Prüfungsziel, Relevanz der Prüfung) sowie die Motive der Studierenden bei der Bewältigung der Prüfung (z.B. „nur“ bestehen, gute Note erreichen, wenig Lernzeit investieren) eine bedeutsame Rolle für die Präferenz zugunsten eines der Formate spielen. Dies sollte in zukünftigen Studien in Form von situationsbezogenen Fragen zusätzlich erfasst werden. Hilfreich wären auch Studien im strukturierten Interview-Verfahren oder in Form von Essay-Bewertungen der Formate, um die Gründe der Präferenz von Studierenden

differenzierter explorieren zu können. Wenngleich die Kombination der Online- und der Labor-Stichproben eine gewisse Generalisierbarkeit der Befunde für den Fachbereich Psychologie nahelegt, steht die Prüfung einer Übertragbarkeit der Befunde auf Studienfächer mit einer möglicherweise anderen Prüfungspraxis aus.

Implikationen für die Hochschulpraxis

Einer Nutzung von MCA sowie CRA in Prüfungen an Hochschulen steht grundsätzlich nichts im Wege. Es scheint zudem jeweils Subgruppen von Studierenden zu geben, die jeweils eines der Formate präferieren. Doch es gibt auch Studierende, die keine spezifische Präferenz zeigen und sich vermutlich mehr Gedanken über die Prüfungsinhalte als über das Prüfungsformat machen. Zur Berücksichtigung der verschiedenen Präferenzen im Sinne einer studierendenfreundlichen Prüfungsgestaltung scheint es dennoch ratsam, eine Kombination aus offenen und geschlossenen Formaten in Klausuren zu nutzen, was sich auch aus messtheoretischer Sicht anbietet (vgl. Lindner et al., 2015; Schult & Sparfeldt, 2018).

Insgesamt weisen die vorliegenden Befunde erneut darauf hin, dass MCA hinsichtlich geprüfter Lernziele und damit verbundener Lernanforderungen besondere Aufmerksamkeit bei der Konstruktion erfordern. Es wäre nämlich denkbar, dass die hier gefundene Tendenz, auch mit geringerem Lernaufwand eine gute Bewertung in MC-Prüfungen zu erreichen, zumindest in Teilen eine suboptimale Prüfungspraxis widerspiegelt. Diese könnte das Lernverhalten unerwünscht beeinträchtigen (vgl. z.B. McCoubrie, 2004; Scouller, 1998; Thiede, 1996), zumal empirische Studien indizieren, dass das Lernen für MCA gegenüber CRA mit oberflächlicheren Lernstrategien (Scouller, 1998; Scouller & Prosser, 1994; Struyven et al., 2005) und kürzeren Lernzeiten (Kellas & Butterfield, 1971) einhergeht. Solchen Effekten kann in der Hochschulpraxis nur durch anspruchsvolle Aufgabenstellungen und substanzielle Energie bei der Aufgabenkonstruktion entgegengewirkt werden (vgl. Lindner et al., 2015; Struyven et al., 2005). Dabei ist auch zu berücksichtigen, dass das Potenzial Leistung zeigen zu können für MCA und CRA als bester Prädiktor mit signifikant höheren Fairnessbewertungen einhergeht.

Inwiefern Lehrende vor dem Hintergrund der Befunde zu erwartende Prüfungsformate vorab offenlegen sollten, ist sicherlich diskutabel und situationsabhängig. Im Sinne der Studierenden halten wir es grundsätzlich für fair, Aufgabenformate anzukündigen; allerdings sollte sich nicht erschließen lassen, welche Inhalte durch welches Format abgeprüft werden. Erwarten Studierende sowohl MC- als auch CR-Formate, ließen sich möglicherweise formatspezifische Einflüsse auf das Lernverhalten reduzieren. Dies

ist jedoch zunächst eine Hypothese, die in zukünftigen Studien empirisch geprüft werden sollte.

Unter Berücksichtigung sämtlicher Ergebnisse möchten wir an Prüfende appellieren, insbesondere Prinzipien der Konstruktion hochwertiger MCA zu studieren (z.B. Haladyna et al., 2002; Haladyna, 2004; Lindner et al., 2015; Waugh & Gonlund, 2013), sofern dieses Format in Prüfungen eingesetzt wird. Dies wäre bei dem zunehmenden Wandel der Prüfungslandschaft zugunsten geschlossener Aufgabenformate im Sinne einer guten Hochschulbildung jedenfalls ausgesprochen wünschenswert.

Literatur

- Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, 32, 347 – 364. doi:10.1007/BF00138871
- Birenbaum, M. & Feldman, R. A. (1998). Relationships between learning patterns and attitudes towards two assessment formats. *Educational Research*, 40, 90 – 98. doi:10.1080/0013188980400109
- Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H. & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I: Cognitive domain*. New York, NY: David McKay.
- Dodeen, H. (2008). Assessing test-taking strategies of university students: Developing a scale and estimating its psychometric indices. *Assessment & Evaluation in Higher Education*, 33, 409 – 419. doi:10.1080/02602930701562874
- Dunlap, W. P., Cortina, J. M., Vaslow, J. B. & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods*, 1, 170 – 177. doi:10.1037/1082-989X.1.2.170
- Gierl, M. J., Bulut, O., Guo, Q. & Zhang, X. (2017). Developing, analyzing, and using distractors for multiple-choice tests in education: A comprehensive review. *Review of Educational Research*, 87, 1082 – 1116. doi:10.3102/0034654317726529
- Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Haladyna, T. M., Downing, S. M. & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15, 309 – 333. doi:10.1207/S15324818AME1503_5
- Haladyna, T. M. & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.
- Huber, L. (2008). „Kompetenzen“ prüfen? In S. Dany, B. Szczybra & J. Wildt (Hrsg.), *Prüfungen auf die Agenda. Hochschuldidaktische Perspektiven auf Reformen im Prüfungswesen* (S. 12 – 26). Bielefeld: Bertelsmann.
- Kastner, M. & Stangl, B. (2011). Multiple choice and constructed response tests: Do test format and scoring matter? *Procedia – Social and Behavioral Sciences*, 12, 263 – 273. doi:10.1016/j.sbspro.2011.02.035
- Kellas, G. & Butterfield, E. C. (1971). Effect of response requirement and type of material on acquisition and retention performance in short-term memory. *Journal of Experimental Psychology*, 88, 50 – 56. doi:10.1037/h0030663
- Kubinger, K. D. (2009). *Psychologische Diagnostik: Theorie und Praxis psychologischen Diagnostizierens* (2. überarbeitete und erweiterte Auflage). Göttingen: Hogrefe.

- Kubinger, K. D. (2014). Gutachten zur Erstellung „gerichtsfester“ Multiple-Choice-Prüfungsaufgaben. *Psychologische Rundschau*, 65, 169 – 178. doi:10.1026/0033-3042/a000218
- Landmann, M., Perels, F., Otto, B., Schnick-Vollmer, K. & Schmitz, B. (2015). Selbstregulation und selbstreguliertes Lernen. In J. Möller & E. Wild (Hrsg.), *Pädagogische Psychologie* (2., vollständig überarbeitete und aktualisierte Auflage, S.45–66). Berlin: Springer. doi:10.1007/978-3-642-41291-2_3
- Lindner, M. A., Strobel, B. & Köller, O. (2015). Multiple-Choice-Prüfungen an Hochschulen? Ein Literaturüberblick und Plädoyer für mehr praxisorientierte Forschung. *Zeitschrift für Pädagogische Psychologie*, 29, 133 – 149. doi:10.1024/1010-0652/a000156
- Liu, O. L., Lee, H. S. & Linn, M. C. (2011). An investigation of explanation multiple-choice items in science assessment. *Educational Assessment*, 16, 164 – 184. doi:10.1080/10627197.2011.611702
- Lukas, J., Melzer, A. & Much, S., unter Mitarbeit von S. Eisentraut (2017). *Auswertung von Klausuren im Antwort-Wahl-Format*. Zugriff am 26.12.2017 http://wiki.llz.uni-halle.de/images/3/38/Handbuch_-_PDF.pdf
- Lukhele, R., Thissen, D. & Wainer, H. (1994). On the relative value of multiple-choice, constructed response, and examinee-selected items on two achievement tests. *Journal of Educational Measurement*, 31, 234 – 250. doi:10.1111/j.1745-3984.1994.tb00445.x
- Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, 34, 207 – 218. doi:10.1207/s15326985ep3404_2
- McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, 26, 709 – 712. doi:10.1080/01421590400013495
- McDaniel, M. A., Blischak, D. M. & Challis, B. (1994). The effects of test expectancy on processing and memory of prose. *Contemporary Educational Psychology*, 19, 230 – 248. doi:10.1006/ceps.1994.1019
- Millman, J., Bishop, C. H. & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, 25, 707 – 726. doi:10.1177/001316446502500304
- Müller, A. & Schmidt, B. (2009). Prüfungen als Lernchance: Sinn, Ziele und Formen von Hochschulprüfungen. *Zeitschrift für Hochschulentwicklung*, 4, 23 – 45. doi:10.3217/zfhe-4-01/03
- Pfister, H.-R., Jungermann, H. & Fischer, K. (2017). *Die Psychologie der Entscheidung. Eine Einführung* (4. Auflage). Berlin: Springer.
- Rodriguez, M. C. (2003). Construct equivalence of multiple-choice and constructed-response items: A random effects synthesis of correlations. *Journal of Educational Measurement*, 40, 163 – 184. doi:10.1111/j.1745-3984.2003.tb01102.x
- Schult, J. & Sparfeldt, J. R. (2018). Reliability and validity of PIRLS and TIMSS. *European Journal of Psychological Assessment*, 34, 258 – 269. doi:10.1027/1015-5759/a000338
- Scouller, K. M. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education*, 35, 453 – 472. doi:10.1023/A:1003196224280
- Scouller, K. M. & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education*, 19, 267 – 279. doi:10.1080/03075079412331381870
- Sparfeldt, J. R., Kimmel, R., Löwenkamp, L., Steingräber, A. & Rost, D. H. (2012). Not read, but nevertheless solved? Three experiments on PIRLS multiple choice reading comprehension test items. *Educational Assessment*, 17, 214 – 232. doi:10.1080/10627197.2012.735921
- Struyven, K., Dochy, F. & Janssens, S. (2005). Students' perceptions about evaluation and assessment in higher education: A review. *Assessment & Evaluation in Higher Education*, 30, 331 – 347. doi:10.1080/02602930500099102
- Thiede, K. W. (1996). The relative importance of anticipated test format and anticipated test difficulty on performance. *The Quarterly Journal of Experimental Psychology*, 49, 901 – 918. doi:10.1080/713755673
- Wan, L. & Henly, G. A. (2012). Measurement properties of two innovative item formats in a computer-based test. *Applied Measurement in Education*, 25, 58 – 78. doi:10.1080/08957347.2012.635507
- Waugh, C. K. & Gronlund, N. E. (2013). *Assessment of student achievement* (10th ed.). Boston, MA: Pearson.
- Winkel, O. (2010). Higher education reform in Germany: How the aims of the Bologna process can be simultaneously supported and missed. *International Journal of Educational Management*, 24, 303 – 313. doi:10.1108/09513541011045245
- Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student's perspective. *Journal of Educational Research*, 80, 352 – 358. doi:10.1080/00220671.1987.10885782

Historie

Manuskript eingereicht: 11. Februar 2018

Manuskript nach Revision angenommen: 3. November 2018

Marlit Annalena Lindner

Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik (IPN)

Olshausenstraße 62

24118 Kiel

0431- 880 4410

mlindner@ipn.uni-kiel.de