Diskussionsforum

Lindner, Marlit Annalena; Sparfeldt, Jörn R.; Köller, Olaf; Lukas, Josef; Leutner, Detlev

Published in:
Psychologische Rundschau

DOI:
10.1026/0033-3042/a000524

Publication date:
2021

Document Version
Verlags-PDF (auch: Version of Record)

Diskussionsforum

Ein Plädoyer zur Qualitätssicherung schriftlicher Prüfungen im Psychologiestudium

Marlit A. Lindner[1], Jörn R. Sparfeldt[2], Olaf Köller[1], Josef Lukas[3] und Detlev Leutner[4]

[1]IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Kiel
[2]Universität des Saarlandes, Saarbrücken
[3]Martin-Luther-Universität Halle-Wittenberg
[4]Universität Duisburg-Essen

Zusammenfassung: Prüfungen sind zentrale, qualitätssichernde Elemente im Hochschulstudium, die hohen Ansprüchen genügen müssen. In der Praxis werden diese Ansprüche jedoch oft nicht idealtypisch erfüllt. Dieser Beitrag skizziert, wie ein evidenzbasiertes Vorgehen bei der Erstellung und Auswertung von schriftlichen Prüfungen aussehen kann, um den Umgang mit typischen Problemen zu erleichtern und kalkulierbare Risiken nach aktuellem Forschungsstand bestmöglich zu reduzieren. Wir diskutieren unter anderem, welche Gütekriterien an Hochschulprüfungen angelegt werden sollten, welche Aufgabenformate mit ihren spezifischen Vor- und Nachteilen wann sinnvoll einsetzbar sind, wie man der Rateproblematik bei geschlossenen Aufgabenformaten begegnen sollte und welche Potenziale computerbasierte Prüfungen bieten. Weiterhin gehen wir auf die Bedeutung von Prüfungen als Lernsteuerungsinstrument ein und diskutieren die Rolle studentischer Testwiseness sowie die Notwendigkeit, durch ineinandergreifende Maßnahmen ein systematisches Qualitätsmanagement für Prüfungen zu implementieren. Darüber hinaus wird im Rahmen des Diskussionsforums ein konstruktiver Dialog über Herausforderungen und Optimierungsbedarfe sowie Voraussetzungen und Gelingensbedingungen auf dem Weg zu besseren Prüfungen im Psychologiestudium fortgeführt.

Schlüsselwörter: Hochschule, Prüfungen, Multiple-Choice-Aufgaben, Studium, Psychologie-Studium

Improving the Quality Management of Written Exams in Psychology Study Programs

Abstract: Exams are central, quality-assuring elements in higher education that must meet high standards. In practice, however, this ideal is often not met. In this article, we outline an evidence-based procedure for the design and scoring of written exams and examinations that facilitates the handling of typical problems and reduces potential risks as much as possible, according to the current state of research. We discuss which quality criteria should be applied to university exams; which item formats can be used, with their specific advantages and disadvantages; how the guessing problem should be addressed in selected response test items; and the potentials of computer-based exams. Furthermore, we discuss the importance of different exam types to control students' learning behavior, the role of students' testwiseness, and the necessity of implementing systematic quality management for written exams. In addition, we aim to further deliberate on a constructive dialogue about challenges and the need for optimization as well as the prerequisites and conditions to construct better exams in university studies of psychology.

Keywords: university, examinations, multiple-choice items, higher education, studies of psychology

Der vorliegende Beitrag ist durch die *Empfehlungen zur Qualitätssicherung in Studium und Lehre* des Vorstandes der Deutschen Gesellschaft für Psychologie (Spinath et al., 2018) inspiriert und widmet sich dem dort skizzierten Anliegen einer angemessenen Erfassung von Prüfungsleistungen im Fach Psychologie. Wir möchten hier an frühere, auf die Konstruktion und Auswertung schriftlicher Prüfungen bezogene Diskussionen (z. B. Kubinger, 2014; Lindner, Stobel & Köller, 2015; Marcus, 2015; Schmidt-Atzert, 2015) anknüpfen und diese weiterführen, um einen stärkeren Erfahrungsaustausch zwischen Lehrenden zur Sicherung einer guten Prüfungspraxis anzuregen und damit zur Qualitätssteigerung im Psychologiestudium beizutragen. Viele dieser Überlegungen lassen sich auf Hochschulprüfungen anderer Fachrichtungen übertragen.

Lehrveranstaltungen und Prüfungen sind zentrale, untrennbare Elemente jeder Hochschulausbildung, die gemeinsam hohen Ansprüchen genügen müssen. Dies spiegelt sich unter anderem in der Grundidee des *Constructive Alignments* (Biggs, 1996) wider, welches einen direkten Bezug von Lehrinhalten, späteren beruflichen Anforderungen und Prüfungsinhalten im Studienkontext fordert. Entsprechende Zusammenhänge zeigen sich besonders deutlich durch die modulbezogene, stetige Leistungsakkreditierung im Zuge der Umstellung des Diplomstudiengangs Psychologie auf das Bachelor-Master-System. Durch

die stetige Akkreditierung von Studienleistungen hat sich das Prüfungsaufkommen im Verlauf des Studiums deutlich erhöht, was sich auch im Arbeitspensum der Lehrenden niederschlägt. Entsprechend gestiegen ist der Einsatz schriftlicher Prüfungsformate in der Psychologie. Insbesondere hat sich dabei ein verstärktes Interesse am Einsatz von standardisierten geschlossenen gegenüber traditionelleren offenen Aufgabenformaten oder mündlichen Prüfungen entwickelt. In diesem Zusammenhang gibt es – teils berechtigte – Bedenken, dass dieser Wandel der Prüfungspraxis auch Risiken für die Qualität der Leistungsprüfungen birgt. Besondere Aufmerksamkeit erhält hier die Diskussion um geschlossene Aufgaben im Antwort-Wahl-Format, da diese viele Fallstricke bei der Erstellung und Administration bieten. Die Problembereiche können aber präzise benannt werden, sodass gerade bei dieser Prüfungsform effektive und begründbare Hinweise und Empfehlungen für Prüfende möglich sind. In diesem Sinne möchten wir einen konstruktiven Dialog anregen, um gemeinsam Herausforderungen und Optimierungsbedarfe sowie Voraussetzungen, Standards und Gelingensbedingungen auf dem Weg zu besseren Prüfungen zu identifizieren.

Gütekriterien

Universitäre Prüfungen können als kriteriumsorientierte Verfahren aufgefasst werden, bei denen eines der wichtigsten Gütemerkmale die Inhaltsvalidität ist. Diese zeichnet sich vor allem durch eine angemessene Auswahl von Prüfungsinhalten mit Blick auf den zu erfassenden Sachbereich und die Breite relevanter Kompetenzen aus (vgl. Marcus, 2015; Schmidt-Atzert, 2015). Wichtigster Faktor ist hierfür neben der sachlichen Qualität der Aufgaben, dass Prüfungsaufgaben eine angemessene Ziehung aus dem theoretischen Aufgabenuniversum darstellen. Einen guten Ausgangspunkt hierfür bieten Modulhandbücher zu Lehrveranstaltungen, die Lehrenden helfen, einen klaren Erwartungshorizont der Veranstaltung zu definieren, konkrete Lehrziele für Studierende abzuleiten und diese transparent zu kommunizieren (vgl. zu Ideen und Technologien zur Begründung von Wissensstrukturen in der psychologischen und kognitionspsychologischen Forschung, z. B. Falmagne, Albert, Doble, Eppstein & Hu, 2013). Bei der Zusammenstellung von Prüfungen bietet sich zudem die Erstellung einer Matrixstruktur zur Sicherung der Inhaltsvalidität an, um die Anzahl der Aufgaben und ihr Anspruchsniveau je Themenfeld auf einer Inhaltsdimension sowie Verhaltens- bzw. Prozessdimension zu verorten. Beispielsweise ließen sich verschiedene klinische Störungsbilder (Lehrinhalte) in einer Matrix mit den Lehrzieltaxonomiestufen (Prozessdimension) kreuzen, um sicherzustellen, dass eine angemessene Verteilung von Prüfungsfragen vorliegt. Im besten Fall sollten Prüfende also eine grobe Einordnung der zur Lösung erforderlichen kognitiven Prozesse und Lösungsschritte wie Reproduktion von Wissen, Verständnis / Anwendung bzw. komplexes Problemlösen vornehmen, sodass diese – wie auch die verschiedenen Inhalte – in den Aufgaben einer Prüfung angemessen repräsentiert sind. Es ist in jedem Fall darauf zu achten, dass die Aufgaben in einer Prüfung eine repräsentative Auswahl von Inhalten und Schwierigkeiten darstellen, was mit einer höheren Aufgabenzahl leichter gelingt. Zudem sollte der Punktwerthorizont eine gute Differenzierung von verschiedenen Leistungsniveaus der Studierenden im Sinne verschiedener Notenstufen (kriteriumsorientierte Klassifikation) erlauben. Neben einer transparenten und objektiven Bewertung spielen auch allgemeine Faktoren der Fairness in Hochschulprüfungen eine wichtige Rolle. Beispielsweise müssen alle Studierenden vorab sämtliche prüfungsrelevanten Informationen erhalten und bei gleicher Kompetenzausprägung zu gleichen Resultaten kommen. Das klingt selbstverständlich, ist aber im Hochschulalltag sicherlich nicht immer (möglicherweise sogar kategorisch niemals idealtypisch) gegeben.

FAZIT: Kernkriterium einer guten universitären Prüfung ist die Abdeckung der fachlichen Breite und Tiefe. Das bedeutet, dass meist ein Bündel fachspezifisch abzuleitender Kompetenzen zu erfassen ist. Daraus ergibt sich auch, dass im Regelfall keine strikte psychometrische Eindimensionalität der Prüfungsaufgaben zu erwarten ist. Entsprechend sind an universitäre Prüfungen teils andere Ansprüche zu stellen als an individualdiagnostische Verfahren zur Messung homogener psychologischer Konstrukte. Wertvolle Hinweise zur Konstruktion von angemessenen Testaufgaben finden sich beispielsweise bei Haladyna (2004) sowie bei Waugh & Gronlund (2013).

Schriftliche Aufgabenformate

In schriftlichen Prüfungen sind zwei wesentliche Kategorien von Aufgabenformaten zu unterscheiden: Offene Aufgaben (*Contructed-Response-Aufgaben*), bei deren Bearbeitung der Prüfling eine eigene Antwort in Reaktion auf einen Aufgabenstamm frei formulieren muss, und geschlossene Aufgaben (*Multiple-Choice-Aufgaben* bzw. Aufgaben mit Antwort-Wahl-Verfahren), in denen standardisierte Antwortmöglichkeiten vorgegeben werden, unter denen der Prüfling nach einer vorgegebenen, formatspezifischen Regel eine oder mehrere Antworten gemäß der Passung zum Aufgabenstamm auswählen muss. Bei der Formatauswahl für Prüfungen sind verschiedene Anwen-

dungsbereiche sowie Vor- und Nachteile zu berücksichtigen, die die diagnostische Aussagekraft beeinflussen (Martinez, 1999).

Anwendungsbereiche. Während eine Abfrage von Wissen mit geschlossenen Aufgaben gleichermaßen erfolgreich oder sogar besser als mit offenen Aufgaben gelingen kann, sind kreativ-schöpferische Leistungen (z. B. die Darlegung einer kreativen Experimentalplanung) ausschließlich über offene Aufgabenformate zu prüfen (vgl. Diskussion bei Lindner et al., 2015). Wenngleich es einfacher ist, komplexe Problemlöseprozesse in offenen Aufgaben abzubilden, sollte das nicht zur Schlussfolgerung führen, dass dies *ausschließlich* mit offenen Aufgabenformaten möglich ist. Eine aufwändige Konstruktion geschlossener Aufgaben macht es durchaus möglich, deutlich über die Abfrage von deklarativem Wissen und Verständnis hinauszugehen, und dies sollte beim Einsatz geschlossener Formate auch angestrebt werden. Vor allem die vorgelegten Aufgabenstämme müssen sorgsam durchdacht und von Seiten der Materialien so ausgestattet sein, dass tatsächlich anspruchsvolle Problemlösungsprozesse bzw. eine Wissensanwendung für die Aufgabenlösung erforderlich sind. Die Konstruktion komplexer geschlossener Aufgaben benötigt daher vergleichsweise hohe zeitliche Ressourcen und differenziertes Fachwissen sowie diagnostische Expertise. Ein konkretes Beispiel, das veranschaulicht, wie eine anspruchsvolle Multiple-Choice-Prüfungsaufgabe aussehen kann, findet man bei Marcus (2015).

Unserer Erfahrung nach werden der Kreativität bei der Erstellung von geschlossenen Aufgabenformaten gegenüber offenen Aufgaben bislang zu enge Grenzen gesetzt. So bieten beispielsweise Zuordnungsaufgaben (Matching) und Anordnungsaufgaben (Ranking) neben Single-Choice-, Multiple-Choice- und Richtig-Falsch-Aufgaben ([multiple] True-False-Aufgaben) interessante Alternativen der Aufgabengestaltung, die jeweils unterschiedliche Kompetenzfacetten abdecken können (vgl. z. B. Lindner et al., 2015; Waugh & Gronlund, 2013). Auch Hotspot-Aufgaben, bei denen vordefinierte Bildbereiche korrekt identifiziert werden müssen (z. B. eine bestimmte Region in einem Hirnschnittbild kennzeichnen), können relevante Aspekte einer zu erfassenden Kompetenz abbilden. Erwogen werden sollten auch halboffene Aufgabenformate (vgl. Rütter, 1978) wie Lückentextvarianten und Kurzantworten, die eine weitgehend automatisierte Auswertung ermöglichen und damit die gewünschten ökonomischen Vorteile bei gleichzeitig vielfältigen Gestaltungsoptionen von Prüfungsaufgaben bieten. Insbesondere computerbasierte Prüfungen potenzieren die Möglichkeiten, komplexe Aufgaben in sämtlichen Formaten zu erstellen, worauf wir später näher eingehen. Studierende sollten in jedem Fall vor der Prüfung durch Übungsaufgaben ein gutes Verständnis erlangen, wie verschiedene Aufgabentypen zu bearbeiten sind, um konstrukt-irrelevante Varianz zu reduzieren.

Vor- und Nachteile der Formate. Zentraler Vorteil geschlossener gegenüber offenen Aufgaben ist die Elimination einer eigenständigen Formulierung von Antworten durch die Studierenden. Dies erhöht die Prüfungseffizienz, da im Durchschnitt eine höhere Anzahl von Aufgaben je Zeiteinheit gestellt werden kann, was eine größere Abdeckung der fachlichen Breite erlaubt und eine höhere diagnostische Informationsdichte bzw. Repräsentativität der curricularen Inhalte in der Prüfung bedingt (Lukhele, Thissen & Wainer, 1994). Zudem ist die Auswertungsobjektivität (und Fairness) besonders hoch, da die Auswertungsregeln standardisiert sind und die Fehleranfälligkeit bei automatisierter Verarbeitung minimal ausfällt.

Problematisch sind geschlossene Aufgabenformate, wenn sie in Prüfungen ausschließlich eingesetzt werden und nicht hinreichend in den Randbereichen differenzieren. Fehlen beispielsweise sehr schwierige Aufgaben, haben besonders leistungsstarke Studierende in geschlossenen Formaten kaum Möglichkeiten, überdurchschnittliche Kompetenzausprägungen umfassend zum Ausdruck zu bringen (z. B. Liu, Lee & Linn, 2011; Rauch & Hartig, 2010).

Offene Aufgaben haben dagegen ihre Stärke bei der Prüfung komplexer Lehrziele, wenn Sachverhalte eigenständig von den Prüflingen erarbeitet und dargestellt werden sollen. Zudem ist die Erstellung offener Aufgaben im Regelfall vermeintlich weniger zeitaufwändig, da zunächst nur ein Aufgabenstamm formuliert werden muss. Der größte Nachteil offener Aufgaben liegt in der späteren Kreditierung der Leistungen, die schwerer objektiv bewertet werden können und zudem einen erheblichen Aufwand bei der Bewertung erfordern. Für einen hohen Standard der Leistungserfassung setzt dies voraus, dass bereits vor der Administration einer Prüfung ein dezidiertes Bewertungsschema (schriftlich) festgehalten wird.

Grundsätzlich kann man davon ausgehen, dass die Konstruktionszeit guter geschlossener Aufgaben gegenüber offenen Aufgaben deutlich länger ausfällt, während sich dieses Verhältnis bei der Kreditierung umkehrt. Ein unmittelbarer zeitlicher Gesamtvorteil durch die Konstruktion von geschlossenen Aufgaben ist daher – zumindest bei sorgsamer Erstellung – nicht unbedingt zu erwarten. Geschlossene Aufgaben werden allerdings umso zeiteffizienter, je häufiger sie in Leistungsprüfungen eingesetzt werden und je größer die Anzahl zu prüfender Personen ist. Der Aufbau einer Aufgabendatenbank mit einem substanziellen Anteil geschlossener Aufgaben lohnt sich daher vor allem für Vorlesungen und mehrgleisige Seminare mit großen Studierendenzahlen. Nicht nur, aber

vor allem geschlossene Aufgaben müssen zudem immer wieder in geeigneter Form abgewandelt werden, um zu gewährleisten, dass ein Auswendiglernen alter Prüfungsfragen nicht automatisch zum Erfolg führt. In Zeiten eines engen Austausches unter Studierenden über soziale Netzwerke dürfte nämlich ein Großteil alter Prüfungsfragen bekannt sein. Dies gilt gleichermaßen für offene Prüfungsfragen, bei denen immerhin noch ein eigener Text in der Klausur (re)produziert werden muss, wobei eine reine Reproduktion sicherlich nicht dem Anspruch des Faches genügen kann. Empirisch sollten wiederholt eingesetzte Aufgaben daher über Kohorten hinweg häufiger richtig gelöst werden. Dies ließe sich in der Praxis testen, um zu einfach gewordene Aufgaben rechtzeitig auszusortieren.

FAZIT: Lehrende sollten sich mit der Vielfalt offener und geschlossener Aufgabenformate, den Herausforderungen der Konstruktion sowie spezifischen Vor- und Nachteilen schriftlicher Prüfungsaufgaben vertraut machen und dies bewusst in die Prüfungskonzeption einbeziehen. Durch eine gezielte Nutzung und den kreativen Einsatz verschiedener Aufgabenformate können auch mit geschlossenen Aufgaben Lehrzieltaxonomiestufen abgebildet werden, die eine reine Wissensreproduktion übersteigen und Anwendungs- bzw. Transferwissen erfordern (vgl. z. B. Marcus, 2015). Dabei ist allerdings zu berücksichtigen, dass Lehrende einen höheren Aufwand bei der Aufgabenkonstruktion einkalkulieren müssen, um langfristig von Ökonomie und Objektivität im Rahmen der Auswertung geschlossener Aufgabenformate zu profitieren. Für eingesetzte offene Aufgaben sollte viel Wert auf eine objektivierbare, schriftlich fixierte Kategorisierung richtiger und falscher Lösungen für die Bewertung der Antworten gelegt werden. In Abhängigkeit von den zu prüfenden Kompetenzen sollte grundsätzlich das jeweils am besten geeignete Format aus den verschiedenen Varianten offener und geschlossener Aufgaben gewählt werden. Oft bietet sich die Kombination offener und geschlossener Aufgabenformate in einer Klausur an.

Prüfungen und Lernverhalten

Hochschulprüfungen haben zwei Hauptfunktionen: (1) diagnostische Aussagen zum aktuellen Lern- und Leistungsstand von Studierenden zu ermöglichen und (2) Studierende zu angemessenem Lernverhalten anzuregen, um fachliche Kompetenzstandards zu erfüllen. Während der erste Aspekt vor allem die psychometrische Qualität von Prüfungen als diagnostisches Instrument betrifft, steht die zweite Funktion vorrangig mit der Annahme eines prüfungsgetriebenen Lernens in Verbindung (*Assessment Drives Learning*; z. B. McCoubrie, 2004; Scouller, 1998; Thiede, 1996). Um tiefgründiges Lernverhalten zu motivieren, ist eine klare Lehrzielorientierung der Prüfung bezogen auf die Themenschwerpunkte der Lehrveranstaltung und spätere berufliche Anforderungen (*Constructive Alignment*) sicherzustellen. Zudem sollten Prüfungsanforderungen im Vorfeld klar kommuniziert werden, um eine Erwartungshaltung der Studierenden zu wecken, die zielführendes, selbstreguliertes Lernverhalten begünstigt.

Formaterwartungseffekte. Die Erwartungshaltung von Studierenden bezüglich der Prüfungsformate kann erheblichen Einfluss auf ihr Lernverhalten nehmen. So zeigten verschiedene Studien, dass geschlossene Aufgaben als weniger anspruchsvoll eingeschätzt werden und die Vorbereitung auf Multiple-Choice-Prüfungen mit oberflächlicherem Lernverhalten einhergeht (z. B. Kellas & Butterfield, 1971; Struyven, Dochy & Janssens, 2005; Thiede, 1996; Zeidner, 1987). Im direkten Vergleich offener und geschlossener Aufgabenformate durch 330 Psychologiestudierende verschiedener Standorte in Deutschland zeigte sich ebenfalls die Erwartungshaltung, dass geschlossene Aufgaben mit einem deutlich geringerem Lernaufwand und einem höheren erwarteten Erfolg von *Testwiseness*-Prüfungsstrategien verbunden sind, während die Objektivität der Bewertung und die Fairness höher eingeschätzt wurden (Lindner, Mayntz & Schult, 2018). Der größte Unterschied zwischen den Formatbewertungen ergab sich bezogen auf das wahrgenommene Potenzial, Leistung zu zeigen, das für offene Aufgaben sehr viel höher eingeschätzt wurde. Eine geringere Anstrengung beim Lernen für eine Klausur mit geschlossenen Aufgaben wäre daher aus lernökonomischer Perspektive durchaus nachvollziehbar.

Geschlossene Aufgaben sind zudem in der Regel tatsächlich einfacher zu lösen als inhaltsgleiche offene Aufgaben (z. B. Rodriguez, 2003; vgl. Schult & Sparfeldt, 2018). Vermutlich trägt auch die Möglichkeit, durch strategisches Abwägen und Taktieren mehr Punkte in geschlossenen Aufgaben zu erreichen, zu der lernungünstigeren Erwartungshaltung bei. Plausibel wäre darüber hinaus, dass Lehrende geschlossene Aufgaben in der Praxis tatsächlich häufiger für einfache Wissensabfragen nutzen und sich somit auch faktisch geringere fachliche Ansprüche in den Bewertungen der Studierenden widerspiegeln. Insgesamt legen die Daten von Lindner et al. (2018) allerdings auch nahe, dass sich die befragten Psychologiestudierenden tendenziell anspruchsvolle Prüfungen wünschen. So ging ein höher eingeschätztes Potenzial, die eigene Leistung in einer Prüfung zeigen zu können, sowohl bei offenen als auch bei geschlossenen Prüfungsaufgaben mit einer größeren wahrgenommenen Fairness einher.

Formative Tests. Grundsätzlich wäre es im Sinne der Lernerfolgssicherung wünschenswert, einen stärkeren Fokus auf formative, semesterbegleitende Tests zu legen, die ein kontinuierliches Lernverhalten fördern und die Bedeutung summativer Prüfungen für die Lernsteuerung reduzieren. Viele Studien zeigen, dass die Bearbeitung von Tests als Lernstrategie gegenüber anderen gängigen Lernstrategien (z. B. wiederholtes Lesen, Mindmapping) besonders erfolgreich ist (*Testing-Effekt*; z. B. Rowland, 2014; Schwieren, Barenberg & Dutke, 2017). Eine weitere Möglichkeit, tiefgründiges Lernverhalten zu begünstigen, kann die Einbindung von Studierenden in die Konstruktion von Testaufgaben für ihre Mitstudierenden sein. Studien weisen auf substanzielle, positive Effekte der Konstruktionsübung eigener geschlossener Aufgaben hin (z. B. Fellenz, 2004), was möglicherweise auch auf offene Aufgaben mit Musterlösungen übertragbar ist. Dass die Studierenden dabei Grundprinzipien der Aufgabenkonstruktion kennenlernen, ist im diagnostisch orientierten Fach Psychologie ein wünschenswerter Nebeneffekt. Selbst konstruierte (optimierte) Aufgaben können dann wiederum für formative Prüfungszwecke eingesetzt werden, womit sich positive Effekte der Aufgabenkonstruktionsübung mit dem Testing-Effekt verbinden ließen.

FAZIT: Prüfende sollten aktiv Einfluss auf die Erwartungshaltung der Studierenden an die Prüfung nehmen und dadurch das Lernverhalten indirekt positiv beeinflussen. Wichtig ist dabei, grundsätzlich anspruchsvolle Prüfungen zu stellen. Prüfungen müssen auch leistungsstarken Studierenden die Möglichkeit geben, ihre Kompetenzen zu zeigen. Um formatspezifische Prüfungserwartungen zu minimieren, sollten offene und geschlossene Aufgabenformate gemischt eingesetzt werden. Semesterbegleitende formative Prüfungen sind besonders sinnvoll (vgl. verteiltes Lernen, Testing-Effekt).

Rateproblematik

Die zunehmende Nutzung geschlossener Antwortformate in Hochschulprüfungen geht auch mit der Diskussion um ein altbekanntes Problem einher: der Möglichkeit, richtige Antworten auch ohne jedes Wissen durch zufällige Auswahl einer Antwort zu produzieren („Raten"). Wer beispielsweise bei einfachen True-False-Aufgaben zufällig immer eine der beiden Antwortoptionen auswählt, ohne die Aufgabe zu lesen, der gibt im Schnitt in 50 % der Aufgaben eine richtige Antwort. Aufgaben mit hoher Ratewahrscheinlichkeit führen so zu Testergebnissen, die viel besser sind, als es basierend auf dem fachlichen Wissen angemessen wäre. Die (unerlässliche) Korrektur des Rateeffektes ist jedoch nicht ganz trivial. Von den vielen Lösungsansätzen, die für das Rateproblem vorgeschlagen wurden, werden die prominentesten hier hinsichtlich ihrer Vor- und Nachteile kurz diskutiert.

Anzahl der Antwortoptionen. Die scheinbar einfachste Lösung zur Ratekorrektur, nämlich geschlossene Aufgaben mit möglichst vielen Antwortoptionen zu konstruieren (vgl. Kubinger, 2014), ist wenig zielführend und erweist sich als kaum praktikabel. Zum einen wird damit der Rateeffekt nicht vollständig eliminiert (selbst bei 10 Antwortoptionen pro Aufgabe erreicht man durch Raten im Durchschnitt noch 10 % der maximalen Punktzahl). Viel problematischer aber ist zum anderen, dass es in aller Regel außerordentlich schwierig ist, zu einer richtigen Antwort genügend viele gute Distraktoren zu formulieren. Schlechte Distraktoren sind nämlich auch ohne einschlägiges Wissen oft als unzutreffend erkennbar, sodass für die Prüflinge nur wenige falsche Antwortalternativen tatsächlich in Frage kommen. Der Versuch, weitere falsche Antwortoptionen zu erfinden, kostet unverhältnismäßig viel Konstruktionszeit und führt selten zum erwünschten Ergebnis (z. B. Haladyna & Downing, 1993). So zeigt sich, dass unattraktive Distraktoren von Studierenden nicht einmal dann gewählt werden, wenn die besten Distraktoren entfernt werden (Shizuka, Takeuchi, Yashima & Yoshizawa, 2006). Die Ratewahrscheinlichkeit wird also nur scheinbar reduziert. Schlechte Distraktoren tragen nicht zu einer Entschärfung der Rateproblematik bei, sondern nur zu deren Verschleierung.

Viele empirische Forschungsbefunde zur Frage der optimalen Anzahl der Antwortoptionen in Multiple-Choice-Aufgaben (vgl. Übersicht bei Lindner et al., 2015) lassen unter Berücksichtigung inhaltlicher und zeitökonomischer Faktoren ableiten, dass eine Anzahl von drei Optionen in klassischen Multiple-Choice-Aufgaben („1 aus 3") oft ausreichend ist, wenn gleichzeitig die geringere Anzahl der Antwortoptionen durch eine höhere Gesamtanzahl der Aufgaben in der Prüfung kompensiert wird (z. B. Vyas & Supe, 2008). Es bleibt zwar grundsätzlich empfehlenswert, so viele *gute* Distraktoren wie möglich vorzugeben, allerdings sollte im Zweifelsfall die Entscheidung zugunsten hochwertiger, trennscharfer Distraktoren fallen und die Qualität immer vor der Quantität der Distraktoren Vorrang haben.

Malus-Punkte. Ein anderer Ansatz zur Ratekorrektur ist die Einführung von Malus-Punkten, also bestimmte vorab definierte Abzüge für falsche Antworten der Prüflinge (vgl. Kubinger, 2014). Die Logik hinter dieser Maßnahme ist, dass Studierende, die bei Unsicherheit raten, mit hoher Wahrscheinlichkeit Punktabzüge bekommen werden. Beim *Formula Scoring* werden für jede falsche

Antwort $\frac{g}{(1-g)}$ Punkte abgezogen (Frary, 1988; Holzinger, 1924). Je höher die Ratewahrscheinlichkeit *g*, desto größer der Punktabzug. Dieses Vorgehen führt dazu, dass der Erwartungswert der erreichten Punkte so gerade dem ratekorrigierten Wissen des Prüflings entspricht, unabhängig von der individuellen Neigung zu raten oder im Zweifel nicht zu antworten (Lukas, Melzer & Much, 2017). Diese effektive statistische Ratekorrektur ist allerdings kaum bekannt. Malus-Punkte werden dagegen in der Praxis oft unabhängig von der Ratewahrscheinlichkeit vergeben (meist ein Minuspunkt für eine falsche Antwort), was zu statistisch grob verzerrten und nicht begründbaren Ergebnissen führt.

Gegen die Anwendung von Malus-Punkten – und das betrifft vermutlich auch das *Formula Scoring* – sprechen zudem juristische Urteile, die eine Anwendung von Malus-Punkten faktisch verbieten (vgl. Kubinger, 2014): Nachweisliches Wissen (d. h. eine richtige Antwort) darf nicht durch falsche Antworten auf ganz andere Fragen diskreditiert werden. Zwar gibt es juristisch zulässige Zwischenregelungen (Vergabe von Malus-Punkten nur innerhalb einer Aufgabe mit Minimalpunktzahl „Null"), die aber aus statistischen Gründen kaum zu rechtfertigen sind.

Malus-Punkte für falsche Antworten werden im Übrigen häufig von Prüfenden nicht in erster Linie als statistische Ratekorrektur verstanden, sondern als erzieherische Maßnahme, um das Raten zu bestrafen. Studierende sollen sich aus Furcht vor Punktabzug im Zweifelsfall gegen das Raten entscheiden und lieber keine Antwort als eine potenziell falsch geratene Antwort geben (vgl. z. B. Frary, 1988; Lesage, Valcke & Sabbe, 2013). Dies ist aus mehreren Gründen problematisch: So dürften Persönlichkeitsunterschiede zwischen Studierenden zu stark unterschiedlichem Rateverhalten führen. Risikoscheue Studierende, die sich bei Unsicherheit trotz vorhandenem (Teil-)Wissen eher gegen das Raten entscheiden, werden z. B. systematisch benachteiligt (vgl. z. B. Lesage et al., 2013). Zudem führen komplexe Regeln der Kreditierung von Leistungen zu einem stark taktisch orientierten Vorgehen, das mit der zu bewertenden Kompetenz nichts zu tun hat und die Fairness bzw. Validität der Prüfungsergebnisse infrage stellt. Maluspunkte sind daher insgesamt kritisch zu sehen. Wenn überhaupt, kommt nur das *Formula Scoring* in Frage.

Alles-oder-Nichts-Prinzip. Eine andere, ebenfalls nicht unproblematische Alternative zum Umgang mit Ratewahrscheinlichkeiten, ist die Verwendung von Multiple-Choice-Aufgaben vom Typ „x aus X" (Multiple-Response) mit einem Alles-oder-Nichts-Scoring (vgl. Kubinger, 2014). Konkret wird ein Punkt nur dann zuerkannt, wenn exakt das richtige Antwortmuster in einer Aufgabe angekreuzt wurde. Bei diesem Vorgehen wird Teilwissen nicht honoriert und man verliert diagnostisch wertvolle Information, da beispielsweise Personen mit 3 von 4 richtigen Kreuzen mit Personen gleichgestellt werden, die kein einziges Kreuz korrekt gesetzt haben. Nur wenn Teilwissen für ein bestimmtes Sachgebiet nicht kreditierungswürdig ist, mag eine solche Wertung in begründeten Ausnahmen sinnvoll sein. Zudem wird die Alles-oder-Nichts-Regelung von Studierenden (zu Recht) als unfair wahrgenommen (vgl. auch Marcus, 2015). Interessanterweise trifft diese nämlich vor allem solche Studierende hart, die vergleichsweise viel Wissen haben, aber durch Flüchtigkeitsfehler oder Unsicherheiten bei einzelnen Antworten einen substanziellen Teil ihrer erbrachten Leistung nicht angerechnet bekommen (siehe Lukas et al., 2017). Auch Fehler bei der Aufgabenkonstruktion erhalten beim Alles-oder-Nichts-Scoring ein unverhältnismäßig hohes Gewicht mit besonders negativen Konsequenzen für leistungsstarke Studierende.

Schmidt-Atzert (2015) stellt in diesem Zusammenhang beispielhaft ein Verfahren mit Multiple-Choice-Aufgaben vom Typ „2 aus 5" vor, in dem er sich für eine Vergabe von bis zu 2 Teilpunkten ausspricht (ein Punkt pro richtiger Antwort), was auf einer abwägenden Betrachtung des ratekorrigierten Punktwertbereichs beruht. Allerdings sind die zugrunde liegenden Überlegungen vergleichsweise komplex und formatabhängig. Dadurch ist eine Übertragbarkeit auf andere Aufgabentypen nicht trivial. Ein zufriedenstellender Einsatz der Vielfalt geschlossener Aufgabenformate ist allerdings nur möglich, wenn ein globaler Ansatz zur Ratekorrektur zugrunde gelegt werden kann, der auf sämtliche Prüfungskonstellationen übertragbar ist.

Ratekorrigierte Bestehens- und Notengrenzen. Ein umfassend begründeter Vorschlag wurde von Lukas et al. (2017) dargelegt, der eine Verbindung der einfachen *Number-Right-Scoring*-Methode (für jede richtige Antwort wird ein Punkt vergeben) mit einer dynamischen Adaptation von Bestehens- und Notengrenzen beinhaltet und damit verwandt mit der Idee eines ratekorrigierten Punktwertbereiches von Schmidt-Atzert (2015) ist, jedoch eine höhere Flexibilität aufweist[1]. Ähnlich wie beim *Formula Scoring* wird in diesem Ansatz eine wahrscheinlichkeitstheoretisch begründete Ratekorrektur vorgeschlagen. Al-

[1] Siehe auch die bei Klauer (1987) dargestellten Ansätze zur Korrektur der Bestehensgrenze unter binomialer Berücksichtigung sowohl der Ratewahrscheinlichkeit als auch der Anzahl der zu bearbeitenden Aufgaben.

lerdings wird nicht der individuelle Punktwert der Prüflinge durch Punktabzug für Falschantworten korrigiert, sondern es werden die für alle Prüflinge geltenden Bestehens- und Notengrenzen angepasst. Je höher die Ratewahrscheinlichkeiten für die Aufgaben einer Klausur sind, desto mehr Punkte muss man erreichen, um zu bestehen (oder eine bestimmte Note zu bekommen). Ausgangspunkt ist immer ein unkorrigiertes Benotungsschema (Notenschlüssel), das gemäß der Ratewahrscheinlichkeiten der einzelnen Aufgaben in einer Klausur so adaptiert wird, dass die angehobene Bestehensgrenze eine effektive Korrektur von Rateeffekten beinhaltet. Bei einer Klausur mit max. 100 Punkten würde man z. B. die übliche Bestehensgrenze von 50 Punkten (für Aufgaben ohne Ratewahrscheinlichkeit) bei Aufgaben mit Ratewahrscheinlichkeit 0.25 (z. B. „1 aus 4“) auf 62.5 Punkte anheben oder bei Aufgaben mit Ratewahrscheinlichkeit 0.5 (z. B. richtig-falsch) auf 75 Punkte. Eine Notengrenze von 95 Punkten für die Bewertung „sehr gut“ würde in den beiden genannten Fällen auf 96.25 bzw. 97.5 Punkte angehoben. Dem Ansatz unterliegt die Annahme, dass Studierende ihre Antworten nach bestem Wissen auswählen und im Zweifel immer raten.[2] Das muss den Prüflingen vor der Prüfung klar kommuniziert werden, um faire Voraussetzungen für das Verfahren zu schaffen. Ausführliche Begründungen für sämtliche gängige Aufgabenformate sowie Tipps und Hilfestellungen zur Umsetzung in der Praxis sind bei Lukas et al. (2017) dargelegt.

Im Vergleich zu den vorangegangenen Verfahren ist die Adaptation von Bestehensgrenzen (1) transparent, fair und allen Beteiligten vermittelbar, (2) statistisch präzise begründbar, (3) flexibel für alle offenen und geschlossenen Aufgabenformate und sämtliche Kombinationen dieser Formate anwendbar, (4) das Verfahren kommt ohne Malus-Punkte und die damit verbundenen juristischen und psychologischen Problemen aus, und es ermöglicht nicht zuletzt (5) neben der Berücksichtigung des Fehlers erster Art (richtige Antwort bei Nicht-Wissen, also: Raten) auch die (optionale) Berücksichtigung einer hypothetischen Fehlertoleranz für den Fehler zweiter Art (falsche Antwort trotz sicheren Wissens). Selbstverständlich muss den Prüflingen das Verfahren zur Adaptation von Bestehensgrenzen vor der Prüfung klar erläutert werden, um faire Voraussetzungen zu schaffen.

Die größte Hürde für ratekorrigierte Bestehens- und Notengrenzen sind Prüfungsordnungen, die eine explizite Vorgabe für die Festlegung von Bestehensgrenzen (z. B. 50 % oder 60 % der maximalen Punktzahl) machen. Hier wäre grundsätzlich eine gewisse Flexibilität wünschenswert, um nach Zusammenstellung der Prüfungsaufgaben (und Kenntnis der klausurspezifischen Ratewahrscheinlichkeiten) eine sinnvolle Wahl der Bestehensgrenze vorzunehmen. Vor allem aber sollte die große Heterogenität von Prüfungsordnungen an verschiedenen Instituten (sowie in juristischen Auslegungen) abgebaut und flächendeckend durch einheitliche, wissenschaftlich gestützte Regelungen ersetzt werden.

Aufgabenanzahl und Aufgabenqualität. Neben einer expliziten Korrektur von zu erwartenden Rateeffekten spielt für die diagnostische Aussagekraft von Prüfungen auch die Anzahl der eingesetzten Testaufgaben eine wichtige Rolle. Insbesondere bei geschlossenen Formaten ist eine deutlich höhere Anzahl an Aufgaben notwendig, um eine diagnostisch belastbare Differenzierung zwischen den Kenntnisständen von Studierenden zu erlauben. So indiziert beispielsweise eine Studie von Pamphlett (2005), dass mindestens 100 diagnostische Informationspunkte (z. B. 100 *True-False*-Aufgaben mit einer Ratewahrscheinlichkeit von jeweils 0.5) erforderlich sind, um eine angemessene Trennung von Leistungsniveaus zu erreichen. Dagegen trug eine deutlich höhere Anzahl an Aufgaben in der Studie nicht mehr substanziell zu einer besseren Diagnostik bei. Wenngleich diese Befunde mit etwas Vorsicht zu interpretieren sind, zeigen unter anderem unsere eigenen Prüfungserfahrungen, dass ein Klausurumfang von mindestens 100 Aufgaben bzw. Informationspunkten auch unter universitären Alltagsbedingungen leicht umsetzbar ist und eine angemessene Differenzierung erlaubt.

Empfehlenswert ist vor diesem Hintergrund eine Nutzung von Multiple-Response-Aufgaben mit x richtigen aus X Antwortoptionen, die jeweils einzeln als richtig oder falsch bewertet werden (*Multiple-True-False*-Format). Dieser geschlossene Aufgabentyp ist vergleichsweise ökonomisch zu konstruieren, da zusätzliche richtige Antworten leichter zu erstellen sind als falsche Antworten. Zudem können die einzelnen Antworten separat bewertet werden und liefern mehr diagnostische Information. In jedem Fall ist neben einer angemessenen Aufgabenschwierigkeit und Aufgabentrennschärfe sicherzustellen, dass auch sämtliche eingesetzte Distraktoren psychometrisch hinreichend gut funktionieren (u. a. die Auswahlhäufigkeit und Trennschärfe der Distraktoren; siehe z. B. Gierl, Bulut, Guo & Zhang, 2017, für nähere Ausführungen). Die Distraktoren sollten für unwissende Studierende ähnlich attraktiv wie die richtige Antwort sein. Nur durch die

[2] Unter „Raten“ verstehen wir hier das rein zufällige Auswählen einer Antwortoption, z. B. durch Werfen einer Münze oder eines Würfels (dazu zählt explizit nicht das Finden einer richtigen Antwort im Ausschlussverfahren oder die Anwendung von *Testwiseness*-Strategien).

Verwendung hochwertiger Distraktoren kann vermieden werden, dass richtige Antworten auf einem Trivial- oder Alltagswissen beruhen, das nichts mit dem Prüfungsstoff zu tun hat.

FAZIT: Der Einsatz geschlossener Prüfungsaufgaben birgt stets das Risiko, dass Prüflinge auch ohne Wissen durch Raten zusätzliche Punkte erhalten. Zum Umgang mit diesem Problem empfiehlt sich ein einfaches *Number-Right-Scoring* (1 Punkt pro richtiger Antwort) bei gleichzeitiger Ratekorrektur der Bestehens- und Notengrenzen. Die Anzahl der Aufgaben und die Aufgabenqualität (insbesondere der Distraktoren) müssen ausreichend hoch sein, um die Inhaltsvalidität zu gewährleisten. Malus-Punkte für Falschantworten sowie das Alles-Oder-Nichts-Prinzip der Bewertung sollten aus den oben genannten Gründen nicht eingesetzt werden. Empfehlenswert erscheint in jedem Fall eine psychometrische Prüfung der Güte der (offenen und geschlossenen) Aufgaben einer Klausur.

Testbearbeitungsstrategien

Neben der Rateproblematik gibt es das Problem der Anwendung systematischer Testbearbeitungsstrategien. Diese stehen seltener im Zentrum der Diskussion, haben allerdings deutlich problematischere Auswirkungen als zufälliges Raten. So genannte *Testwiseness*-Strategien sind bei geschlossenen Formaten besonders erfolgreich, wenn die Aufgaben Konstruktionsfehler aufweisen (vgl. z. B. Downing, 2005). *Testwiseness* wurde von Millman, Bishop und Ebel (1965) als Personeneigenschaft definiert, die sich durch die Exposition mit standardisierten Prüfungen ausprägt und sich durch Wissen um bestimmte Heuristiken auszeichnet, die helfen, Hinweisreize in den Antworten besser zu interpretieren, um diese mit höherer Wahrscheinlichkeit als eher richtige oder eher falsche Antworten zu klassifizieren.

Zu klassischen *Testwiseness*-Strategien zählen beispielsweise das Wissen, dass Antwortoptionen, die in mittleren Positionen stehen, statistisch gesehen häufiger richtig sind, da (uninformierte) Prüfende eine Tendenz haben, diese in der Mitte „zu verstecken". Es ist daher grundsätzlich empfehlenswert, eine randomisierte Zuteilung der richtigen Antwortoptionen auf die Antwortpositionen vorzunehmen (Bar-Hillel, Budescu & Attali, 2005), sofern keine inhaltlich logische (z. B. numerische) Anordnung der Antworten möglich ist. Auch lange Antwortoptionen sind oft ein Hinweis auf die richtige Lösung, da Prüfende besonders ausführliche Erläuterungen oder Angaben in richtigen Antworten nutzen, um sicherzustellen, dass die als richtig intendierte Antwort auch faktisch richtig ist. Besonders vorsichtig müssen Prüfende mit Signalwörtern in Antwortoptionen umgehen, die entweder durch ihren extremen Charakter (z. B. alle, immer, niemals) Hinweise auf falsche Antworten liefern oder durch Relativierungen (z. B. eher, manchmal, meistens) Hinweise auf richtige Antworten geben. Damit *Testwiseness*-Strategien anwendbar sind, müssen allerdings Schwächen in der Aufgabenkonstruktion vorliegen, was vermutlich nicht nur im medizinischen Bereich (vgl. z. B. Downing, 2005; Tarrant & Ware, 2008), sondern auch in der Psychologie eher Regelfall als Ausnahme sein dürfte.

Die genannten *Testwiseness*-Strategien treten bei offenen Aufgaben nicht in vergleichbarer Form auf, da sich Lösungshinweise hauptsächlich auf Antwortoptionen bzw. die Passung von Antworten und Aufgabenstamm beziehen. Allerdings werden auch für offene Aufgaben einschlägige prüfungsstrategische Verhaltensweisen berichtet (vgl. z. B. Waugh & Gronlund, 2013) – wie ein sprachlich elaboriertes Wiederholen der Aufgabenstellung, ein absichtlich mehrdeutiges Formulieren der Antwort oder das Aufzählen von mehr Argumenten als gefordert (in der Hoffnung, dass x richtige Argumente unter den Einzelantworten sind und goutiert werden).

Testkonstruktionskompetenz. Wenngleich Lehrende der Psychologie qua Ausbildung kundig im Feld der Psychologischen Diagnostik sind, ist offen, ob dieses Wissen immer konsequent auf die Erstellung und Administration von Prüfungen im Rahmen der Hochschullehre übertragen wird. Neben mangelnden zeitlichen Ressourcen fehlt es manchmal an spezifischer Expertise zur Testaufgabenkonstruktion mit dem Ziel der Erfassung von Fachwissen bzw. berufsbezogenen Kompetenzen. Entsprechende wertvolle Hinweise hierzu finden sich beispielsweise in *Item-Writing Guidelines* (z. B. Haladyna, 2004; Haladyna, Downing & Rodriguez, 2002; Lindner et al., 2015; Waugh & Gronlund, 2013). Die Rolle der Testkonstruktionskompetenz wird insbesondere für geschlossene Aufgabenformate oft unterschätzt, deren Erstellung ein besonders hohes Maß an Fach- und Konstruktionswissen, Kreativität und Geschick sowie sprachliches Feingefühl erfordert.

Die darüber hinausgehende wichtige Kenntnis von allgemeinen Aufgabenkonstruktionsprinzipien zeigt sich beispielsweise in Untersuchungen zur Aufgabenqualität in der medizinischen Lehre, in der die Verwendung von geschlossenen Aufgabenformaten eine lange Tradition hat. Selbst hier lassen sich viele Konstruktionsfehler identifizieren, die uninformierten Testerstellenden nicht bewusst sind (z. B. Downing, 2005; Tarrant & Ware, 2008), wobei spezifische Trainings die Qualität der Aufgaben nachweislich verbessern können (Jozefowicz et al., 2002). Vielfach hilft eine Vermittlung typischer Konstruktionsfehler, die durch ein Bewusstsein für die Problematik ef-

fektiv reduziert werden können. Dazu gehören insbesondere unterschwellige Hinweise auf die korrekte Lösung durch grammatikalische Fehler, herausstechende Antworten oder sprachliche Hinweisreize, die von Studierenden mit hoher *Testwiseness* ausgenutzt werden (vgl. z. B. Case & Swanson, 2002; Lindner et al., 2015; Sparfeldt, Kimmel, Löwenkamp, Steingräber & Rost, 2012). Wenn Prüfende nicht die zeitlichen Kapazitäten haben, um sicherzustellen, dass hochwertige und weitgehend fehlerfreie geschlossene Aufgaben erstellt werden können, sollte auf den Einsatz geschlossener Aufgaben verzichtet werden. Jedoch weisen im universitären Alltag eingesetzte offene Aufgaben ebenfalls häufig erhebliche Schwächen auf (z. B. missverständliche oder nicht eindeutige Formulierungen in der Aufgabenstellung).

FAZIT: Um die erfolgreiche Anwendung von Testbearbeitungsstrategien zu unterbinden, müssen (geschlossene) Aufgaben vor ihrem Einsatz besonders penibel auf sprachliche, grammatikalische und logische Fehler untersucht werden (vgl. auch Abschnitt *Qualitätssicherung*).

Computerbasierte Prüfungen

Computerbasierte Prüfungen stellen hohe technische, organisatorische und juristische Anforderungen. Dennoch wurden an manchen Universitäten (z. B. Göttingen, Bremen) in den letzten Jahren bereits umfangreiche infrastrukturelle Einrichtungen geschaffen, die eine zeitgleiche Prüfung von über 200 Studierenden in groß angelegten Computerpools erlauben. In den kommenden Jahren ist mit einer zunehmenden Verbreitung digitaler Prüfungen an Universitäten zu rechnen, weshalb es essenziell ist, sich schon jetzt mit entsprechenden Konzepten zu beschäftigen. Informationsangebote und Ansätze zum Austausch für Lehrende zur Digitalisierung von Lehre und Prüfungen bieten beispielsweise einschlägige Online-Plattformen wie *e-teaching.org* oder *hochschulforumdigitalisierung.de*.

Neue Darstellungsformen. Einer der wichtigsten Vorteile computerbasierter Prüfungen ist das Potenzial einer höheren Praxisrelevanz von Testaufgaben durch neuartige Möglichkeiten der Darstellung komplexer Sachverhalte, die in traditionellen Papierprüfungen nur schwer oder gar nicht umsetzbar sind. Dazu gehört insbesondere der Einsatz multimedialer, multimodaler und dynamischer Materialien, wie beispielsweise Kombinationen aus Texten, Bildern, Animationen, Videos und Audioaufnahmen, und die Möglichkeit zur interaktiven Steuerung. Diese neuen Darstellungsformen erlauben eine Revolution der Aufgabengestaltung, die vor allem eine praxisorientierte Herangehensweise und damit eine höhere ökologische Validität der Prüfungsaufgaben gewährleisten kann. Vor allem die Verbindung von geschlossenen Aufgabenformaten mit komplexen Stimuli (z. B. Videovignetten, interaktive Simulationsaufgaben) ist vielversprechend, um die Grenzen standardisierter Aufgaben bestmöglich auszuschöpfen und höhere Lehrziele abzudecken. Potenzial bietet sich jedoch nicht nur hinsichtlich neuer Aufgabeninhalte, sondern auch für die Nutzung neuer Aufgabenformate mit mehrstufigen Antwortprozessen (z. B. *Explanation-Multiple-Choice*-Aufgaben, *Answer-Until-Correct*-Format; z. B. Liu et al., 2011; Wilcox, 1981). Weiterhin können beispielsweise gängige statistische Programme (z. B. R, G-Power) für realitätsnahe Aufgabenstellungen zum Einsatz kommen. Auch spielerische Elemente (*Gaming*) und Feedback könnten vor allem im Rahmen von formativen computerbasierten Tests eingesetzt werden, um einen motivations- und lernförderlichen Einfluss auf die Studierenden zu nehmen.

Adaptive Prüfungen. Computerbasierte Aufgaben erlauben eine Umsetzung von adaptiven Testverfahren, bei denen basierend auf dem individuell festgestellten Leistungsniveau jeweils Aufgaben mit einer Lösungswahrscheinlichkeit von z. B. etwa 0.5 dargeboten werden, was eine genaue und gleichzeitig effiziente Leistungsdiagnostik erlaubt (Frey, 2012). Allerdings braucht es für adaptive Testverfahren hinreichend große und gut normierte Aufgabenpools, die in der Praxis bislang wohl nicht bestehen und höchstens über kollegiale Zusammenschlüsse in zufriedenstellender Qualität und Quantität erstellbar wären. Weiterhin ist die Vergleichbarkeit der Prüfungsvoraussetzungen für Studierende durch die Darbietung völlig unterschiedlicher Testaufgaben in unterschiedlicher Zusammenstellung (vgl. auch Aufgabenpositionseffekte durch Ermüdungs- und Lerneffekte, z. B. Nagy, Lüdtke & Köller, 2016) in summativen Prüfungen nicht nur aus juristischer Sicht fragwürdig. Problematisch sehen kann man auch, dass für adaptive Prüfungen ausschließlich geschlossene Aufgabenformate oder halboffene Aufgaben mit Kurzantworten in Frage kommen, was einen negativen Einfluss auf das Lernverhalten nehmen könnte. Abgesehen von administrativen Hürden eignen sich adaptive Tests daher eher für den Bereich semesterbegleitender Prüfungen zu Lernzwecken (z. B. dargeboten über Lernplattformen), da adaptive Tests eine vergleichsweise differenzierte und effiziente Rückmeldung zum Leistungsstand geben und Studierende so effektiv beim Lernen unterstützen könnten.

Automatisierte Auswertungen. Einer der wichtigsten Vorteile computerbasierter Prüfungen ist die automatisierbare Auswertung geschlossener Aufgabenformate. Dieser

effizienzförderliche Aspekt der Prüfungsadministration ist ein wichtiges Argument zugunsten der Nutzung computerbasierter Prüfungen, da es den Lehrenden Arbeitsbelastung in der Korrektur abnimmt und gleichzeitig den Studierenden durch eine schnelle Rückmeldung der Ergebnisse beispielsweise bei Nichtbestehen mehr Zeit für die Vorbereitung auf die Nachklausur bietet. Hinzu kommt, dass basierend auf den Antworten der Studierenden eine automatische Errechnung von Aufgabenparametern vorgenommen werden kann, die Prüfenden eine unkomplizierte Kontrolle der Aufgabenqualität erlauben und eine Anpassung von zu schwierigen bzw. zu leichten Aufgaben, wenig trennscharfen Distraktoren oder negativen Aufgabentrennschärfen ermöglichen, ohne substanziellen Aufwand in die Datenanalyse zu investieren. Weiterhin könnte eine solche psychometrische Auswertung helfen, bei Studierenden verbreitete Fehlkonzepte oder Kompetenzlücken zu identifizieren; in künftigen Lehrveranstaltungen könnte dann differenzierter darauf eingegangen werden.

Bei aller zu erwartender Effizienzsteigerung darf aber nicht übersehen werden, dass die erste Administration von digitalen Prüfungen zunächst vergleichsweise viel Aufwand erfordert und ein Effizienzvorteil erst nach vollständiger Etablierung und mehreren Prüfungsdurchgängen eintreten dürfte, selbst wenn administrative und technische Rahmenbedingungen von Seiten der Universität geschaffen werden. Allein die Hinterlegung von Aufgaben in einer digitalen Aufgabendatenbank ist zunächst eine substanzielle zeitliche Investition, die manche Prüfende von einer Umstellung altbewährter papierbasierter Prüfungen auf digitale Prüfungen abschrecken mag. Wichtig wäre es, von Seiten der Hochschulen personelle und technische Ressourcen für den Umstieg bereitzustellen, da andernfalls eine zeitliche Verzögerung der Nutzung bereits verfügbarer Technologien zu erwarten wäre.

FAZIT: Computerbasierte Prüfungen erlauben es, die Qualität und den Anspruch gestellter Aufgaben deutlich zu erhöhen. Neben einer automatisierten Auswertung geschlossener Aufgaben machen die Möglichkeiten der Darstellung komplexer Materialien, beispielsweise unter Einsatz multimedialer Elemente, den Einsatz computerbasierter Prüfungen besonders attraktiv. Eine Orientierung in Richtung digitaler Prüfungen ist insgesamt positiv zu sehen, wenngleich es vermutlich noch Jahre dauern wird, bis das Potenzial an Universitäten voll ausgeschöpft werden kann. Bis dahin besteht beispielsweise hinsichtlich der diagnostischen Eignung neuer Aufgabenformate, dem Einsatz multimedialer Elemente und Fragen der Usability digitaler Prüfungen noch viel Forschungsbedarf.

Qualitätssicherung

Wie die vorangehenden Abschnitte nahelegen, ist die Konstruktion guter Prüfungen eine anspruchsvolle und oft unterschätzte Aufgabe. Während an Hochschulen inzwischen zurecht ein sehr großer Fokus auf die Qualität der Lehre gelegt wird, ist die Qualität von Prüfungen mindestens gleichermaßen bedeutungsvoll, wenngleich diese seltener im Fokus stehen. Das ist insofern verwunderlich, als dass der Zusammenhang zwischen Lernverhalten und Prüfungsanforderungen umso wichtiger zu berücksichtigen ist, je weniger Kontrolle durch die Lehrenden in den Veranstaltungen ausgeübt werden kann. Durch die weitgehende Abschaffung der Anwesenheitspflicht an einigen Standorten ist davon auszugehen, dass der Stellenwert einer angemessenen Gestaltung von Prüfungen nochmals gestiegen ist, da diese im Zweifelsfall wichtigstes Element zur Sicherung der fachlichen Qualität bleiben, sofern Studierende sich überwiegend für ein Selbststudium entscheiden.

Korrekturschleifen. Insbesondere die Nutzung geschlossener Aufgaben erfordert ein Vier-Augen-Prinzip, das im Idealfall dokumentiert sein sollte. Dies ist vor dem Hintergrund geboten, dass die Äußerungsmöglichkeiten der Prüflinge bei geschlossenen Aufgaben per Definition stark eingeschränkt sind und eine Bewertung der Antworten bereits implizit in der Vorlage der Aufgaben inbegriffen ist, da diese vor der Klausur vollständig feststehen. Somit ist es Studierenden in der Prüfung nicht möglich, Konstruktionsfehler der Lehrenden durch differenzierte Äußerungen auszugleichen. Eine explizite Qualitätssicherung durch Korrekturschleifen mit Fachkolleginnen und -kollegen ist daher sehr zu empfehlen.

In der Regel ist es Aufgabe der Lehrenden, den informellen Austausch proaktiv zu suchen, um die Aufgabenqualität sicherzustellen. Solche Kooperationen werden meistens nicht systematisch von Seiten der Institute begleitet. Allerdings wäre es hilfreich, eine übergeordnete Struktur zur gegenseitigen Validierung von Prüfungsaufgaben einzurichten, denn selbst eine sehr sorgsame Erstellung von Testaufgaben ist kein Garant für Fehlerfreiheit; vor allem subtile sprachliche Hinweise und fachliche Unklarheiten fallen oft nur auf, wenn man unter Kolleginnen oder Kollegen unabhängig voneinander nach entsprechenden Konstruktionsfehlern sucht. In diesem Zusammenhang sollte auch das Wissen über geschlossene Aufgabenformate und Konstruktionsprinzipien zukünftig noch stärker durch hochschuldidaktische Weiterbildungsangebote verbreitet werden. Darüber hinaus sollten Kommentare von Studierenden zu missverständlichen Formulierungen der Aufgaben zugelassen werden, um unklare Aufgabenstellungen identifizieren und diese ggf. nach-

träglich aus der Wertung nehmen zu können. Dies muss den Studierenden aus Gründen der Chancengleichheit vor der Prüfung bekannt sein und zum Vorteil aller angewandt werden (vgl. Beaucamp & Buchholz, 2010).

Institutionelle Zusammenarbeit. Ein strukturelles Problem der Qualitätssicherung liegt in mangelnden zeitlichen bzw. personellen Ressourcen der einzelnen Lehrenden. Wünschenswert wäre daher, die Ressourcen und die Expertise in der Psychologie zu bündeln, um eine allgemein bessere Prüfungsqualität zu gewährleisten. Denkbar wären verschiedene Formen der Zusammenarbeit, beispielsweise institutsinterne Korrekturkreisläufe innerhalb des Fachs und über Fachgebiete hinweg sowie eine Erstellung gemeinsamer, institutsübergreifender Aufgabenpools in den Fachgemeinschaften.

Ein gewisses Vorbild für eine optimierte psychologische Prüfungspraxis könnten hier die vergleichsweise sorgsam erstellten medizinischen Prüfungen sein. Da in der Medizin seit Jahrzehnten international und auch in deutschen Staatsexamina zu einem großen Anteil mit geschlossenen Aufgabenformaten gearbeitet wird, hat in diesem Bereich eine lange Auseinandersetzung mit den positiven und negativen Konsequenzen dieser Prüfungspraxis stattgefunden (z. B. Case & Swanson, 2002; Haladyna, 2004). In Deutschland werden bereits seit 1972 sämtliche Staatsexamensaufgaben in dem eigens für diesen Zweck gegründeten *Institut für Medizinische und Pharmazeutische Prüfungsfragen* (IMPP) entwickelt, um den besonderen Anforderungen an (geschlossene) Prüfungsaufgaben gerecht zu werden (Beaucamp & Buchholz, 2010). Darüber hinaus gibt es überinstitutionelle Zusammenschlüsse für Prüfungen jenseits der Staatsexamina, die den kooperativen Austausch sowie ein gemeinsames Review von eingestellten Prüfungsfragen über ein digitales Item-Management-System erlauben (z. B. der Prüfungsverbund Medizin; UCAN, 2019). Durch diese institutionelle Kooperation können Prüfende in der Medizin auf einen kollektiv gepflegten, wachsenden Aufgabenpool zurückgreifen und ohne allzu großen Aufwand bessere Prüfungen stellen sowie eine solide empirische Qualitätssicherung der Aufgaben und Distraktoren gewährleisten. Voraussetzung für die Nutzung ist beispielsweise die Rückmeldung erlangter Aufgabenkennwerte sowie ggf. die Bereitstellung eigener Aufgaben, die dem gemeinschaftlichen Pool zugeführt werden.

Gemäß den Empfehlungen des Vorstandes der Deutschen Gesellschaft für Psychologie (Spinath et al., 2018) ist eine Angleichung von basalen psychologischen Curricula über Studienstandorte hinweg sowie die Vergleichbarkeit von Prüfungsleistungen wünschenswert. Um diesem Ziel in der Psychologie näher zu kommen, wäre es naheliegend, sich im Hinblick auf die gestellten Anforderungen in psychologischen Prüfungen abzustimmen, wenngleich die Freiheit der Lehrenden dadurch selbstverständlich nicht eingeschränkt werden darf. Ein gemeinsam gepflegter Aufgabenpool unter federführender Koordination durch die Deutsche Gesellschaft für Psychologie würde jedoch die Vergleichbarkeit der Leistungen an verschiedenen Studienstandorten erleichtern und Inspiration für die eigene Prüfungspraxis bieten. Allem voran würde es den Lehrenden die Möglichkeit geben, mit individuell überschaubarem Aufwand gute Prüfungen zu erstellen und die Aufgabenqualität in größerem Maßstab zu prüfen bzw. auf routinemäßige psychometrische Analysen der bereitgestellten Aufgaben zurückgreifen zu können. Dies hätte vermutlich Vorteile für die Effizienz und die Qualität der Prüfungen und würde zu einer weiteren Etablierung gemeinsamer Standards in den Psychologie-Studiengängen beitragen. Eine professionalisierte Erarbeitung von Prüfungsfragen könnte es auch ermöglichen, komplexe Aufgaben mit hohem Konstruktionsaufwand umzusetzen (z. B. im Rahmen computerbasierter Prüfungen).

FAZIT: Ausgangspunkt einer guten Prüfungspraxis ist eine Sicherstellung der Qualitätsprüfung eingesetzter Testaufgaben. Dafür wären Kooperationsstrukturen innerhalb von und zwischen Instituten hilfreich, die Korrekturzirkel, eine Pflege gemeinsamer Aufgabendatenbanken oder technische Lösungen unterstützen. Nachdrücklich zu empfehlen ist es, bei geschlossenen Aufgaben stets ein Vier-Augen-Prinzip umzusetzen, um Konstruktionsfehler zu korrigieren und den Erfolg von Rate- und Testbearbeitungsstrategien systematisch einzudämmen. Ein großflächiger Zusammenschluss im Fach Psychologie zum Austausch von Testaufgaben und der Wahrung von Prüfungsstandards wäre wünschenswert.

Schlussbemerkungen

Die Bedeutung einer guten Prüfungspraxis im Fach Psychologie ist für die Qualität und für die Reputation der Studiengänge von höchster Bedeutung. Um zu gewährleisten, dass Prüfungen den Ansprüchen an ihre diagnostische und lernförderliche Funktion gerecht werden, müsste neben einer weiteren Verbreitung von Testkonstruktionswissen auch die Relevanz von Prüfungen in der universitären Lehre stärker in den Fokus rücken und eine engere Zusammenarbeit forciert werden. Die Etablierung gemeinsamer Kriterien, die empirische Befunde für eine fachlich ausgewogene und diagnostisch sinnvolle Prüfungsgestaltung einbeziehen, würde erheblich zur Sicherung der Studienqualität beitragen. Einige Ansatzpunkte wurden in diesem Diskussionsbeitrag vorgestellt. Die angesprochenen Punkte zur Qualität und Qualitätssicherung

von Hochschulprüfungen sind dabei nicht nur für die Psychologie-Hauptfachstudiengänge bedeutsam: Zum einen lehren und prüfen Kolleginnen und Kollegen der Psychologie in weiteren Studiengängen mit teilweise sehr vielen Studierenden und entsprechender Relevanz einer qualitativ hochwertigen und effizienten Prüfungsgestaltung (z. B. Lehramtsstudiengänge, Pädagogik, Wirtschaftswissenschaft). Zum anderen bietet die in der Psychologie vorhandene diagnostische Expertise eine wesentliche Grundlage für die hochschuldidaktische Aufgabe einer Verbesserung von Hochschulprüfungen in unterschiedlichsten Fächern. Möglicherweise wäre es eine lohnende Aufgabe für die Deutsche Gesellschaft für Psychologie, die Kompetenzen unseres Fachs explizit zu bündeln und für eine fachübergreifende Qualitätssicherung des Prüfungsgeschehens an Hochschulen zur Verfügung zu stellen. In diesem Zusammenhang besteht allerdings noch viel praxisorientierter Forschungsbedarf (siehe auch Lindner et al., 2015).

Da hier vor allem Kernaspekte der Konstruktion schriftlicher Prüfungen angesprochen wurden, sind darüber hinausgehende, ausführliche Diskussionen über die allgemeine Prüfungspraxis in der Psychologie wünschenswert, um die übergreifende Zusammenarbeit im Rahmen der Lehre auf den Bereich der Prüfungen auszuweiten und neue Strukturen zu schaffen, die eine kollektive Verbesserung der Lehr- und Prüfungspraxis im Psychologiestudium erlauben. Dabei sollte auch die Rolle anderer, beispielsweise mündlicher und digitaler Prüfungsformate sowie die allgemeine Ausrichtung der Prüfungsordnungen weiter diskutiert werden. Auch im Rahmen der Einführung des neuen Psychotherapiestudiengangs ist eine Auseinandersetzung mit angemessenen Approbationsprüfungen unter Einsatz neuer Formate hoch relevant. Das mit diesem Artikel verbundene Diskussionsforum bietet eine Plattform, um den fachinternen Austausch zu beginnen. In diesem Sinne freuen wir uns auf Diskussionsbeiträge, die weitere Themen, Problemstellungen und Ideen einbringen, die eine differenziertere Sicht auf Prüfungen in den Psychologiestudiengängen erlauben und möglicherweise aktuelle Best-Practice-Beispiele sowie innovative Konzepte für die Zukunft skizzieren.

Literatur

Bar-Hillel, M., Budescu, D. & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind & Society*, *4*, 3 – 12. https://doi.org/10.1007/s11299-005-0001-z

Beaucamp, G. & Buchholz, J. A. (2010). Rechtsfragen bei der Einführung von Multiple-Choice-Prüfungen (Antwort-Wahl-Verfahren). *Wissenschaftsrecht*, *43*, 56 – 67.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*, 347 – 364. https://doi.org/10.1007/BF00138871

Case, S. M. & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences* (3rd ed.). Philadelphia, PA: National Board of Medical Examiners.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*, 133 – 143. https://doi.org/10.1007/s10459-004-4019-5

Falmagne, J.-C., Albert, D., Doble, C., Eppstein, D. & Hu, X. (Eds.). (2013). *Knowledge Spaces: Applications in Education.* Heidelberg: Springer.

Fellenz, M. R. (2004). Using assessment to support higher level learning: The multiple choice item development assignment. *Assessment & Evaluation in Higher Education*, *29*, 703–719. https://doi.org/10.1080/0260293042000227245

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice*, *7*(2), 33 – 38.

Frey, A. (2012). Adaptives Testen. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2., aktual. und überarb. Aufl., S. 275 – 293). Berlin: Springer.

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. https://doi.org/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. (1993). How many options is enough for a multiple choice test item? *Educational and Psychological Measurement*, *53*, 999 – 1010. https://doi.org/10.1177/0013164493053004013

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. https://doi.org/10.1207/S15324818AME1503_5

Holzinger, K. J. (1924). On scoring multiple response tests. *Journal of Educational Psychology, 15*, 445 – 447. https://doi.org/10.1037/h0073083

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D. & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine*, *77*, 156 – 161. https://doi.org/10.1097/00001888-200202000-00016

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. https://doi.org/10.1037/h0030663

Klauer, K. J. (1987). *Kriteriumsorientierte Tests.* Göttingen: Hogrefe. Siehe auch die bei Klauer (1987) dargestellten Ansätze zur Korrektur der Bestehensgrenze unter binomialer Berücksichtigung sowohl der Ratewahrscheinlichkeit als auch der Anzahl der zu bearbeitenden Aufgaben.

Kubinger, K. D. (2014). Gutachten zur Erstellung „gerichtsfester" Multiple-Choice-Prüfungsaufgaben. *Psychologische Rundschau*, *65*, 169 – 178. https://doi.org/10.1026/0033-3042/a000218

Lesage, E., Valcke, M. & Sabbe, E. (2013). Scoring methods for multiple choice assessment in higher education – Is it still a matter of number right scoring or negative marking? *Studies in Educational Evaluation*, *39*, 188 – 193. https://doi.org/10.1016/j.stueduc.2013.07.001

Lindner, M. A., Mayntz, S. M. & Schult, J. (2018). Studentische Bewertung und Präferenz von Hochschulprüfungen mit Aufgaben im offenen und geschlossenen Antwortformat. *Zeitschrift*

für Pädagogische Psychologie, 32, 239 – 248. https://doi.org/10.1024/1010-0652/a000229

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. https://doi.org/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. https://doi.org/10.1080/10627197.2011.611702

Lukas, J., Melzer, A. & Much, S., Eisentraut, S. (2017). *Auswertung von Klausuren im Antwort-Wahl-Format.* Halle (Saale): @LLZ, Zentrum für multimediales Lehren und Lernen. http://nbn-resolving.de/urn:nbn:de:gbv:3:2-66099

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. https://doi.org/10.1111/j.1745-3984.1994.tb00445.x

Marcus, B. (2015). Multiple-Choice-Prüfungsaufgaben in der Psychologie: Eine Erwiderung auf Kubingers (2014) Gutachten. *Psychologische Rundschau*, *66*, 166 – 170. https://doi.org/10.1026/0033-3042/a000253

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist*, *34*, 207 – 218. https://doi.org/10.1207/s15326985ep3404_2

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

Millman, J., Bishop, C. H. & Ebel, R. (1965). An analysis of test-wiseness. *Educational and Psychological Measurement*, *25*, 707 – 726. https://doi.org/10.1177/001316446502500304

Nagy, G., Lüdtke, O. & Köller, O. (2016). Modeling test context effects in longitudinal achievement data: Examining position effects in the longitudinal German PISA 2012 assessment. *Psychological Test and Assessment Modeling*, *58*, 641 – 670. https://www.pedocs.de/frontdoor.php?source_opus=12804

Pamphlett, R. (2005). It takes only 100 true-false items to test medical students: true or false? *Medical Teacher*, *27*, 468 – 470. https://doi.org/10.1080/01421590500097018

Rauch, D. & Hartig, J. (2010). Multiple-choice versus open-ended response formats of reading test items: A two-dimensional IRT analysis. *Psychological Test and Assessment Modeling*, *52*, 354 – 379.

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. https://doi.org/10.1111/j.1745-3984.2003.tb01102.x

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, *140*, 1432 – 1463. https://doi.org/10.1037/a0037559

Rütter, T. (1978). Formen der Testaufgabe. In K. J. Klauer (Hrsg.), *Handbuch der Pädagogischen Diagnostik* (Bd. 1, S. 257 – 280). Düsseldorf: Schwann.

Schmidt-Atzert, L. (2015). „Kommentar zum Gutachten zur Erstellung ‚gerichtsfester' Multiple-Choice-Prüfungsaufgaben" von Klaus D. Kubinger. *Psychologische Rundschau*, *66*, 170 – 173. https://doi.org/10.1026/0033-3042/a000254

Schult, J. & Sparfeldt, J. R. (2018). Reliability and validity of PIRLS and TIMSS. *European Journal of Psychological Assessment*, *34*, 258 – 269. https://doi.org/10.1027/1015-5759/a000338

Schwieren, J., Barenberg, J. & Dutke, S. (2017). The testing effect in the psychology classroom: A meta-analytic perspective. *Psychology Learning & Teaching*, *16*, 179 – 196. https://doi.org/10.1177/1475725717695149

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. https://doi.org/10.1023/A:1003196224280

Shizuka, T., Takeuchi, O., Yashima, T. & Yoshizawa, K. (2006). A comparison of three-and four-option English tests for university entrance selection purposes in Japan. *Language Testing*, *23*, 35 – 57. https://doi.org/10.1191/0265532206lt319oa

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. https://doi.org/10.1080/10627197.2012.735921

Spinath, B., Antoni, C., Bühner, M., Elsner, B., Erdfelder, E., Fydrich, T. et al. (2018). Empfehlungen zur Qualitätssicherung in Studium und Lehre: Verabschiedet vom Vorstand der DGPs am 20. April 2018. *Psychologische Rundschau*, *69*, 183 – 192. https://doi.org/10.1026/0033-3042/a000408

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. https://doi.org/10.1080/02602930500099102

Tarrant, M. & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, *42*, 198 – 206. https://doi.org/10.1111/j.1365-2923.2007.02957.x

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. https://doi.org/10.1080/713755673

Umbrella Consortium for Assessment Networks (2019, Dezember). *Fachbereich Medizin der Justus-Liebeig-Universität Gießen.* Verfügbar unter: https://www.ucan-assess.org/cms/de

Vyas, R. & Supe, A. (2008). Multiple-choice questions: A literature review on the optimal number of options. *National Medical Journal of India*, *21*, 130 – 133.

Waugh, C. K. & Gronlund, N. E. (2013). *Assessment of student achievement* (10th ed.). Boston, MA: Pearson.

Wilcox, R. (1981). Solving measurement problems with an answer-until-correct scoring procedure. *Applied Psychological Measurement*, *5*, 399 – 414. https://doi.org/10.1177/014662168100500313

Zeidner, M. (1987). Essay versus multiple-choice type classroom exams: The student's perspective. *Journal of Educational Research*, *80*, 352 – 358. https://doi.org/10.1080/00220671.1987.10885782

Förderung

Open Access-Veröffentlichung ermöglicht durch IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik, Kiel.

Dr. Marlit Annalena Lindner
IPN – Leibniz-Institut für die Pädagogik der Naturwissenschaften und Mathematik
Olshausenstraße 62
24118 Kiel
mlindner@leibniz-ipn.de

https://doi.org/10.1026/0033-3042/a000524