Zeitschrift für Pädagogische Psychologie, 29 (3–4), 2015, 133–149

Übersichtsartikel

Multiple-Choice-Prüfungen an Hochschulen?

Ein Literaturüberblick und Plädoyer für mehr praxisorientierte Forschung

Marlit A. Lindner, Benjamin Strobel und Olaf Köller

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

Zusammenfassung: Multiple-Choice-Aufgaben (MCA) sind bei der Leistungsmessung großer Personengruppen besonders ökonomisch. Im Zuge des hohen Prüfungsaufkommens im Bachelor-Master-System werden MCA-Klausuren auch an deutschen Hochschulen immer häufiger eingesetzt. Doch welche diagnostische Qualität haben Prüfungen mit MCA und wo liegen Vorteile und Probleme? In diesem Literaturüberblick kommen wir zu vier zentralen Ergebnissen: (1) MCA von hoher Qualität sind in vielen Fällen diagnostisch vergleichbar zu Constructed-Response-Aufgaben. (2) Es existieren effektive Strategien, um Rateeffekten zu begegnen. (3) Der Einfluss des Prüfungsformats auf Lern- und Prüfungsstrategien ist kaum vermeidbar. (4) Besonders geeignet für den Hochschulkontext sind die MC-Formate Multiple-Response und Multiple-True-False sowie insbesondere computerbasierte Testaufgaben. Zusätzlich zeigen wir einen Mangel an Forschungsarbeiten auf, die für belastbare Aussagen über den diagnostischen Wert von MCA in realen Kontexten unerlässlich sind und leiten daraus Forschungsfragen ab.

Schlüsselwörter: Multiple-Choice-Aufgaben, Prüfungen an Hochschulen, Testkonstruktion, Leistungsdiagnostik, Prüfungsstrategien

Are Multiple-Choice Exams Useful for Universities? A Literature Review and Argument for a More Practice Oriented Research

Abstract: Multiple-choice questions (MCQ) are particularly efficient in measuring achievement in large student groups. Due to the high number of tests in the bachelor-master-system, German universities administer MCQ exams with increasing frequency. So what is the diagnostic quality of exams using MCQ and which assets and drawbacks are associated with MCQ application? In the course of this literature review we draw four essential conclusions: (1) High quality MCQ share similar diagnostic characteristics with constructed-response questions in many cases. (2) There are potent strategies to address (the problem of) guessing in MCQ. (3) Effects of MCQ on learning and testing strategies are hardly avoidable. (4) The multiple-response and multiple-true-false format as well as computer-based MCQ-formats are particularly suitable for university exams. Additionally, we identify a considerable lack of research in this area and propose research desiderata so that the diagnostic value of MCQ in higher education can be reliably evaluated in the future.

Keywords: multiple-choice items, high-stakes exams, item writing, achievement assessment, testwiseness

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

1 Relevanz von Multiple-Choice-Aufgaben (MCA) an Hochschulen

Durch die Einführung von Bachelor-Master-Studiengängen an deutschen Hochschulen ist ein gesteigertes Prüfungsaufkommen zu verzeichnen, das auf die Notwendigkeit einer Kreditierung jedes Leistungsmoduls zurückgeht (Winkel, 2010). Infolge dieser Entwicklung ist das Interesse an MCA aufgrund ökonomischer Vorteile gegenüber Fragen mit offenem Antwortformat (*constructed response* [CR]; auch *open response*) und mündlichen Prüfungen merklich gestiegen. Wurden Klausuren mit MCA bis vor wenigen Jahren vorwiegend und mit langer Tradition in medizinischen und rechtswissenschaftlichen Studiengängen eingesetzt, finden sie mittlerweile Einzug in den Prüfungsalltag vieler Fächer. Dabei stellt sich die Frage, wie gut sich das MC-Format für verlässliche leistungsdiagnostische Aussagen eignet und welche Vorteile und Herausforderungen mit dem Einsatz von MCA in Hochschulprüfungen verbunden sind.

DOI 10.1024/1010-0652/a000156 © 2015 Verlag Hans Huber, Hogrefe AG, Bern

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

Die Forschung zur Qualität von MCA hat eine lange Tradition, die bis zum Anfang des 20. Jahrhunderts zurückreicht. Gleichwohl zeichnet sich die Forschungslage vor allem durch eine große Heterogenität der Befunde aus. Ziel des vorliegenden Reviews ist es, die Befundlage zu evaluieren und vier für die Nutzung von MCA an Hochschulen zentrale Fragen zu adressieren. Wir widmen uns zunächst (1) der diagnostischen Vergleichbarkeit von MC- und CR-Prüfungsformaten und beleuchten die Befunde für den Hochschulkontext. Anschließend diskutieren wir, (2) wie problematisch das potenzielle Erraten richtiger Lösungen bei MCA ist und stellen Möglichkeiten zum Umgang mit Rateeffekten vor. Danach gehen wir darauf ein, (3) welchen Einfluss der Einsatz von MCA auf Lern- und Lösungsstrategien der Studierenden haben kann. Weiterhin (4) stellen wir eine große Vielfalt von MC-Formaten vor und wägen ihre Eignung für Hochschulprüfungen ab. Auf diesem Weg zeigen wir auch bestehende Forschungslücken auf. Schon hier möchten wir darauf hinweisen, dass viele dargestellte Befunde und unsere Schlussfolgerungen zu weiten Teilen auf Felder jenseits der Hochschule übertragbar sind, in denen high-stakes Leistungsüberprüfungen mit Hilfe von MCA stattfinden (z. B. Eignungsdiagnostik). In Abgrenzung zu anderen Einsatzgebieten und zur kritischen Einordnung der Befunde diskutieren wir zum Abschluss die besonderen Herausforderungen der Verwendung von MCA im Hochschulkontext ausführlicher, geben fünf konkrete Praxisempfehlungen und leiten künftigen Forschungsbedarf ab.

2 Strategien der Literatursuche

Die Literaturrecherche erfolgte vorrangig über Google Scholar, PubPsych, PsychInfo, ERIC und Bibliothekskataloge. Neben dem Begriff *multiple-choice* als übergeordnetes Suchwort waren zahlreiche spezifische deutsche und englische Schlüsselwörter Ausgangspunkt der Anfragen (z. B. *guessing*, *testwiseness*). Basierend auf vorliegenden Quellen zu einzelnen Themenbereichen haben wir dann im Schneeballsystem einschlägige Arbeiten möglichst vollständig zusammengetragen. Neben Artikeln aus Fachzeitschriften haben wir publizierte Konferenzbeiträge und Bücher berücksichtigt. Während wir für dieses Review über 450 Arbeiten gesichtet haben, beschränkt sich die Auswahl zitierter Quellen auf von uns besonders relevant und hochwertig eingeschätzte Arbeiten. In Teilbereichen war die Auswahl jedoch begrenzt.

3 Sind MC- und CR-Klausuren diagnostisch gleichwertig?

Durch eine systematische Betrachtung der Stärken und Schwächen von MC- und CR-Prüfungen zeigen wir im Folgenden ihr diagnostisches Potenzial auf und leiten mit Blick auf den Hochschulkontext weiteren Forschungsbedarf für eine zufriedenstellende Antwort auf die umstrittene Frage der Äquivalenz von MCA und CR-Aufgaben (CRA) ab.

3.1 Formale Charakteristika

MCA setzen sich mindestens aus zwei Elementen zusammen: (1) einer Problem- bzw. Fragestellung (*Aufgabenstamm*) und (2) mehreren vorgegebenen Lösungsmöglichkeiten (*Alternativen* oder *Optionen*). Je nach MC-Format (vgl. Abschnitt 6) umfasst die Liste der Alternativen eine oder mehrere richtige Antworten und verschiedene inkorrekte Optionen (*Distraktoren*), die von den Prüflingen als richtig bzw. falsch identifiziert werden müssen. Im Vergleich dazu bestehen CRA lediglich aus einem Aufgabenstamm, der mit einer Frage oder konkreten Aufgabenstellung abschließt. Das CR-Format verlangt von dem Prüfling somit eine eigenständige Konstruktion der Antwort, üblicherweise in Form eines Freitextes.

3.2 Ökonomie und Psychometrie

Der Vorteil von MCA ist, dass die Schreibarbeit der Prüflinge gegenüber CRA entfällt, weil diese lediglich Kreuze setzen. Damit reduziert sich gleichzeitig eine Quelle konstrukt-irrelevanter Varianz (Haladyna & Downing, 2004), da die Schreibleistung bei CRA der zu messenden Leistung oft nicht zuzuordnen ist. Die eingesparte Zeit erlaubt eine höhere MC-Aufgabenanzahl in der Prüfung (Burton, 2001; Wainer & Thissen, 1993), wodurch ein breites Spektrum der behandelten Inhalte abgeprüft werden kann. Somit liefern MCA typischerweise deutlich mehr diagnostische Information pro Zeiteinheit als CRA (Wan & Henly, 2012). Neben dem daraus resultierenden testtheoretischen Vorteil im Hinblick auf die Reliabilität (vgl. Spearman-Brown-Formel) und Konstruktvalidität (vgl. Rost, 2004), liegt der offensichtliche Vorteil von MCA in dem geringen Auswertungsaufwand mit maximaler Auswertungsobjektivität (Haladyna, 2004; Martinez, 1999; Simkin & Kuechler, 2005). Während offene Antworten einer zeit- und kostenintensiven Kategorisierung durch Prüfende unterzogen werden müssen, von der Bewertungsfehler und Verzerrungen z. B. aufgrund der Lesbarkeit der Handschrift, Eloquenz oder personenbezogener Erwartungseffekte ausgehen können (Dunbar, Koretz & Hoover, 1991; Hollingworth, Beard & Proctor, 2007), kann bei entsprechender technischer Ausstattung das MC-Format vollständig automatisiert ausgewertet werden. Dieser Vorteil fällt für Lehrende umso stärker ins Gewicht, je größer die zu prüfende Gruppe ist. Eine zügige Leistungsrückmeldung kommt aber vor allem den Prüflingen zugute, die zusätzliche Lernzeit im Falle des Nichtbestehens gewinnen.

Wenngleich es zunehmend bessere Softwareansätze zur automatischen Bewertung von CRA gibt (vgl. Bridgeman, Trapani & Attali, 2012; Shermis & Burstein, 2002), sind diese noch nicht in der Praxis angekommen und stehen

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

hauptsächlich für eine Auswertung basierend auf Schlüsselworten zur Verfügung; der Sinngehalt von Geschriebenem kann nicht erfasst und bewertet werden, sobald von diesen Schlüsselworten abgewichen wird. Beim gegenwärtigen Standard einer manuellen Bewertung von CRA lässt sich daher im Hinblick auf Ökonomie und damit einhergehende psychometrische Vorteile (v. a. Objektivität und Reliabilität) ein breiter Konsens in der Literatur identifizieren, der MC-Klausuren Überlegenheit gegenüber CR-Klausuren bescheinigt. Wie hoch dieser ökonomische Vorteil in der Praxis tatsächlich ausfällt, bleibt empirisch zu klären. Hier wären Studien zur Kosten-Nutzen-Abwägung wünschenswert, die vor allem Personalkosten für den gesamten Prüfungsprozess (Konstruktion, Darbietung, Auswertung und Wiederverwertung der Aufgaben) in realen Hochschulkontexten erfassen und diese Kosten der diagnostischen Qualität (z. B. operationalisiert durch psychometrische Parameter, Parallelmessungen in anderen Prüfungsformaten und externe Validitätshinweise) gegenüberstellen.

3.3 Befunde zur Messäquivalenz

Auf die Frage, ob MCA und CRA eine vergleichbare inhaltliche Validität aufweisen und tiefes fachliches Verständnis sensitiv erfassen können, gibt die Literatur keine klare Antwort. Das Potenzial der Formate, ein bestimmtes Konstrukt mit gleicher diagnostischer Qualität zu erfassen, bezeichnen wir im Folgenden als *Messäquivalenz*. Skepsis gegenüber der Messäquivalenz von MCA und CRA (z. B. Hancock, 1994; Lane, 2004; Nickerson, 1989) ist unter anderem darauf zurückzuführen, dass die Wiedererkennung einer richtigen Antwort (*recognition*) im Allgemeinen andere, weniger komplexe kognitive Prozesse als die eigenständige Reproduktion des Wissens (*free recall*) erfordert (vgl. Anderson & Bower, 1972; Kintsch, 1970). Mit dieser Annahme konform erweisen sich MCA gegenüber CRA in einigen Studien zwar als leichter, doch das Ausmaß der Effekte variiert mit dem verwendeten MC-Format (z. B. Bonner, 2013; Hohensinn & Kubinger, 2011; Liu, Lee & Linn, 2011). Zu bemerken ist auch, dass vornehmlich leistungsschwächere Studierende Prüfungen mit MCA bevorzugen, während leistungsstarke Studierende CRA und andere Formate favorisieren (Birenbaum & Feldman, 1998). Dies könnte darauf zurückgehen, dass MCA in Randbereichen der Kompetenz schlechter differenzieren als CRA (Lee, Liu & Linn, 2011). So werden Personen im unteren Leistungsbereich durch die Möglichkeit, richtige Antworten zu erraten, wohl eher zu vorteilhaft bewertet, während leistungsstarke Studierende ihr Potenzial nicht vollständig zeigen können. Ad hoc eignen sich CRA (oder z. B. mündliche Prüfungen) somit besser als MCA für eine genaue diagnostische Aussage in den Extrembereichen der Leistungsfähigkeit (Lee et al., 2011; Rauch & Hartig, 2010).

Auch jenseits der Messeigenschaften in Randbereichen kommen Studien, die das allgemeine Potenzial der Messäquivalenz von MCA und CRA untersuchen, zu keinem eindeutigen Ergebnis (vgl. z. B. Bennett, Rock & Wang, 1991; Hancock, 1994; Kastner & Stangl, 2011; Lee et al., 2011; Martinez, 1999; Rauch & Hartig, 2010; Traub & Fisher, 1977; Wan & Henly, 2012). Daraus wird deutlich, dass bei der Interpretation von Ergebnissen zur Messäquivalenz der Formate viele potenziell einflussreiche Kontextfaktoren (z. B. fachliche Domäne, angestrebter Messbereich, diagnostische Zielsetzung, Auswertungsprozeduren) sowie die große Heterogenität *innerhalb* der Formate (z. B. inhaltliche Gestaltung, MC-Formatwahl, Aufgabenqualität) stärker berücksichtigt werden müssen (vgl. auch Martinez, 1999; Simkin & Kuechler, 2005). Ein großes Problem bei der Aufklärung der Effekte dieser Rahmenbedingungen besteht darin, dass die Berichterstattung vieler publizierter Arbeiten mangelhaft ist. Beispielsweise werden verwendete Aufgaben nur oberflächlich beschrieben oder Kreditierungsregeln gar nicht oder zu unspezifisch benannt. Kastner und Stangl (2011) zeigen, dass die Vergleichbarkeit der Messergebnisse stammäquivalenter MCA und CRA durchaus von der Kreditierung der Formate abhängt, was die Bedeutsamkeit einer aussagekräftigen Berichterstattung nochmals hervorhebt.

In der groß angelegten Literatur-Analyse von Rodriguez (2003) findet sich der Versuch, die Heterogenität der Studienlage zu systematisieren und Ergebnisse zur Messäquivalenz von MCA und CRA aus 67 Studien (29 Korrelationsstudien) zu integrieren. Als wichtige Dimensionen zur Aufklärung heterogener Befundmuster wurden dabei (1) die Übereinstimmung des Aufgabenstamms, (2) die erwartete Ausführlichkeit der Antwort auf die CRA (kurz vs. lang) und (3) die inhaltliche Äquivalenz der verglichenen MCA und CRA identifiziert. Die Analyse ergab, dass Aufgaben mit gleichem Aufgabenstamm und kurzer erwarteter Antwort im direkten Vergleich sehr ähnliche Messeigenschaften aufweisen und die (korrigierte) Korrelation der Testwerte hoch bis sehr hoch ($r > .90$) ausfällt. Korrelationen zwischen MCA und CRA, die nicht stammäquivalent sind, sich im Inhalt leicht unterscheiden oder lange Antworten im Essay-Format vorsehen, waren geringer, aber substanziell ($r = .66–.95$). Rodriguez (2003) schlussfolgert aus diesen Ergebnissen, dass Konstruktäquivalenz von MCA und CRA unter anderem von der Intention des Konstrukteurs abhängt: Möchte man, dass MCA und CRA dasselbe Konstrukt messen, ist dies durch entsprechende Gestaltung der Aufgaben möglich.

3.4 Messäquivalenz und Messbereich

Ein wichtiger Faktor für die Messäquivalenz von MCA und CRA liegt in den Eigenschaften des angestrebten Messbereichs. So kann insbesondere *Wissen* im Sinne der Taxonomie von Bloom, Engelhart, Furst, Hill und Krathwohl (1956) durch MCA optimal erfasst werden (Haladyna, 2004). In diesem Bereich gelten die Formate diagnostisch als mindestens gleichwertig bzw. MCA aufgrund ihrer Objektivität und Ökonomie tendenziell als überlegen

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

(Haladyna & Rodriguez, 2013). Wohl auch deshalb werden in vielen MC-Prüfungen überwiegend Reproduktionsaufgaben eingesetzt (vgl. Tarrant, Knierim, Hayes & Ware, 2006), die abgesehen vom Memorieren und Identifizieren gelernter Fakten nur geringe kognitive Anforderungen stellen. Möglicherweise speist diese häufige Verwendung anspruchsarmer Aufgaben generelle Zweifel an der Eignung von MCA, komplexe Wissensstrukturen und Kompetenzen abzuprüfen (z. B. Kubinger, 2014; Nickerson, 1989). Viele Experten vertreten demgegenüber die Meinung, dass eine sorgfältige und kreative Konstruktion von MCA es durchaus ermöglicht, höhere Denkprozesse anzuregen und entsprechend Leistungen wie *Verständnis*, *Interpretation* oder *Wissensanwendung* (vgl. Bloom et al., 1956) zu erfassen, die über das bloße Wiedererkennen von Informationen hinausgehen (z. B. Bible, Simkin & Kuechler, 2008; Haladyna, 1997; Haladyna, Downing & Rodriguez, 2002; McCoubrie, 2004; Simkin & Kuechler, 2005). So sollte auch bezogen auf komplexere Leistungen Messäquivalenz von MCA und CRA erreichbar sein, wenngleich mehr Kreativität und ein höherer Konstruktionsaufwand erforderlich sind: Praxisnahe Fallvignetten, Matching-Formate, kontext-abhängige Item-Sets oder figurale Antwortformate sind Beispiele für anspruchsvollere MCA-Gestaltungsmöglichkeiten (vgl. Case & Swanson, 2002; Haladyna & Rodriguez, 2013; vgl. Abschnitt 6). Dass es zur Messung höherer kognitiver Fähigkeiten nicht auf das Format per se ankommt, sondern vor allem auf die investierte Gestaltungsleistung (Rodriguez, 2002), zeigt sich beispielsweise im Kontext von Intelligenztests, die figurale Matrizenaufgaben im MC-Format nutzen (z. B. Ravens Standard Progressive Matrices [SPM]; Raven, Raven & Court, 1998). In diesem Kontext wird kaum angezweifelt, dass MCA höhere kognitive Prozesse anregen und messen können. Eine klare Grenze von MCA liegt hingegen in der Erfassung von kreativen und schöpferischen Leistungen (z. B. das Schreiben einer Geschichte), während CRA an dieser Stelle keineswegs an Grenzen stoßen (Haladyna, 1997).

Bei Erwägung des Einsatzes von MCA in sozial- und geisteswissenschaftlichen Hochschulfächern ist diese Beschränkung zu beachten, da zu erbringende Leistungen im Vergleich zu naturwissenschaftlichem Faktenwissen häufiger schöpfungsorientiert sind. Die Dominanz von MCA in naturwissenschaftlichen Fächern (v. a. Medizin) mag auch einer besseren Passung von Fachinhalten und Format geschuldet sein; wir sehen aber keine grundsätzlichen Hürden für den Einsatz von MCA in geisteswissenschaftlichen Fächern. Einschlägige empirische Befunde zur Tauglichkeit von MCA in verschiedenen Fachdomänen fehlen bislang.

3.5 Die Rolle der Aufgabenqualität

Der wichtigste Faktor für die diagnostische Wertigkeit von MC-Prüfungen gegenüber äquivalenten CR-Prüfungen ist die faktische Qualität eingesetzter MCA. Dies ist unmittelbar an die Fähigkeit des Prüfenden, hochwertige MCA zu gestalten, gekoppelt. So erfordert die MC-Erstellung fachliche Expertise, ein Mindestmaß psychometrischer Kenntnisse, spezifisches Wissen über Gestaltungsrichtlinien sowie Feingefühl und Kreativität (vgl. Haladyna et al., 2002; Haladyna, 2004). Bei derart hohen Anforderungen ist es nicht verwunderlich, dass die Qualität von MCA oft mangelhaft ist. Gerade unerfahrene Konstruierende machen Fehler, die massiven Einfluss auf die psychometrische Güte der MCA haben können (vgl. Downing, 2002a, 2002b, 2005; Tarrant et al., 2006; Tarrant & Ware, 2008), beispielsweise durch unbedachte sprachliche Hinweise auf die richtige Lösung, die die Ratewahrscheinlichkeit deutlich erhöhen (vgl. Abschnitte 4 und 5). Die Kenntnis und Anwendung von mittlerweile gut ausgearbeiteten und zu großen Teilen empirisch belegten MC-Konstruktionsregeln, den *item writing guidelines*, die erstmals von Haladyna et al. (2002) expliziert wurden, kann dabei nachweislich ein hilfreicher Leitfaden sein, um die Qualität von MCA signifikant zu erhöhen (Jozefowicz et al., 2002; Wallach, Crespo, Holtzman, Galbraith & Swanson, 2006). Fraglich ist, inwieweit diese Richtlinien an deutschen Hochschulen bekannt sind und wenn, ob sie unter vorherrschendem Zeit- und Kostendruck im Alltagskontext tatsächlich Akzeptanz und Anwendung finden. Wir sehen in einer Analyse der Prüfungspraxis bzw. der typischen Qualität von MC- und CR-Klausuren an Hochschulen ein elementares und vernachlässigtes Forschungsfeld, das für eine befriedigende Klärung der Äquivalenz von MCA und CRA dringend zu bearbeiten ist.

3.6 Fazit

Insgesamt gibt die Literatur Anlass zu Optimismus bezüglich des Potenzials von MCA sich unter Einhaltung wissenschaftlicher Konstruktionsstandards in der Hochschulpraxis zu etablieren bzw. fallabhängig gegenüber CRA durchzusetzen. Für verlässliche Aussagen zur Messäquivalenz der Formate unter verschiedenen Rahmenbedingungen (z. B. Domäne, Messbereich, diagnostische Zielstellung, Kreditierung) besteht aber weiterhin erheblicher Forschungsbedarf. Beispielsweise wurden viele Studien an Schulen und vorwiegend mit mathematisch-naturwissenschaftlichen Aufgaben durchgeführt, wodurch die Übertragbarkeit auf komplexe Inhalte an Hochschulen und andere Fächer unsicher ist. Weitere Forschungsanstrengungen sollten sich auf kontrollierte Feld- oder experimentelle Studien – vor allem an Hochschulen – mit hochvergleichbaren MCA und CRA konzentrieren, um Rahmenfaktoren systematisch und differenziert zu untersuchen.

4 Wie groß ist «das Rateproblem» bei MC-Prüfungen?

Eine mit der Vorgabe von Antwortoptionen verbundene und in der Literatur vieldiskutierte Eigenschaft von MCA

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

liegt in der Möglichkeit, die richtige Lösung vollständig oder teilweise im Ausschlussprinzip zu erschließen oder sie zu erraten (z. B. Angoff, 1989; Budescu & Bar-Hillel, 1993; Bush, 2015; Burton, 2001; Dirkzwager, 2003; Espinosa & Gardeazabal, 2010; Frary, 1988, 1989; Grosse & Wright, 1985; Haladyna & Rodriguez, 2013). Rateeffekte haben bei unzureichenden Vorkehrungen einen negativen Einfluss auf die Reliabilität, Diskrimination und Validität von MCA (Zimmerman & Williams, 2003). Durch attraktive Distraktoren, viele hochwertige Aufgaben, spezielle Scoring-Systeme und an die Ratewahrscheinlichkeit angepasste Bestehensgrenzen lassen sich Rateeffekte aber relativieren (z. B. Ben-Simon, Budescu & Nevo, 1997; Frary, 1988; Haladyna, 2004; Haladyna et al., 2002). Diese Vorkehrungen teilen sich in (1) Maßnahmen im Rahmen der Aufgabenkonstruktion und (2) Maßnahmen im Rahmen der Kreditierung.

4.1 Maßnahmen im Rahmen der Aufgabenkonstruktion

Die **Auswahl geeigneter Distraktoren** ist die wichtigste Voraussetzung für eine funktionierende MCA und eine faktische Begrenzung auf die a priori Ratewahrscheinlichkeit. Ein Distraktor ist geeignet, wenn er Personen ohne erforderliches Wissen plausibel erscheint, von Personen mit erforderlichen Fähigkeiten und Kenntnissen jedoch als falsch erkannt wird. Er ist ungeeignet, wenn er auch ohne erwartetes Wissen leicht als Distraktor erkannt und ausgeschlossen werden kann. In der Praxis sollte jeder Distraktor zumindest von einem Teil der Personen (> 5 %; vgl. Tarrant, Ware & Mohammed, 2009) tatsächlich gewählt werden. Owens, Hanna und Coppedge (1970) beschreiben drei Methoden, um Distraktoren zu konstruieren: (1) Bei der *beurteilenden Methode* erfindet der Testkonstrukteur die Distraktoren. Diese orientieren sich z. B. an typischen Fehlvorstellungen von Studierenden. (2) Die *Häufigkeitsmethode* ist ein empirisches Verfahren, bei dem zuerst ein Test mit offenen Antworten durchgeführt wird. Distraktoren werden anschließend aus den häufigsten falschen Antworten entwickelt. (3) Die *Diskriminationsmethode* basiert ebenfalls auf einer Pilotierung der Aufgaben im offenen Format. Hier werden Distraktoren aus den falschen Antworten generiert, die am besten zwischen Personen mit guter und schlechter Leistung diskriminieren (hohe negative Korrelation mit dem Testwert). Owens et al. (1970) fanden keine Unterschiede in der Testgüte in Abhängigkeit der drei Methoden, wogegen Hanna und Johnson (1978) eine geringere Reliabilität für die Häufigkeitsmethode fanden. Aktuellere Studien zu dieser Thematik fehlen. Eine Arbeitserleichterung in der näheren Zukunft könnten Softwarepakete bieten, die basierend auf Textmaterial verschiedene MC-Optionen anhand linguistischer Algorithmen vorschlagen und so einen guten Ausgangspunkt für die Aufgabenkonstruktion schaffen (Mitkov, Ha & Karamanis, 2006).

Weiterhin hat die **Anzahl der Antwortoptionen** einen erheblichen Einfluss auf die a priori Ratewahrscheinlichkeit. In der Praxis findet man häufig vier oder fünf Optionen (Delgado & Prieto, 1998; Haladyna & Downing, 1993). Kubinger (2014) rät, möglichst viele Antwortoptionen (z. B. x aus 5 oder 7) zu verwenden, um die a priori Ratewahrscheinlichkeit zu minimieren. Auch wenn dies in manchen Situationen sinnvoll sein kann, weisen Studien konsistent darauf hin, dass Items mit drei Antwortalternativen aus verschiedenen Gründen vorzuziehen sind (z. B. Baghaei & Amrahi, 2011; Bruno & Dirkzwager, 1995; Lord, 1977; Rodriguez, 2005; Tarrant et al., 2009; Trevisan, Sax & Michael, 1991, 1994; Vyas & Supe, 2008; Yaman, 2011). Die oft schlechte Distraktorqualität ist einer dieser Gründe: So fanden Haladyna und Downing (1993), dass die meisten Aufgaben nur einen bis höchstens zwei sinnvolle Distraktoren enthalten und viele Themen tatsächlich kaum mehr erlauben. Weiterhin konnten Levine und Drasgow (1983) zeigen, dass fähigere Personen im Regelfall nur wenige Distraktoren tatsächlich wählen. Selten gewählte Distraktoren profitieren selbst dann nicht, wenn man den besten Distraktor eliminiert (Shizuka, Takeuchi, Yashima & Yoshizawa, 2006). Der Entwurf plausibler falscher Antworten stellt zudem die größte und zeitaufwendigste Herausforderung bei der MC-Konstruktion dar (Haladyna, 2004; Haladyna & Downing, 1993). Wird die Anzahl der Antwortoptionen reduziert, kann die gesparte Zeit, z. B. für eine intensivere Auseinandersetzung mit der Qualität der anderen Distraktoren, genutzt werden (Haladyna & Downing, 1989).

Die Abnahme der Optionsanzahl je Item bedingt zwar allgemein eine geringere Reliabilität, allerdings zeigte die Metaanalyse von Rodriguez (2005), dass dies im Fall einer Reduktion von vier auf drei Optionen nicht zuzutreffen scheint; hier stieg die Reliabilität sogar an. Dies deckt sich mit Erkenntnissen von Grier (1975), wonach Reliabilitätsverluste aufgrund von weniger Antwortoptionen durch eine höhere mögliche Gesamtaufgabenanzahl in gleicher Prüfungszeit ausgeglichen werden können (vgl. auch Haladyna & Downing, 1993; Vyas & Supe, 2008). Auch die Abdeckung eines breiteren Spektrums der inhaltlichen Thematik wird begünstigt (Trevisan et al., 1994; Vyas & Supe, 2008), was zu Befunden der Metaanalyse von Rodriguez (2005) passt, der keine Beeinträchtigung der inhaltlichen Validität durch eine Reduktion der Antwortalternativen (von fünf oder vier auf drei) fand. Neben ökonomischen Vorteilen und empirischen Befunden gibt es auch theoretische Arbeiten, die für eine Anwendung des Drei-Optionen-Formats sprechen (Bruno & Dirkzwager, 1995; Lord, 1977; Tversky, 1964).

Kritisch zu bedenken ist, dass Konstruierende tatsächlich in der Lage sein müssen, funktionierende von nicht funktionierenden Distraktoren zu unterscheiden, um bei wenigen Antwortoptionen eine gute Aufgabe zu erhalten. Selbst wenn Fachexperten scheinbar ausreichend gut in der Lage dazu sind (Cizek & O'Day, 1994; Swanson, Holtzman & Allbee, 2008), ist die Distraktorattraktivität nicht

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

immer leicht zu beurteilen. Somit ist es auch bei der Verwendung von drei Optionen ratsam, zunächst möglichst viele passende Distraktoren zu suchen und ihre Trennschärfe und Attraktivität im Rahmen einer Pilotierung zu untersuchen (vgl. Haladyna et al., 2002). Die Funktionalität von Distraktoren kann empirisch geprüft werden, indem (1) die tatsächliche Wahlhäufigkeit sowie (2) Korrelationen zwischen allen Antwortoptionen und dem Testscore betrachtet werden. Diese Koeffizienten sollten für Distraktoren negativ und für die richtige Antwort positiv ausfallen.

Für die Praxis ist zudem entscheidend, dass bei drei Antwortalternativen tatsächlich eine **höhere Aufgabenanzahl** realisiert wird, da viele der Befunde zur Favorisierung von drei Optionen auf dieser Annahme beruhen. Somit sollte jede MC-Prüfung über ausreichend viele MCA verfügen, um die Ratewahrscheinlichkeit für die Prüfung als Ganzes zu verringern. Während eine Aussage zur Mindestanzahl von MCA schwer abzustecken ist, weist Burton (2006) darauf hin, dass die in der Hochschulpraxis häufige Größenordnung von ca. 30 Aufgaben keinesfalls angemessen ist und eher 100 MCA und mehr für eine stabile Messung notwendig sind. Die konkret angemessene Anzahl hängt dabei von dem MC-Format (vgl. Abschnitt 6) und dem Umfang des Curriculums ab. Eine Studie von Pamphlett (2005), in der die Anzahl von Aufgaben im Multiple-True-False-Format (True-False-Aussagen einzeln gescored) zwischen 100 und 300 Aufgaben systematisch manipuliert wurde, zeigt indessen, dass 100 Aufgaben für eine stabile Schätzung der Personenfähigkeit ausreichten und eine deutliche Erhöhung (auf bis zu 300 Aufgaben) die Schätzung kaum verbesserte. Mit Blick auf die Hochschulpraxis, in der Ressourcen zur Aufgabenkonstruktion von Prüfenden sowie die Klausurdauer deutlich begrenzt sind, werden selbst Aufgabenzahlen von 100 häufig nicht realisierbar sein. Umso wichtiger ist äußerste Sorgfalt bei der Konstruktion der Aufgaben.

4.2 Maßnahmen im Rahmen der Kreditierung

Die einfachste Methode der Bewertung von MCA ist die Summierung der Anzahl richtiger Antworten zu einem Rohwert (*Number-Right-Scoring; NR*), während viele andere Auswertungsmethoden vorgeschlagen wurden, beispielsweise zur Anerkennung von Teilwissen bzw. der Berücksichtigung der Sicherheit beim Antworten in Partial-Credit Formaten (z. B. *Confidence Based Testing, Multiple-Evaluation*; Ben-Simon et al., 1997; Bush, 2001; Dirkzwager, 2003) oder der nachträglichen Korrektur von Rateeffekten durch definierte Punktabzüge (z. B. *Formula-Scoring, Negative Marking*; Budescu & Bar-Hillel, 1993; Lord, 1975; Frary, 1988). Diese Bewertungssysteme werden allerdings aufgrund einer komplexen Handhabung und Interpretation seit Jahrzenten kontrovers diskutiert (z. B. Higham & Arnold, 2007; Lesage, Valcke & Sabbe, 2013; Lord, 1975): So stellen die Methoden explizite Annahmen zum Rateverhalten auf, die im Regelfall eine Anpassung der Instruktion für Studierende nach sich ziehen. Ob die zugrundeliegenden Rateannahmen tatsächlich zutreffen und inwiefern Prüflinge differenziell auf die spezifischen Instruktionen reagieren, bestimmt dabei entscheidend, wie valide, reliabel und fair die Testergebnisse sind (Budescu & Bar-Hillel, 1993). Beim Formula-Scoring werden Prüflinge beispielsweise vor Punktabzug für Rateverhalten (falsche Antworten) gewarnt, was neben der mathematischen Korrektur gleichermaßen dazu führen soll, dass faktisch weniger geraten wird. Abgesehen von juristischen Problemen mit Maluspunkten (vgl. Kubinger, 2014) verweist bereits Frary (1988) auf Probleme dieses Vorgehens, da Prüflinge aufgrund von Persönlichkeitseigenschaften sehr unterschiedlich auf die Anweisung nicht zu raten reagieren und sich systematische Unter- und Überschätzungen der tatsächlichen Leistungen ergeben (z. B. Higham & Arnold, 2007). So korrigiert Formula-Scoring nur für zufälliges, uninformiertes Raten, was insbesondere im Hochschulkontext nicht plausibel ist. Bei Ausschlussmöglichkeit nur einer Antwort lohnt es sich rechnerisch bereits, zwischen den übrigen Alternativen zu raten (Frary, 1988). Derartige «Regeln» erkennen jedoch nicht alle Studierenden oder lassen Aufgaben bei Unsicherheit dennoch lieber aus (Lesage et al., 2013). Diese personenbezogene Varianz in der Rateneigung kann zu systematischen Verzerrungen der Prüfungsergebnisse führen, die auf Persönlichkeitseigenschaften (z. B. Ängstlichkeit, Gewissenhaftigkeit) oder auf das Geschlecht zurückgehen. Auch wenn Befunde diesbezüglich inkonsistent sind, zeigen viele Studien, dass männliche Prüflinge bei Unwissen eher raten als weibliche, wodurch sie in MCA tendenziell besser abschneiden (Ben-Shakhar & Sinai, 1991; DeMars, 2000; Prieto & Delgado, 1999a, 1999b). Eine umfangreiche Analyse verschiedener Klassenstufen der Jahrgänge 1980–2000 eines breiten Testprogramms im US-Staat Iowa indizierte hingegen keinen relevanten Einfluss des Geschlechts auf die Neigung, Aufgaben auszulassen (Von Schrader & Ansley, 2006). Geschlechtsbezogene Rateeffekte waren indessen bei jüngeren, mit dem MC-Format unerfahrenen Schülern ausgeprägter.

Um personenbezogenen Ratetendenzen zu begegnen, sollten Prüfende unmissverständlich offenlegen, unter welchen Bedingungen es sich bei der verwendeten Scoring-Prozedur zu raten ‹lohnt› (Bar-Hillel, Budescu & Attali, 2005). Wenngleich es aus pädagogischer Sicht suboptimal ist, kann eine derartige Aufklärung zumindest Konstruktirrelevante Varianz verringern (vgl. Haladyna & Downing, 2004; Messick, 1989) und die Testfairness erhöhen (Zieky, 2006).

Im Rahmen der Item-Response-Theory (IRT; z. B. Embretson & Reise, 2000) bietet das Drei-Parameter-Modell (3PL), in dem neben dem Schwierigkeits- und Trennschärfe-Parameter ein Rateparameter modelliert wird, ebenfalls eine Korrekturmöglichkeit an. Die zuverlässige Schätzung des Rateparameters und damit verbunden von Personenparametern setzt aber eine erhebliche Anzahl an Prüflingen voraus, die in vielen Anwendungsfällen kaum vorliegen

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

werden. Auch ist die Gültigkeit des 3PL-Modells an weitere Voraussetzungen gebunden, die bei Anwendung im Hochschulbereich oft verletzt sein dürften (vgl. Embretson & Reise, 2000). Vor allem psychometrischen Laien empfehlen wir aber grundsätzlich von IRT-Skalierungen im Hochschulbereich abzusehen.

4.3 Fazit

Offensichtlich existieren Stellschrauben, um auf potenzielle Rateeffekte zu reagieren. Somit sehen wir im Raten für die Hochschulpraxis kein ernsthaftes Problem. Blindes Raten findet in high-stakes Situationen im Gegensatz zu rateverwandten, problematischen Prüfungsstrategien (vgl. Abschnitt 5.2) ohnehin selten statt. Zur Beschränkung von Rateeffekten bewerten wir Ansätze im Rahmen der Aufgabenkonstruktion als deutlich positiver gegenüber einer Verwendung komplexer Scoringsysteme, die für die Hochschulpraxis aufgrund (1) rechtlicher Einschränkungen der Kreditierung (vgl. Beaucamp & Buchholz, 2010; Kubinger, 2014), (2) unkalkulierbarer Konstrukt-irrelevanter Varianz und (3) hohem Anspruch an psychometrisches Wissen der Prüfenden wenig geeignet sind. Dagegen ist das NR-Scoring eine transparente und fairere Bewertungsstrategie; selbst wenn es nicht zur Minimierung von Rateeffekten beiträgt, macht es diese besser kalkulierbar. Als wichtigste Strategie sehen wir jedoch den Einsatz einer hohen Anzahl guter MCA (dabei: *Distraktorqualität* vor *Distraktorquantität*) in Kombination mit angemessenen Bestehensgrenzen, um die Rateeffekte für eine Gesamtprüfung im Rahmen zu halten. Auch der Rückgriff auf innovative MC-Formate kann lohnenswert sein (s. Abschnitt 6).

Die umfangreiche Forschungslage zum Raten, der Anzahl von Antwortoptionen und Kreditierungsregeln sticht vergleichsweise positiv heraus. In diesem Bereich sehen wir keine fundamentalen Erkenntnislücken; aktuellere Studien zu Konstruktionsprinzipien geeigneter Distraktoren sowie die Entwicklung rateresistenter Testaufgaben sind ungeachtet dessen erstrebenswert.

5 Beeinflusst der Einsatz von MCA Lern- und Prüfungsstrategien?

5.1 Prüfungsformat und Lernstrategien

Studien konnten wiederholt zeigen, dass das Format einer angekündigten Leistungsüberprüfung einen Einfluss auf das Verhalten von Lernenden bei der Prüfungsvorbereitung hat (z. B. Scouller, 1998). Daraus erwächst die Befürchtung, dass ‹MC-Prüfungen› auch zu ‹MC-Lernen› führen (vgl. Shepard, 2000) und die Aneignung von Wissen mehr auf die Wiedererkennung von Fakten als auf die Fähigkeit zur eigenständigen Reproduktion ausgelegt wird. In diesem Zusammenhang unterscheiden Struyven, Dochy und Janssens (2005) zwischen oberflächlichen (*surface learning*), tiefen (*deep learning*) und leistungsorientierten (*achieving approach*) Lernstrategien. Oberflächliche Strategien zeichnen sich durch eine selektive Aneignung prüfungsrelevanter Informationen unter Minimierung des persönlichen Einsatzes aus, wohingegen tiefe Lernstrategien den wünschenswerten Versuch des Lernenden bezeichnen, vertieftes Verständnis für die Materie zu entwickeln. Leistungsorientierte Strategien sind auf die Maximierung des Outcomes (der Note) ausgerichtet und damit vor allem situationsangepasst.

Die Wahl einer Lernstrategie ist, neben persönlichen Gewohnheiten, vor allem auch von der Wahrnehmung der Anforderungen der erwarteten Prüfungsform abhängig: So zeigen Untersuchungen, dass Studierende MCA eher als leichtes Prüfungsformat bewerten und bei der Vorbereitung auf MC-Prüfungen tatsächlich häufiger von oberflächlichen Lernstrategien Gebrauch machen, wohingegen Prüfungen mit CRA als schwieriger gelten und häufiger mit einer tiefen Lernstrategie assoziiert sind (z. B. McCoubrie, 2004; Scouller & Prosser, 1994; Scouller, 1998; Struyven et al., 2005). Diese Wahrnehmung bedingt auch, dass Studierende Prüfungen mit MCA vorziehen, da sie mit dem Format unter anderem eine größere Erfolgserwartung verbinden (Struyven et al., 2005; Zeidner, 1987). Ergebnisse von Scouller (1998) zeigen sogar, dass die Anwendung tiefer Lernstrategien zu schlechteren MC-Prüfungsleistungen führte. Dies erklären wir uns dadurch, dass Studierende mit fundiertem Wissen auch auf kleinste inhaltliche Fehler oder Ungereimtheiten in MCA aufmerksam werden und damit z. B. die (als richtig intendierte) Antwort – kritisch hinterfragt – nicht ganz korrekt oder zumindest diskutierbar ist. Oberflächlich vorbereitete Studierende stolpern über solche Feinheiten vermutlich seltener. Empirische Belege zur Stützung dieser Annahmen fehlen jedoch vollständig, was den Forschungsbedarf aufzeigt.

Die beste Lösung für die aufgeworfenen Probleme sehen wir in einer Kombination diagnostischer Strategien, durch Mischung von MCA und CRA in Klausuren (vgl. auch Wainer & Thissen, 1993), sodass Studierende im Unklaren bleiben, welche Inhalte durch welches Format geprüft werden. Während dieses Vorgehen in der Praxis häufig beobachtbar ist, gibt es unseres Wissens bislang keine Arbeiten, die den Einfluss der Formatkombination auf Lernstrategien von Studierenden untersuchen. Auch wären aufgrund ihrer ökologischen Validität lernbegleitende Fragebogen- oder Tagebuchstudien zur differenzierteren Analyse der Effekte von Prüfungsformaten auf Lernstrategien wünschenswert.

5.2 MC-Prüfungsstrategien und Testwiseness

Neben einer Lernstilveränderung besteht vor allem in high-stakes Prüfungen mit MCA die Gefahr, dass Eigenschaften des Formats strategisch ausgenutzt werden, um ein besseres Testergebnis zu erzielen. Diese Fähigkeit, in der sich Prüflinge vor allem aufgrund ihrer Testerfahrung

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

unterscheiden (z. B. Dodeen, 2008; Dolly & Williams, 1986), wird im englischen Sprachraum als *Testwiseness* bezeichnet und geht u. a. auf Arbeiten von Millman, Bishop und Ebel (1965) zurück. Nährboden für eine erfolgreiche Anwendung von Testwiseness-Strategien bieten vor allem nachlässig erstellte MCA, da Konstruktionsfehler typischerweise Rückschlüsse auf die korrekte Antwort zulassen oder helfen, falsche Optionen direkt auszuschließen (Case & Swanson, 2002; Martinez, 1999), wodurch sich die a priori Ratewahrscheinlichkeit unkalkulierbar erhöht. Da Testwiseness die Leistung unabhängig vom fachlichen Wissen systematisch verzerren kann, entsteht Konstrukt-irrelevante Varianz; Fairness, Reliabilität und Validität der Prüfung sind gefährdet (Cohen, 2006; Haladyna & Downing, 2004).

Ein Beispiel für die Relevanz solcher Lösungsstrategien bieten experimentelle Studien mit Matrizen- (Mittring & Rost, 2008) und Leseaufgaben (Rost & Sparfeldt, 2007; Sparfeldt, Kimmel, Löwenkamp, Steingräber & Rost, 2012), in denen Testpersonen angewiesen wurden, ausschließlich aufgrund der vorgegebenen Antwortoptionen einer Aufgabe die richtige Antwort auszuwählen. Während der Aufgabenstamm in diesen Settings also nie präsentiert wurde, lösten die Personen die Aufgaben dennoch überzufällig häufig. Dies verdeutlicht, dass MCA durch unerwünschte Strategien, wie einfache logische Schlüsse (z. B. Abzählstrategien in Matrizen) oder aufgrund der Konstellation und Gestaltung von Antwortoptionen teils viel leichter zu lösen sind als intendiert.

Neben der Attraktivität der Distraktoren (vgl. Abschnitt 4), gelten eine saubere Sprachwahl (z. B. Case & Swanson, 2002; Sarnacki, 1979), das Verhältnis der Antwortoptionen zueinander bzw. ihre Homogenität (z. B. Martínez, Moreno, Martín & Trigo, 2009) und die Positionierung der richtigen Antwort (z. B. Attali & Bar-Hillel, 2003) als besonders relevante Aspekte der Aufgabenkonstruktion im Zusammenhang mit Testwiseness-Strategien. So zeigt sich, dass Prüflingen mit hoher Aufmerksamkeit für sprachliche Korrektheit (bzw. Muttersprachlern) Vorteile aus falscher Grammatik und Rechtschreibung oder unangemessener Wortwahl in MCA entstehen können (Case & Swanson, 2002; Sarnacki, 1979). Weiterhin stellen extreme Worte, wie *‹alle›*, *‹immer›*, *‹nie›* vor allem in Distraktoren durch ihre selten angemessene, implizierte Absolutheit eine Lösungshilfe dar (Haladyna, et al., 2002; Tarrant & Ware, 2008). Prüflinge können auch von einem Vergleich der Antwortoptionen profitieren, da z. B. ein abweichender Satzbau oder Detailgrad manchmal Hinweise auf die richtige Antwort geben kann (Haladyna et al., 2002; Millman, et al., 1965). Lösungshinweise, die auf heterogene Antwortoptionen zurückgehen, bedürfen daher einer expliziten Prüfung (vgl. z. B. Tarrant et al., 2006). Distraktoren mit hoher Ähnlichkeit zur richtigen Antwort erweisen sich hingegen als vorteilhaft für die Itemdiskrimination (Martínez et al., 2009) und sorgen für eine leichte, wünschenswerte Erhöhung der Schwierigkeit (Ascalon, Meyers, Davis & Smits, 2007).

Auch die Position der Antwortoptionen kann als Lösungshinweis dienen, da es eine so starke implizite Tendenz von Prüfenden gibt, die richtige Antwort in der Mitte zu platzieren, dass bei Aufgaben mit richtiger Antwort in der Mitte sogar die Diskrimination beeinträchtigt wird (Attali & Bar-Hillel, 2003). Ginge es nach einem Konsens von 37 Autoren, den Haladyna und Downing (1989) dokumentieren, sollte die richtige Antwort auf jeder Position ausbalanciert platziert werden (*key balancing*). Das Ausbalancieren folgt jedoch impliziten oder expliziten Regeln (z. B. richtige Antwort nicht zu oft hintereinander an derselben Stelle), die sich von Prüflingen gleichermaßen ausnutzen lassen (Bar-Hillel et al., 2005). Eine zufällige Positionierung der richtigen Antwort (*key randomization*) erlaubt dagegen keine Antwortstrategien, kann automatisiert und ohne Rücksicht auf übrige Aufgaben umgesetzt werden. Bar-Hillel et al. (2005) fassen entsprechend zusammen, dass es keine Dimension gibt, auf der ein Ausbalancieren der Randomisierung überlegen ist.

Um versteckte Lösungshinweise in MCA aufzudecken, empfehlen Mittring und Rost (2008), eine Präsentation der Antworten ohne Aufgabenstamm auch in der Konstruktionspraxis zu nutzen. Besonders geeignet sind zudem ‹Lautes Denken Protokolle› (Ericsson & Simon, 1984), um im Rahmen der Testerprobung Aufschluss über Lösungswege und unerwünschte Strategien zu erhalten (Bonner, 2013; Cohen, 2006; Leighton, 2004; Reich, 2013). Wir halten es dabei für zielführend, diese Methode zukünftig auch zu Forschungszwecken viel häufiger einzusetzen, um Teststrategien und Lösungsprozesse im Allgemeinen besser zu verstehen und wissensunspezifisches Lösungsverhalten wirksamer abzuwenden. Ein aus unserer Sicht vielversprechender und wissenschaftlich stärker zu verfolgender Ansatz zur Reduktion des Erfolgs von Testwiseness- und Ratestrategien ist ein Einsatz von innovativen MC-Formaten, wie z. B. dem Discrete-Option-Format, das den simultanen Vergleich von Antworten durch eine sequenzielle Präsentation der Optionen unterbindet (Willing, Ostapczuk & Musch, 2014; vgl. Abschnitt 6).

Zur Begrenzung der Auswirkungen von Testwiseness sollten neben einer sorgsamen Aufgabenkonstruktion alle Prüflinge vor Verwendung von MCA mit dem Format vertraut gemacht werden, um zumindest Nachteile für Studierende zu reduzieren, die in ihrem Werdegang keine oder wenig MC-Erfahrung gesammelt haben (Sarnacki, 1979; Zieky, 2006). Beispielsweise könnten ‹unproblematische› Tipps zur Lösung von MC-Klausuren gegeben werden (z. B. Zeitmanagement, Rateempfehlung, Fehlervermeidung), wie sie bereits in der Arbeit von Millman et al. (1965) zu finden sind. Inwieweit eine Aufklärung von Prüflingen über Teststrategien tatsächlich helfen kann, die Fairness von MC-Prüfungen zu steigern oder ob ein solches Vorgehen kontraproduktiv ist, bleibt zu untersuchen. Auch ist unbekannt, welche Bedeutung Testwiseness-Effekte an deutschen Hochschulen haben, da die meisten Befunde internationaler Studien aus den USA stammen, wo MCA bereits im Schulsystem stetige Verwendung fin-

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

den und damit von einer sehr viel höheren Erfahrung der Prüflinge mit MCA auszugehen ist, die für die Entwicklung einer «Testwiseness-Kompetenz» als entscheidend gilt (Dodeen, 2008). Untersuchungen zur Testwiseness an deutschen Hochschulen könnten dabei einen wertvollen Beitrag für die Aufklärung der Relevanz des Problems hierzulande leisten.

5.3 Fazit

Studien finden deutliche Effekte des MC-Formats auf das Lernverhalten und Prüfungsstrategien von Studierenden. Wir möchten hervorheben, dass eine unerwünschte Anpassung von Lernstrategien insbesondere im Hochschulkontext ein gewichtiges Problem darstellt. In anderen Einsatzgebieten von MCA (z. B. Large-Scale-Assessment, Eignungsdiagnostik) steht allein die *Kompetenzmessung* im Fokus, nicht der *Kompetenzerwerb*. Dagegen sollen Hochschulprüfungen zum intensiven Lernen der Fachinhalte anregen, während die Leistungsmessung vergleichsweise sekundär ist. Es ist daher essenziell, die studentische Wahrnehmung von MCA als ‹leichtes Prüfungsformat› zu verändern. Dies gelingt nur durch konsequenten Einsatz anspruchsvoller MCA. Weiterhin ist zu berücksichtigen, dass Studierende im Hochschulkontext extrem motiviert sind, gut abzuschneiden; mit dem Versuch strategischer Testwertmaximierung ist damit immer zu rechnen. Beide Faktoren erhöhen die Anforderungen an die Aufgabenkonstruktion erheblich.

Als wichtigste Gegenmaßnahme unerwünschter Lern- und Prüfungsstrategien empfehlen wir daher (1) die Etablierung von (informellen) Qualitätssicherungssystemen an Hochschulen (z. B. Aufgabenkontrolle durch MC-erfahrene Lehrende und Fachkollegium; vgl. Downing & Haladyna, 1997; Haldyna & Rodriguez, 2013), die neben dem Ausschluss von Konstruktionsfehlern ein angemessenes Anforderungsniveau der Prüfungsaufgaben sicherstellen. Zudem ist (2) die Kombination von MCA und CRA in Hochschulprüfungen zu präferieren, um oberflächlichen Lernstrategien entgegenzuwirken.

6 Welche MC-Formate eignen sich für die Hochschulpraxis?

Neben prototypischen, konventionellen MCA gibt es zahlreiche MC-Varianten, die mehr Aufmerksamkeit verdienen. Im Folgenden beschreiben wir zunächst Charakteristika sowie Stärken und Schwächen der Formate und geben anschließend eine Empfehlung für die Hochschulpraxis. Um ihre Eignung für diesen Kontext zu bewerten, legen wir folgende Anforderungen zugrunde: Ein gutes MC-Format sollte (1) psychometrisch hochwertig sein, (2) resistent gegenüber Teststrategien und Raten, (3) ökonomisch in Konstruktion und Darbietung sowie (4) die Erfassung komplexeren Wissens erlauben. Zusätzlich berücksichtigen wir spezifische Vor- und Nachteile im Vergleich zu anderen MC-Formaten. Unsere Einschätzungen stützen sich einerseits auf wenige existierende Befunde einschlägiger empirischer Studien und ergänzen sich durch Hinweise aus der Literatur sowie eigene Schlussfolgerungen. Die Aussagen sind dabei nicht für alle Formate gleichermaßen umfangreich, da wir uns aus Platzgründen auf besonders zentrale Aspekte konzentrieren und zudem nicht für jedes Format hinreichende Informationen in der Literatur vorliegen. Die mangelnde empirische Befundlage zeigt dabei wichtige Forschungslücken auf, an denen verstärkt gearbeitet werden sollte, um MC-typische Probleme (vgl. Abschnitte 3 bis 5) zu mindern und das diagnostische Potenzial von MCA voll zu nutzen.

Unsere schematische Darstellung der MC-Formate in Abbildung 1 soll helfen, die differenzierenden Charakteristika leichter nachzuvollziehen. Jedes Format ist daher durch korrespondierende Buchstabenreferenzen (A-N) gekennzeichnet. Inhaltliche Beispiele für einige Formate sind in Arbeiten von Case und Swanson (2002), Haladyna (2004) sowie Haladyna und Rodriguez (2013) zu finden.

6.1 MC-Varianten im Überblick

MCA im **(A) konventionellen Format** (*Single-Choice*) sind am häufigsten in der Praxis vertreten (Haladyna & Rodriguez, 2013). Sie bestehen aus einem Stamm und mindestens drei Alternativen mit einer richtigen Antwort. Ein Spezialfall ist das **(B) Alternate-Choice-Format** (AC) mit nur zwei Antwortalternativen, von denen eine richtig und die andere falsch ist. Bei der komplexen konventionellen Variante, dem **(C) Multiple-Response-Format** (MR; *Multiple-Mark*; *Multiple-Multiple-Choice)*, kann dagegen eine beliebige Anzahl von Alternativen korrekt sein. Der Prüfling muss richtige von falschen Antworten trennen. MR-Items haben daher eine geringere a priori Ratewahrscheinlichkeit und sind schwieriger als konventionelle MCA. Sie können sogar ein ähnliches Schwierigkeitsniveau wie CRA aufweisen, was sie zu einer geeigneten Alternative macht (Hohensinn & Kubinger, 2011).

Beim **(D) True-False-Format** (TF; *Two-Choice*; *Binary Choice*) wird eine Aussage präsentiert, die Prüflinge als richtig (*true*) oder falsch (*false*) bewerten sollen. Es werden also keine konkurrierenden Alternativen angeboten, so wie es im klassischen oder Alternate-Choice-Format der Fall ist. Die Vorteile des TF-Fomats liegen in einer kurzen Lesezeit sowie schnellen Konstruktion und Auswertung (Haladyna, 2004). Problematisch ist neben einer sehr hohen Ratewahrscheinlichkeit die unterschiedliche Wirkung richtiger und falscher Aufgaben (Cronbach, 1942). Prüflinge haben beim Raten die Tendenz, Aussagen als wahr zu bewerten (Grosse & Wright, 1985). Tatsächlich falsche Aussagen werden dadurch reliabler als tatsächlich wahre Aussagen. Dieser unerwünschte Effekt kann sich in Abhängigkeit der Antworttendenz in Interaktion mit dem Item-Typ auch umkehren. Markieren Prüflinge eine Aus-

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

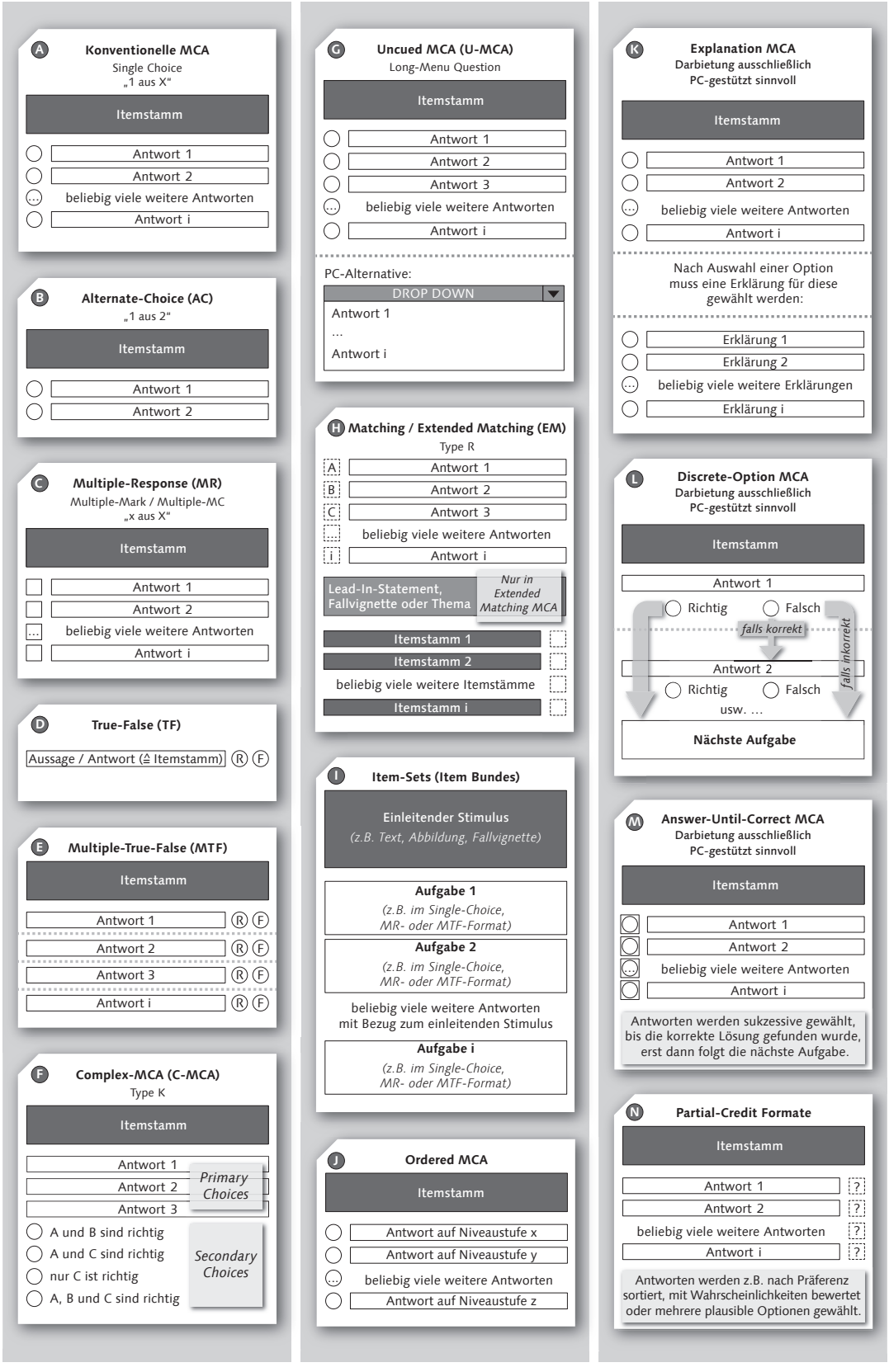

Abbildung 1. Schematische Übersicht verschiedener MC-Formate. Runde Antwortfelder repräsentieren eine auszuwählende Option (*1 aus x*), eckige Antwortfelder indizieren, dass potenziell mehrere Optionen (*x aus X*) gewählt werden können.

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

sage als falsch, bedeutet das zudem nicht, dass sie die korrekte Antwort kennen (Chandratilake, Davis, & Ponnamperuma, 2011; Hancock, Thiede, Sax & Michael, 1993). Gleichzeitig sind TF-Aufgaben weniger reliabel und trennscharf als korrespondierende AC-Items (Hancock et al., 1993; Oosterhof & Glasnapp, 1974). Empfehlenswerter ist aus unserer Sicht das **(E) Multiple-True-False-Format** (MTF; *Type X)* bei dem mehrere TF-Aussagen zu einem einleitenden Aufgabenstamm präsentiert werden (Albanese, Kent & Whitney, 1979). Der Vorteil gegenüber einfachen TF-Aufgaben liegt in der Scoring-Methode, insofern eine MTF-Aufgabe ausschließlich kreditiert wird, wenn der Prüfling *alle* zugehörigen Aussagen korrekt bewertet. So kann die Ratewahrscheinlichkeit, äquivalent zum MR-Format, effektiv gesenkt werden.

Eng verwandt mit MR und MTF ist das **(F) Complex MCA-Format** (C-MCA; *Type K*), bei dem ebenfalls mehrere Optionen korrekt sein können. Prüflinge entscheiden sich allerdings nicht unmittelbar für die Alternativen (*primary responses*), sondern wählen aus einer Liste von Antwortkombinationen (*secondary choices*) die korrekte Kombination aus (Albanese, 1993). Durch die doppelte Präsentation von Antworten benötigen C-MCA mehr Platz und Lesezeit als andere Formate, wodurch weniger Aufgaben bei gleicher Testzeit zu einer schlechteren Reliabilität führen (Haladyna, 2004). Äquivalente Aufgaben in anderen Formaten zeigen bessere psychometrische Kennwerte: So sind C-MCA schwieriger (Albanese, 1993; Tripp & Tollefson, 1985) und diskriminieren schlechter (Rodriguez, 1997). Problematisch sind auch Interdependenzen zwischen den Antworten, die als Hinweis auf die richtige Lösung genutzt werden können, da die Listen üblicherweise nicht alle denkbaren Antwortkombinationen enthalten (Albanese et al., 1979; Haladyna et al., 2002). Kann eine primäre Alternative sicher als richtig oder falsch klassifiziert werden, können sekundäre Antwortkombinationen leicht eliminiert und die Ratewahrscheinlichkeit deutlich erhöht werden. Von der Verwendung von C-MCA möchten wir abraten, da es neben vielen Problemen aus unserer Sicht keinen diagnostisch plausiblen Grund zugunsten einer Verwendung gibt.

(G) Uncued MCA (U-MCA; *Long-Menu-Question*) bedienen sich umfangreicher Antwortlisten anstatt einzelner Antwortalternativen. Durch bis zur Unüberschaubarkeit lange Listen (z. B. am PC durch Drop-Down Menüs umgesetzt) gibt es keine Hinweise (*cues*) auf die richtige Lösung (Schuwirth, van der Vleuten, Stoffers & Peperkamp, 1996). Um Aufgaben zu bearbeiten, muss der Prüfling die korrekte Antwort kennen und aktiv suchen. Somit sind eigenständige Rekonstruktionsleistungen erforderlich (Fajardo & Chan, 1993). Auch gegenüber blindem Raten und Testwiseness ist das Format resistenter. In der Studie von Fenderson, Damjanov, Robeson, Veloski und Rubin (1997) erwiesen sich U-MCA im Vergleich zu konventionellen MCA als reliabler. Allerdings ist das Format aufgrund des hohen Platzbedarfs und der Möglichkeit Filter- und Suchfunktionen einzusetzen nur am Computer ökonomisch anwendbar. Eine Studie von Schuwirth et al. (1996) zeigt, dass PC-gestützte U-MCA eine bessere Alternative zu CRA darstellen als konventionelle MCA. Bei Paper-Pencil U-MCA ist die Liste der Antworten dagegen natürlicherweise beschränkt. So kommt der Distraktorqualität wieder eine tragende Rolle zu, wodurch sich der Konstruktionsaufwand erheblich erhöht und Vorteile des Formats verloren gehen.

Bei **(H) Matching-Aufgaben** werden je ein Set aus Optionen und ein Set aus Stämmen (Aussagen, Entitäten o. ä.) dargeboten, wobei Prüflinge die Alternativen den Stämmen korrekt zuordnen müssen. Der Einsatz bietet sich besonders an, wenn die inhaltliche Fragestellung natürlicherweise eine Zuordnungssituation bietet (z. B. Symptome – Diagnosen, Jahreszahlen – Ereignisse usw.). Im Vergleich zu äquivalenten konventionellen MCA benötigt Matching weniger Platz und Lesezeit. Zimmerman und Williams (1982) konnten zeigen, dass Matching-MCA höhere Reliabilitäten und eine geringe Ratewahrscheinlichkeit aufweisen. Um Hinweise auf die richtige Antwort zu vermeiden, sollte eine ungleiche Anzahl von Stämmen und Alternativen verwendet werden, sodass sich durch Zuordnung der ersten Paare die Folgenden nicht automatisch ergeben (Haladyna, 2004). Eine komplexere Variante ist das **Extended Matching** (EM; *Type R*). In diesem Format werden nicht nur Stämme und Alternativen vorgegeben, sondern die Aufgabe wird durch ein Thema (*theme*) bzw. eine einleitende Aussage (*lead-in statement*) ergänzt (Case & Swanson, 2002). Eine solche Einleitung kann z. B. eine Frage, ein Szenario oder eine Fall-Vignette sein, um der Aufgabe einen Kontext zu geben. Um Rateeffekte zu minimieren, werden längere Antwortlisten eingesetzt. Bisherige Studien zu Matching-Formaten weisen auf eine gute Eignung der Aufgaben hin, insbesondere zur Erfassung komplexer Kompetenzen (Haladyna & Rodriguez, 2013). Die Ergebnisse zweier Studien von Swanson und Kollegen (Swanson, Holtzman, Albee & Clauser, 2006; Swanson et al., 2008) zeigen darüber hinaus, dass EM-Aufgaben durch Rückgriff mehrerer Aufgaben auf die gleiche Antwortliste konventionellen MCA hinsichtlich der Bearbeitungszeit überlegen sind und das Format mehr diagnostische Information pro Zeiteinheit liefert. Die Autoren plädieren beim EM-Format für eine höhere Anzahl an Aufgaben (mit z. B. acht Optionen) anstatt sehr lange Antwortlisten zu nutzen. Für Prüfungen mit EM-Items sind, verglichen mit konventionellen MCA, tendenziell weniger Aufgaben (ca. 60) für eine zuverlässige Schätzung der Leistung ausreichend (Case, Swanson & Ripkey, 1994). Eine einfach zu konstruierende Abwandlung des EM-Formats ist das **Modified-Matching**, bei dem alle richtigen Antworten eines Tests sowie einige Distraktoren auf einer separaten Seite stehen und den Testaufgaben korrekt zugeordnet werden müssen. Dieses von McAllister und Guidice (2012) vorgeschlagene Format ist sehr gut auf bereits bestehende MCA (v. a. konventionelle Formate) anwendbar und kann ohne großen Konstruktionsaufwand die blinde Ratewahrscheinlichkeit effektiv minimieren. In einer empirischen Untersuchung

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

der Autoren führte das Format vor allem zu einer höheren Messgenauigkeit in Randbereichen der Leistungsverteilung.

(I) Kontext-abhängige Item-Sets (*Item Bundles, Testlets, Super Items*) bestehen aus einem längeren einleitenden Stimulus (Abbildung, Text, Fallvignette o. ä.), gefolgt von mehreren Aufgaben oder Fragen in einem beliebigen MC-Format, die sich auf den gemeinsamen Stimulus beziehen. Item-Sets wird die Eigenschaft zugeschrieben, komplexe Denkprozesse zu messen (Haladyna & Rodriguez, 2013), diese Annahme ist jedoch nicht weiter empirisch belegt. Das Format ist besonders nützlich, um verschiedene Aspekte eines komplexeren Sachverhalts gemeinsam zu prüfen, wie z. B. das Verständnis einzelner Textteile im Rahmen eines Lesetests (Haladyna, 1992). Problematisch sind jedoch auf den gemeinsamen Stimulus zurückgehende, lokale Dependenzen der Aufgaben (z. B. DeMars, 2012; Eckes, 2014; Tuerlinckx & De Boeck, 2001): Versteht eine Person den Itemstamm beispielsweise nicht, wirkt sich dies negativ auf die Lösungswahrscheinlichkeit aller zugehörigen Aufgaben aus, wobei die ‹gemeinsame Ursache› in der Bewertung der Leistung nicht berücksichtigt wird. Solche Dependenzen können positiv oder negativ sein und treten umso eher auf, je mehr Aufgaben den gleichen Stimulus nutzen (Eckes, 2014). Testscores und Itemparameterschätzungen werden bedeutsam verzerrt. Da psychometrische Korrekturprozeduren (s. z. B. Wilson & Adams, 1995) aufwendig und eher im Rahmen von Large-Scale Studien von Relevanz sind, empfinden wir es als entscheidend für eine Anwendung des Formats im Hochschulkontext, zumindest inhaltliche Abhängigkeiten in den Item-Sets zu minimieren und nur eine begrenzte Aufgabenanzahl zu einem Stimulus darzubieten.

Eine in der Konstruktion begründete, differenziertere diagnostische Aussage zum Kenntnisstand eines Prüflings kann durch **(J) Ordered MCA** getroffen werden (Briggs, Alonzo, Schwab & Wilson, 2006). Hier werden die Antwortoptionen des konventionellen Formats so konstruiert, dass jede Antwort eine theoretisch definierte Niveaustufe konzeptuellen Verständnisses widerspiegelt. Das Format eignet sich jedoch nur für wenige Themenfelder und erfordert einen theoriegeleiteten, langwierigen Konstruktionsprozess. Unserer Einschätzung nach ist es damit zwar diagnostisch sehr interessant und für eine effiziente Messung komplexen Wissens attraktiv, doch für den Hochschulalltag wenig geeignet.

In den vergangenen Jahren wurden viele innovative Itemformate entwickelt, die auf einer computergestützten Darbietung des Testmaterials beruhen und neue Interaktionswege zwischen Prüfling und Testmaterial eröffnen (s. z. B. Boyle & Hutchison, 2009; Zenisky & Sireci, 2002). Hieraus ergibt sich vor allem die Möglichkeit einer sequenziellen Darbietung von Informationen. Beispielsweise bauen **(K) Explanation-MCA** darauf auf, dass Prüflinge nach Wahl einer Antwortoption aus einer nachträglich erscheinenden Liste eine Erklärung für Ihre Antwort auswählen, Die Aufgabe wird gemäß der Wertigkeit dieser Begründung kreditiert (Liu et al., 2011). Für den Hochschulkontext ist das Format aufgrund des hohen Konstruktionsaufwandes (Antworten und Begründungen) eher nicht geeignet. Vielversprechend ist dagegen das **(L) Discrete-Option MCA-Format**, bei dem eine sequenzielle Präsentation der Antwortoptionen einer konventionellen MCA so lange erfolgt, bis eine Option als «richtig» gewählt wird oder die tatsächlich richtige Option als «falsch» zurückgewiesen wird (Kingston, Tiemann, Miller & Foster, 2012). Der Erfolg von Testwiseness- und Ratestrategien kann so deutlich reduziert werden, da jede Antwort für sich genommen, ohne kontextuelle Hinweise bewertet werden muss (Willing et al., 2014).

Auch das **(M) Answer-Until-Correct MCA-Format** *(AUC*; Wilcox, 1981) kann in computergestützten Prüfungen sinnvoll eingesetzt werden: Hierbei sind alle Optionen einer konventionellen MCA sichtbar, wobei das folgende Item nur bei richtiger Antwort erscheint und andernfalls so lange gewählt wird, bis diese gefunden wurde (Muñiz & Menéndez, 2011). Die Bewertung kann dichotom erfolgen oder aber die Anzahl falscher Antworten berücksichtigen. Der entscheidende Vorteil des Formats liegt in dem direkten Feedback über die richtige Lösung, wodurch einer Entstehung falschen Wissens durch die Präsentation falscher Informationen in den Distraktoren vorgebeugt werden kann. Beispielsweise zeigen nämlich Roediger und Marsh (2005) im experimentellen Setting, dass Prüfungen mit MCA eine Lernsituation darstellen und die Erinnerung falscher Fakten eine realistische Gefahr ist, die vor allem durch umgehendes Feedback vermieden werden kann (s. auch Butler & Roediger, 2008).

Eine eigene Kategorie sind die variantenreichen **(N) Partial-Credit Formate**, die auf einer Abwandlung konventioneller MCA beruhen und Prüflingen eine differenziertere Rückmeldung zu einzelnen Antwortoptionen erlauben. Beispielsweise kann dies durch Sortierung der Optionen nach Präferenz (*Strict/Partial Ordering*), Zuschreibung von Wahrscheinlichkeiten der Richtigkeit aller Lösungen (*Confidence/Probability Weighting; Multiple Evaluation*), Auswahl mehrerer plausibel empfundener Optionen (*Subset Selection*) und Varianten davon erfolgen (s. auch Ben-Simon et al., 1997; Bush, 2015). Die so ausgedrückte Sicherheit bei der Beantwortung wird bei der Bewertung berücksichtigt (vgl. Lesage et al., 2013). Obwohl diese Formate die Rateproblematik gewissermaßen umgehen, ist ihre hohe Komplexität für Prüflinge und Auswertende ein Manko. So profitieren mit dem Format gut vertraute sowie leistungsstarke Studierende, die ihr Wissen angemessen einschätzen und ihr Antwortverhalten taktisch besser steuern (Ben-Simon et al., 1997). Die größere Variationsmöglichkeit beim Antworten kann dennoch angenehm für Prüflinge sein und eine detailliertere diagnostische Aussage zum Leistungsstand ermöglichen (Bush, 2015; Dirkzwager, 2003). Für informelle Leistungsrückmeldungen sind Partial-Credit Formate daher gut geeignet. In wichtigen Prüfungen empfiehlt sich der Einsatz unserer Ansicht nach nicht bzw. höchstens bei bester Vertrautheit aller Studierender mit dem Format.

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

6.2 Fazit

Zusammenfassend kommen wir zu der Einschätzung, dass für Papier-Bleistift-Prüfungen an Hochschulen vor allem die Formate MR (C) und MTF (E) geeignet sind. Beide können komplexeres Wissen erfassen, haben eine geringe a priori Ratewahrscheinlichkeit, sind leicht zu konstruieren und aufgrund geringer Interdependenzen der Antworten vergleichsweise resistent gegen Bearbeitungsstrategien. Um in Randbereichen besser zu differenzieren und für Zuordnungssituationen, sind Matching-Formate (H) eine hervorragende Alternative.

Nach Möglichkeit empfehlen wir insbesondere PC-gestützte Prüfungen durchzuführen, die vielversprechende MC-Varianten (vgl. 6.1, L und M) zulassen. Diese haben neben Ökonomie in Darbietung und Auswertung hohes Potenzial, Probleme wie z. B. Raten, Testwiseness oder geringe Aufgabenkomplexität zu überwinden. Dabei bleibt der Konstruktionsaufwand überschaubar, da es oft ausreicht, bestehende MCA in anderer Art und Weise zu präsentieren (vgl. 6.1, L).

Aufgrund der lückenhaften Befundlage sehen wir formatspezifische Analysen psychometrischer Eigenschaften, des erreichbaren Messbereichs, des faktischen Konstruktionsaufwands und der Akzeptanz bei Studierenden als zentrale Fragestellungen weiterer Forschung. Ein Vergleich verschiedener MC-Formate zu äquivalenten CRA und konventionellen MCA (vgl. Willing et al., 2014) ist wünschenswert. Ergänzend sehen wir großen Wert in kognitionspsychologischen Studien, die Lösungsstrategien und angeregte kognitive Prozesse verschiedener MC-Formate (auch vs. CRA) ins Visier nehmen und eine Abwägung auf tieferer Ebene, z. B. durch Protokolle Lauten Denkens (z. B. Leighton, 2004), ermöglichen.

7 Diskussion und Forschungsperspektiven

Im vorliegenden Literaturüberblick haben wir MCA hinsichtlich ihrer Messeigenschaften, diagnostischen Eignung für Prüfungen im Hochschulkontext und damit verbundene Probleme von verschiedenen Seiten beleuchtet. Es zeigt sich, dass die Bewertung vor allem von der faktischen Güte eingesetzter MCA abhängt: Nur sorgsam konstruierte Aufgaben können valide Informationen zur Leistungsbewertung liefern und so gegenüber anderen Prüfungsformaten (z. B. CRA) diagnostische Angemessenheit beanspruchen. Wie gut MCA neben einfacher Reproduktion komplexe Wissensstrukturen und Kompetenzen erfassen, ist dabei vor allem von Ressourcen und Konstruktionskompetenz einzelner Prüfender abhängig.

Hier sehen wir im Hochschulkontext ein zentrales Problem, welches bei der Einordnung der Befunde verglichen mit anderen Einsatzgebieten von MCA (z. B. im Large-Scale-Assessment-Kontext) berücksichtigt werden muss: Lehrende verfügen über geringe zeitliche und personelle Ressourcen für die Aufgabenerstellung. Gleichzeitig fehlen oft fundiertes psychometrisches Wissen, Kenntnis von MCA-Konstruktionsprinzipien und Erfahrung mit der (MC-)Testkonstruktion. Weiterhin sind zu messende Kompetenzen an Hochschulen vergleichsweise komplex und erfordern damit besondere Kreativität bei der Aufgabenerstellung. Erschwerend kommt hinzu, dass Studierende eine hohe Motivation haben, ihren Testwert (z. B. durch Prüfungsstrategien) zu maximieren und eventuelle Konstruktionsfehler auszunutzen. Wie bereits in Abschnitt 5.3 diskutiert, ist auch eine Anpassung von Lernstrategien an das MC-Testformat im Hochschulkontext sehr kritisch zu bedenken, da eine Anregung zum fundierten Wissenserwerb in der Regel primäres Ziel von Hochschulprüfungen ist, dem das Prüfungsformat keinesfalls entgegenwirken sollte.

Trotz eher schwierigen Rahmenfaktoren sehen wir dennoch das Potenzial in der Nutzung von MCA an Hochschulen und möchten nicht von einem Einsatz des Formats abraten. Wir haben vielmehr fünf Praxisempfehlungen aggregiert, die auch bei knappen Ressourcen zu diagnostisch vertretbaren Prüfungen führen sollten: Wir empfehlen, (1) MCA und CRA in Prüfungen gemischt darzubieten, (2) Kommentare der Prüflinge bei wahrgenommenen Fehlern in der Aufgabenstellung zuzulassen, (3) eine Qualitätssicherung der Aufgaben durch Checklisten und andere Lehrende durchzuführen sowie (4) einfach zu handhabende Formate (z. B. MR/MTF) und (5) Number-Right-Scoring mit ausreichend hohen Bestehensgrenzen einzusetzen. Sollte eine hinreichende Aufgabenqualität nicht sichergestellt sein, möchten wir von MCA in benoteten Prüfungen abraten, da unerwünschte Eigenschaften des Formats verstärkt zum Tragen kommen und den diagnostischen Wert in Frage stellen.

Vor allem die praxisorientierte Perspektive hat in unserem Literaturüberblick viele Forschungslücken aufgetan, die für belastbare Aussagen zur diagnostischen Wertigkeit von MCA unter «Alltagsbedingungen» im Hochschulkontext geschlossen werden sollten. So fehlen empirische Belege für wichtige Annahmen (wie z. B. zur ‹wahren› Ökonomie von MCA) sowie gut dokumentierte Studien, die differenzierte Aussagen zur Messäquivalenz von MCA und CRA unter verschiedenen Rahmenbedingungen erlauben (vgl. Abschnitt 3.6). Als positive Entwicklung des vergangenen Jahrzehnts ist die weithin gelungene Erforschung allgemeiner Regeln zur MC-Konstruktion hervorzuheben (vgl. Haladyna et al., 2002; Haladyna & Rodriguez, 2013). Während hier viele Befunde vorliegen, sehen wir einen nächsten Schritt darin, diese Regeln bei Lehrenden an deutschen Hochschulen großflächig bekannt zu machen. Durch begleitende Forschung wäre zu evaluieren, welche Qualität MCA in der Praxis erreichen (können) und welche Faktoren entscheidend für das Qualitätsniveau sind. Ein Fokus sollte darauf liegen, ob bestimmte Konstruktionsfehler in der Hochschulpraxis gegenüber besonders kritischen Fehlern verschmerzbar sind.

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

Vor allem für geistes- und sozialwissenschaftliche Studienfächer der Hochschule sehen wir weiteren Bedarf, die Tauglichkeit des MC-Formats zu evaluieren. Da MCA in vielen Bereichen bislang selten eingesetzt werden, stellt sich die Frage, an welche Grenzen MCA in diesen Domänen tatsächlich stoßen. Wir kommen zu der Überzeugung, dass vielfach nicht die inhaltliche Passung die seltene Nutzung des Formats bedingt, sondern möglicherweise eine pauschal kritische Einstellung gegenüber dem Messbereich von MCA zum Tragen kommt (vgl. Abschnitt 3.3). Eine vernachlässigte Perspektive richtet sich demnach auf die Entscheidung von Prüfenden, MCA einzusetzen oder nicht. Hier wäre es von Interesse, den Einfluss von Hintergrundvariablen (z. B. Studienfach, Hochschulpolitik) und persönlichen Charakteristika (z. B. psychometrische Kenntnisse, Einstellung gegenüber MCA) zu untersuchen. Das wichtigste langfristige Ziel sehen wir darin, praxistaugliche Wege aufzuzeigen, um unter begrenzten Ressourcen diagnostisch hochwertige MCA zu entwickeln. Dafür sollte vor allem der Computerisierung von Prüfungen und den damit verbundenen Freiheiten der MC-Formatgestaltung mehr Aufmerksamkeit geschenkt und das erhebliche Potenzial für Arbeitserleichterungen und höhere Qualitätsstandards von MCA genutzt werden.

Autorenhinweise

Die Autoren danken dem Hauptherausgeber Prof. Jörn Sparfeldt sowie drei anonymen Begutachtenden für ihre hilfreichen Anregungen zu diesem Manuskript.

Literatur

Albanese, M. A. (1993). Type K and other complex multiple-choice items: An analysis of research and item properties. *Educational Measurement: Issues and Practice, 12* (1), 28–33.

Albanese, M. A., Kent, T. H. & Whitney, D. R. (1979). Cluing in multiple-choice test items with combinations of correct responses. *Academic Medicine, 54*, 948–950.

Anderson, J. R. & Bower, G. H. (1972). Recognition and retrieval process in free recall. *Psychological Review, 79*, 97–132.

Angoff, W. H. (1989). Does guessing really help? *Journal of Educational Measurement, 26*, 323–336.

Ascalon, M. E., Meyers, L. S., Davis, B. W. & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education, 20*, 153–170.

Attali, Y. & Bar-Hillel, M. (2003). Guess where: The position of correct answers in multiple-choice test items as a psychometric variable. *Journal of Educational Measurement, 40*, 109–128.

Baghaei, P. & Amrahi, N. (2011). The effects of the number of options on the psychometric characteristics of multiple choice items. *Psychological Test and Assessment Modeling, 53*, 192–211.

Bar-Hillel, M., Budescu, D. & Attali, Y. (2005). Scoring and keying multiple choice tests: A case study in irrationality. *Mind & Society, 4*, 3–12.

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

Bennett, R. E., Rock, D. A. & Wang, M. (1991). Equivalence of free-response and multiple-choice items. *Journal of Educational Measurement, 28*, 77–92.

Ben-Shakhar, G. & Sinai, Y. (1991). Gender differences in multiple choice tests: The role of differential guessing tendencies. *Journal of Educational Measurement, 28*, 23–35.

Ben-Simon, A., Budescu, D. V. & Nevo, B. (1997). A comparative study of measures of partial knowledge in multiple-choice tests. *Applied Psychological Measurement, 21*, 65–88.

Bible, L., Simkin, M. G. & Kuechler, W. L. (2008). Using multiple-choice tests to evaluate students' understanding of accounting. *Accounting Education: An International Journal, 17*, 55–68.

Birenbaum, M. & Feldman, R. A. (1998). Relationships between learning patterns and attitudes towards two assessment formats. *Educational Research, 40*, 90–98.

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.

Bonner, S. M. (2013). Mathematics strategy use in solving test items in varied formats. *The Journal of Experimental Education, 81*, 409–428.

Boyle, A. & Hutchison, D. (2009). Sophisticated tasks in e-assessment: What are they and what are their benefits? *Assessment & Evaluation in Higher Education, 34*, 305–319.

Bridgeman, B., Trapani, C. & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*, 27–40.

Briggs, D. C., Alonzo, A. C., Schwab, C. & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment, 11*, 33–63.

Bruno, J. E. & Dirkzwager, A. (1995). Determining the optimal number of alternatives to a multiple-choice test item: An information theoretic perspective. *Educational and Psychological Measurement, 55*, 959–966.

Budescu, D. & Bar-Hillel, M. (1993). To guess or not to guess: A decision theoretic view of formula scoring. *Journal of Educational Measurement, 30*, 277–291.

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education, 26*, 41–50.

Burton, R. F. (2006). Sampling knowledge and understanding: How long should a test be? *Assessment & Evaluation in Higher Education, 31*, 569–582.

Bush, M. (2001). A multiple choice test that rewards partial knowledge. *Journal of Further and Higher Education, 25*, 157–163.

Bush, M. (2015). Reducing the need for guesswork in multiple-choice tests. *Assessment & Evaluation in Higher Education, 40*, 218–231.

Butler, A. C. & Roediger, H. L., III (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*, 604–616.

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.

Case, S. M., Swanson, D. B. & Ripkey, D. R. (1994). Comparison of items in five-option and extended-matching formats for assessment of diagnostic skills. *Academic Medicine, 69*, 1–3.

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

Chandratilake, M., Davis M. & Ponnamperuma G. (2011). Assessment of medical knowledge: The pros and cons of using true/false multiple choice questions. *The National Medical Journal of India, 24*, 225–228.

Cizek, G. J. & O'Day, D. M. (1994). Further investigation of nonfunctioning options in multiple-choice test items. *Educational and Psychological Measurement, 54*, 861–872.

Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly, 3*, 307–331.

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology, 33*, 401.

Delgado, A. R. & Prieto, G. (1998). Further evidence favoring three-option items in multiple-choice tests. *European Journal of Psychological Assessment, 14*, 197–201.

DeMars, C. E. (2000). Test stakes and item format interactions. *Applied Measurement in Education, 13*, 55–77.

DeMars, C. E. (2012). Confirming testlet effects. *Applied Psychological Measurement, 36*, 104–121.

Dirkzwager, A. (2003). Multiple evaluation: A new testing paradigm that exorcizes guessing. *International Journal of Testing, 3*, 333–352.

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.

Dolly, J. P. & Williams, K. S. (1986). Using test-taking strategies to maximize multiple-choice test scores. *Educational and Psychological Measurement, 46*, 619–625.

Downing, S. M. (2002a). Construct-irrelevant variance and flawed test questions: Do multiple-choice item-writing principles make any difference? *Academic Medicine, 77* (10), 103–104.

Downing, S. M. (2002b). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education, 7*, 235–241.

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.

Downing, S. M. & Haladyna, T. M. (1997). Test item development: Validity evidence from quality assurance procedures. *Applied Measurement in Education, 10*, 61–82.

Dunbar, S. B., Koretz, D. M. & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education, 4*, 289–303.

Eckes, T. (2014). Lokale Abhängigkeit von Items im TestDaF-Leseverstehen. *Diagnostica, 61*, 93–106.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists.* Mahwah, NJ: Lawrence Erlbaum.

Ericsson, K. A. & Simon, H. A. (1984). *Protocol analysis*. Cambridge, MA: MIT press.

Espinosa, M. P. & Gardeazabal, J. (2010). Optimal correction for guessing in multiple-choice tests. *Journal of Mathematical Psychology, 54*, 415–425.

Fajardo, L. L. & Chan, K. M. (1993). Evaluation of medical students in radiology. Written testing using uncued multiple-choice questions. *Investigative Radiology, 28*, 964–968.

Fenderson, B. A., Damjanov, I., Robeson, M. R., Veloski, J. J. & Rubin, E. (1997). The virtues of extended matching and uncued tests as alternatives to multiple choice questions. *Human Pathology, 28*, 526–532.

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

Frary, R. B. (1989). Partial-credit scoring methods for multiple-choice tests. *Applied Measurement in Education, 2*, 79–96.

Grier, J. B. (1975). The number of alternatives for optimum test reliability. *Journal of Educational Measurement, 12*, 109–113.

Grosse, M. E. & Wright, B. A. (1985). Validity and reliability of true-false tests. *Educational and Psychological Measurement, 45*, 1–13.

Haladyna, T. M. (1992). Context-dependent item sets. *Educational Measurement: Issues and Practice, 11* (1), 21–25.

Haladyna, T. M. (1997). *Writing test items to evaluate higher order thinking*. Boston, MA: Allyn and Bacon.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. New York, NY: Routledge.

Haladyna, T. M. & Downing, S. M. (1989). Validity of a taxonomy of multiple-choice item-writing rules. *Applied Measurement in Education, 2*, 51–78.

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.

Haladyna, T. M. & Downing, S. M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice, 23* (1), 17–27.

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.

Haladyna, T. M. & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hancock, G. R. (1994). Cognitive complexity and the comparability of multiple-choice and constructed-response test formats. *The Journal of Experimental Education, 62*, 143–157.

Hancock, G. R., Thiede, K. W., Sax, G. & Michael, W. B. (1993). Reliability of comparably written two-option multiple-choice and true-false test items. *Educational and Psychological Measurement, 53*, 651–660.

Hanna, G. S. & Johnson, F. R. (1978). Reliability and validity of multiple-choice tests developed by four distractor selection procedures. *The Journal of Educational Research, 71*, 203–206.

Higham, P. A. & Arnold, M. M. (2007). Beyond reliability and validity: The role of metacognition in psychological testing. In R. A. Degregorio (Ed.), *New developments in psychological testing* (pp. 139–162). Hauppauge, NY: Nova Science.

Hohensinn, C. & Kubinger, K. D. (2011). Applying item response theory methods to examine the impact of different response formats. *Educational and Psychological Measurement, 71*, 732–746.

Hollingworth, L., Beard, J. J. & Proctor, T. P. (2007). An investigation of item type in a standards-based assessment. *Practical Assessment, Research & Evaluation*, *12*(18). Retrieved from http://pareonline.net/genpare.asp?v=12&n=18

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.

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.

Kingston, N. M., Tiemann, G. C., Miller Jr, H. L. & Foster, D. (2012). An analysis of the discrete-option multiple-choice item type. *Psychological Test and Assessment Modeling, 54*, 3–19.

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

Kintsch, W. (1970). Models for free recall and recognition. In D. A. Norman (Ed.), *Models of human memory* (pp. 331–373). New York, NY: Academic Press.

Kubinger, K. D. (2014). Gutachten zur Erstellung «gerichtsfester» Multiple-Choice-Prüfungsaufgaben. *Psychologische Rundschau, 65*, 169–178.

Lane, S. (2004). Validity of high-stakes assessment: Are students engaged in complex thinking? *Educational Measurement: Issues and Practice, 23* (3), 6–14.

Lee, H. S., Liu, O. L. & Linn, M. C. (2011). Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Applied Measurement in Education, 24*, 115–136.

Leighton, J. P. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23* (4), 6–15.

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.

Levine, M. V. & Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement, 43*, 675–685.

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.

Lord, F. M. (1975). Formula scoring and number right scoring. *Journal of Educational Measurement, 12*, 7–11.

Lord, F. M. (1977). The optimal number of choices per item: a comparison of four approaches. *Journal of Educational Measurement, 14*, 33–38.

Martinez, M. E. (1999). Cognition and the question of test item format. *Educational Psychologist, 34*, 207–218.

Martínez, R. J., Moreno, R., Martín, I. & Trigo, M. E. (2009). Evaluation of five guidelines for option development in multiple-choice item-writing. *Psicothema, 21*, 326–330.

McAllister, D. & Guidice, R. M. (2012). This is only a test: A machine-graded improvement to the multiple-choice and true-false examination. *Teaching in Higher Education, 17*, 193–207.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

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

Mitkov, R., Ha, A. L. & Karamanis, N. (2006). A computer-aided environment for generating multiple-choice test items. *Natural Language Engineering, 12* (2), 177–194.

Mittring, G. & Rost, D. H. (2008). Die verflixten Distraktoren. *Diagnostica, 54*, 193–201.

Muñiz, J. & Menéndez, F. (2011). The answer-until-correct item format revisited. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 7*, 103–110.

Nickerson, R. S. (1989). New directions in educational assessment. *Educational Researcher, 18*, 3–7.

Oosterhof, A. C. & Glasnapp, D. R. (1974). Comparative reliabilities and difficulties of the multiple-choice and true-false formats. *The Journal of Experimental Education, 42*, 62–64.

Owens, R. E., Hanna, G. S. & Coppedge, F. L. (1970). Comparison of multiple-choice tests using different types of distractor selection techniques. *Journal of Educational Measurement, 7*, 87–90.

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

Prieto, G. & Delgado, A. R. (1999a). The role of instructions in the variability of sex-related differences in multiple-choice tests. *Personality and Individual Differences, 27*, 1067–1077.

Prieto, G. & Delgado, A. R. (1999b). The effect of instructions on multiple-choice test scores. *European Journal of Psychological Assessment, 15*, 143.

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.

Raven, J., Raven, J. C. & Court, J. H. (1998). *Manual for Raven's Progressive Matrices and Vocabulary Scales. Section 3, The Standard Progressive Matrices.* Oxford: Oxford Psychologists Press.

Reich, G. A. (2013). Imperfect models, imperfect conclusions: An exploratory study of multiple-choice tests and historical knowledge. *The Journal of Social Studies Research, 37*, 3–16.

Rodriguez, M. C. (1997, April). *The art & science of item writing: A meta-analysis of multiple-choice item format effects.* Paper presented at the Annual Meeting of the American Education Research Association, Chicago, IL.

Rodriguez, M. C. (2002). Choosing an item format. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 211–229). Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24* (2), 3–13.

Roediger, H. L., III & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1155–1159.

Rost, D. H. & Sparfeldt, J. R. (2007). Leseverständnis ohne Lesen? *Zeitschrift für Pädagogische Psychologie, 21*, 305–314.

Rost, J. (2004). *Lehrbuch Testtheorie-Testkonstruktion* (2., vollst. überarb. u. erw. Aufl.). Bern: Hans Huber.

Sarnacki, R. E. (1979). An examination of test-wiseness in the cognitive test domain. *Review of Educational Research, 49*, 252–279.

Schuwirth, L. W. T., van der Vleuten, C., Stoffers, H. E. J. H. & Peperkamp, A. G. W. (1996). Computerized long-menu questions as an alternative to open-ended questions in computerized assessment. *Medical Education, 30*, 50–55.

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.

Scouller, K. M. & Prosser, M. (1994). Students' experiences in studying for multiple choice question examinations. *Studies in Higher Education, 19*, 267–279.

Shepard, L. A. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*, 4–14.

Shermis, M. D. & Burstein, J. C. (Eds.). (2002). *Automated essay scoring: A cross-disciplinary perspective.* New York, NY: Routledge.

https://econtent.hogrefe.com/doi/pdf/10.1024/1010-0652/a000156 - Friday, April 26, 2024 12:35:51 PM - IP Address:18.191.202.45

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.

Simkin, M. G. & Kuechler, W. L. (2005). Multiple-choice tests and student understanding: What is the connection? *Decision Sciences Journal of Innovative Education, 3*, 73–98.

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.

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.

Swanson, D. B., Holtzman, K. Z. & Allbee, K. (2008). Measurement characteristics of content-parallel single-best-answer and extended-matching questions in relation to number and source of options. *Academic Medicine, 83*, 21–24.

Swanson, D. B., Holtzman, K. Z., Allbee, K. & Clauser, B. E. (2006). Psychometric characteristics and response times for content-parallel extended-matching and one-best-answer items in relation to number of options. *Academic Medicine, 81*, 52–55.

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.

Tarrant, M., Knierim, A., Hayes, S. K. & Ware, J. (2006). The frequency of item writing flaws in multiple-choice questions used in high stakes nursing assessments. *Nurse Education in Practice, 6*, 354–363.

Tarrant, M., Ware, J. & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: A descriptive analysis. *BMC Medical Education, 9*, 40.

Traub, R. E. & Fisher, C. W. (1977). On the equivalence of constructed-response and multiple-choice tests. *Applied Psychological Measurement, 1*, 355–369.

Trevisan, M. S., Sax, G. & Michael, W. B. (1991). The effects of the number of options per item and student ability on test validity and reliability. *Educational and Psychological Measurement, 51*, 829–837.

Trevisan, M. S., Sax, G. & Michael, W. B. (1994). Estimating the optimal number of options per item using an incremental option paradigm. *Educational and Psychological Measurement, 54*, 86–91.

Tripp, A. & Tollefson, N. (1985). Are complex multiple-choice options more difficult and discriminating than conventional multiple-choice options? *The Journal of Nursing Education, 24* (3), 92–98.

Tuerlinckx, F. & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods, 6*, 181–195.

Tversky, A. (1964). On the optimal number of alternatives at a choice point. *Journal of Mathematical Psychology, 1*, 386–391.

Von Schrader, S. & Ansley, T. (2006). Sex differences in the tendency to omit items on multiple-choice tests: 1980–2000. *Applied Measurement in Education, 19*, 41–65.

Vyas, R. & Supe, A. (2008). Multiple-choice questions: A literature review on the optimal number of options. *National Medical Journal of India, 21* (3), 130–133.

Wainer, H. & Thissen, D. (1993). Combining multiple-choice and constructed-response test scores: Toward a marxist theory of test construction. *Applied Measurement in Education, 6*, 103–118.

Wallach, P. M., Crespo, L. M., Holtzman, K. Z., Galbraith, R. M. & Swanson, D. B. (2006). Use of a committee review process to improve the quality of course examinations. *Advances in Health Sciences Education, 11*, 61–68.

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.

Wilcox, R. (1981). Solving measurement problems with an answer-until-correct scoring procedure. *Applied Psychological Measurement, 5*, 399–414.

Willing, S., Ostapczuk, M. & Musch, J. (2014). Do sequentially-presented answer options prevent the use of testwiseness cues on continuing medical education tests? *Advances in Health Sciences Education,* Advance online publication. doi: 10.1007/s10459-014-9528-2

Wilson, M. & Adams, R. J. (1995). Rasch Models for item bundles. *Psychometrica, 60*, 181–198.

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.

Yaman, S. (2011). The optimal number of choices in multiple-choice tests: Some evidence for science and technology education. *New Educational Review, 23*, 227–241.

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

Zenisky, A. L. & Sireci, S. G. (2002). Technological innovations in large-scale assessment. *Applied Measurement in Education, 15*, 337–362.

Zieky, M. (2006). Fairness review in assessment. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 359–376). Mahwah, NJ: Lawrence Erlbaum Associates.

Zimmerman, D. W. & Williams, R. H. (1982). Element of chance and comparative reliability of matching tests and multiple-choice tests. *Psychological Reports, 50*, 975–980.

Zimmerman, D. W. & Williams, R. H. (2003). A new look at the influence of guessing on the reliability of multiple-choice tests. *Applied Psychological Measurement, 27*, 357–371.

Marlit Annalena Lindner

Leibniz-Institut für die Pädagogik der Naturwissenschaften
und Mathematik (IPN)
Olshausenstraße 62
24118 Kiel
Deutschland
mlindner@ipn.uni-kiel.de