Handlungsempfehlungen zur Erstellung guter Multiple-Choice-Beispiele für Prüfer/innen auf dem Prüfstand

Dobrovits, Ingrid

Published in:
Facetten der Entrepreneurship Education - Festschrift Josef Aff

Published: 01/01/2016

Document Version
Publisher's PDF, also known as Version of record

Ingrid Dobrovits (WU Wien)

Handlungsempfehlungen zur Erstellung guter Multiple-Choice-Beispiele für Prüfer/innen auf dem Prüfstand

In Zeiten von überfüllten Hörsälen und unzureichenden Kapazitäten an österreichischen Universitäten ist das Streben nach Prüfungsökonomie entsprechend hoch. Mittlerweile sind standardisierte computergestützte (Massen-)Prüfungen sowohl für das Universitätspersonal als auch für die Studierenden ein gewohntes Prüfungsformat und deshalb wird an Österreichs Universitäten einer Multiple-Choice(MC)-Klausur oft der Vorrang gegenüber einer offenen Prüfung gegeben. Das für Massenprüfungen zuständige Universitätspersonal findet sich daher oft sehr plötzlich und unvermittelt in der Rolle als Multiple-Choice-Prüfer/in.

Unzählige Studien gehen der Frage nach, ob MC-Prüfungen (genauso wie offene Klausuren) überhaupt in der Lage sind, höhere kognitive Lernziele wie Verständnis, Analyse, Synthese und Evaluation abzutesten (vgl. z. B. Mandernach 2003; Gatterer 2013; Bacon 2003). Obgleich diese Untersuchungen zu gänzlich unterschiedlichen Ergebnissen in diesem Punkt gelangen, so zeigt sich doch als Grundtendenz, dass gut gestellte MC-Beispiele dieses (Prüfer/innen)Ziel sehr wohl erreichen können: “They do if we write them to do so.” (Torres 2011: 4). „Bei sorgsamer Konstruktion durch Fachexpert/innen sind Prüfungen im MC-Format mit jenen im offenen Format jedenfalls auf eine Qualitätsstufe zu stellen.“ (Gatterer 2013: 150). Wie aber können höherwertige MC-Beispiele durch die Prüfer/innen konstruiert werden?

1 Taxonomie zur Erstellung von MC-Beispielen nach Haladyna und Downing

Die wohl umfassendste Sammlung von Kriterien für die Erstellung von (guten) MC-Beispielen ist die von Haladyna, Downing und Rodriguez entwickelte Taxonomie „A Revised Taxonomy of Multiple-Choice (MC) Item-Writing Guidelines“ (vgl. Haladyna et al. 2002: 312 ff.).

Haladyna/Downing haben bereits im Jahr 1989 die von unterschiedlichsten Expert/innen aufgestellten Regeln zur Erstellung von MC-Beispielen zu einer Taxonomie zusammengefasst. Die ursprüngliche Variante umfasste 43 Richtlinien, die aus 46 Textbüchern und anderen Quellen der Prüfungsliteratur zusammengetragen wurden (vgl. Haladyna/Downing 1989: 37).

Diese Taxonomie wurde im Jahr 2002 überarbeitet und umfasst nunmehr 31 Richtlinien, welche auch gleich auf ihre Gültigkeit hin untersucht wurden (vgl. Haladyna et al. 2002: 312). Diese Taxonomie kann nach eigenen Angaben sowohl im Klassenverband als auch für groß angelegte Prüfungen Verwendung finden.

Die im Jahr 2002 überarbeitete Taxonomie teilt die Richtlinien zur besseren Übersicht für die Prüfungsersteller/innen in fünf Kategorien (Inhalt, Format, Stil, Fragenstamm und Antwortoptionen) ein. Es folgt hier eine stichwortartige Zusammenfassung der Richtlinien (eigene deutsche Übersetzung):

Inhalt	
1	jedes Beispiel auf der Grundlage eines Leitgedankens
2	keine trivialen Inhalte
3	neuartige Inhalte zur Überprüfung höherer Lernziele
4	voneinander unabhängige Beispiele
5	keine übermäßig speziellen oder übermäßig allgemeinen Inhalte
6	keine Fragen auf der Grundlage von Meinungen
7	keine Fangfragen (absichtliche „Fallen“)
8	einfaches Vokabular
Format	
9	Verwendung von verschiedenen MC-Typen, aber komplexen *Typ K*[24] vermeiden
10	vertikale Formatierung der Items anstatt horizontaler Formatierung
Stil	
11	Beispiele überprüfen und gegebenenfalls redigieren
12	korrekte Grammatik, Rechtschreibung, Zeichensetzung, Großschreibung
13	Leseaufwand gering halten
Fragenstamm	
14	klarer Fragenstamm
15	zentrale Frage im Fragenstamm
16	keine unnötigen Informationen im Fragenstamm
17	positive Formulierung des Fragenstammes, Negationen gänzlich vermeiden
Antwortoptionen	
18	so viele sinnvolle Antwortoptionen als möglich
19	nur eine Antwort ist eindeutig richtig/am besten
20	Platzierung der richtigen Antwort variieren
21	logische/numerische Anordnung
22	keine Überlappungen
23	inhaltliche und grammatikalische Homogenität
24	in etwa idente Länge
25	Antwortoption *„keine Antwort ist richtig“* vorsichtig verwenden
26	Antwortoption *„alle der genannten“* vermeiden
27	positive Formulierung, Negationen vermeiden
28	keine Hinweise zu der richtigen Antwort geben
29	nur plausible Distraktoren
30	typische Fehler als Grundlage der Distraktoren
31	Humor sparsam verwenden

Tabelle 1: A Revised Taxonomy of MC Item-Writing Guidelines (vgl. Haladyna et al. 2002: 312)

[24] Im Fall der *komplexen MC-Angabe* (Typ K) muss der Studierende aus mehreren Möglichkeiten eine oder mehrere richtige identifizieren, wobei aus Kombinationen auszuwählen ist.

Die Autoren haben in einer zweiteiligen Studie alle 31 Richtlinien auf ihre Validität hin überprüft (vgl. Haladyna et al.: 2002). Einerseits wurden dazu von den Autoren 27 aktuelle Textbücher zur Erstellung von MC-Beispielen analysiert und andererseits 27 Studien seit dem Jahr 1990 rezensiert. Für die in grau hinterlegten Richtlinien (Nummern 7, 8, 10, 15, 17, 18, 21, 23, 25, 26 sowie 31) fanden sich allerdings widersprüchliche Ergebnisse. Die Autoren Haladyna et al. halten dennoch weiterhin (theoretisch) an diesen fest. Im Detail zeigten sich zu diesen Richtlinien folgende Unstimmigkeiten:

Richtlinie 7 – Keine Fangfragen (vgl. Haladyna et al. 2002: 315)

Da in den von den Autoren untersuchten Textbüchern dem Problem der Fangfragen kaum Aufmerksamkeit geschenkt wurde, war des Weiteren zu überprüfen, woran dies liegen könnte. Es wird vermutet, dass dies auf die mangelnde Definition von Fangfragen zurückzuführen sein könnte. Einzig Roberts wagte sich an eine Definition, die von den Autoren um Haladyna einstimmig übernommen wurde. Er befragte 226 Personen nach ihrer Definition zu „Fangfragen“ (engl. *Trick Items*) und konnte sieben Kernpunkte ausfindig machen: demnach sind Fangfragen von der Intention des Testschreibers abhängig (absichtlich verwirrend oder irreführend formuliert) und dadurch gekennzeichnet, dass sie triviale bzw. unwichtige Inhalte abprüfen, zu geringe Unterschiede in den Antwortoptionen, unnötige Informationen im Fragenstamm (engl. *Window Dressing*), mehrere extrem ähnliche Antwortoptionen (Unterscheidung z. B. erst ab der dritten Kommastelle), Inhalte, die gegenteilig unterrichtet wurden oder hohe Mehrdeutigkeit aufweisen (vgl. Roberts 1993: 334 f.). Roberts konnte in seiner Untersuchung zeigen, dass Beispiele mit Fangfragen signifikant schwerer zu lösen sind als Beispiele ohne „Tricks“. Für Macher (2005: 42) sind Fangfragen „keinesfalls akzeptabel“, denn die Lösungswahrscheinlichkeit eines Beispiels „soll allein durch die Komplexität der zugrunde liegenden Problematik, den kognitiven Anspruch […] und die Feinheit der erforderlichen Differenzierung (inhaltliche Nähe der Antwortalternativen zueinander) bestimmt werden, und nicht durch künstlich eingebaute ‚Schikanen‘“.

Auch die Universität von Oregon empfiehlt ihren Testschreiber/innen, den Fragenstamm kurz zu halten, da unnötige Informationen die Lesezeit während der Prüfung verlängern und dies zulasten der Reliabilität geht, da dann weniger Beispiele abgefragt werden können (vgl. Universität von Oregon 2011).

Impara und Foster (2009: 98) halten dagegen: “The simpler the item, the more likely an examinee can memorize and recall it.” Sie sprechen sich für längere und komplexere Aufgabenstellungen aus zweierlei Gründen aus. Erstens kann damit dem Auswendiglernen von Fragen entgegengewirkt werden und zweitens müssen im wirklichen Leben Probleme ebenfalls in einem Kontext gelöst werden, wodurch es dem Prüfling durchaus abverlangt werden kann, relevante von nicht relevanten Informationen herauszufiltern.

Richtlinie 8 – Einfaches Vokabular (vgl. Haladyna et al. 2002: 315)

Hier kommen die Verfasser/innen der MC-Textbücher und die Ersteller/innen diverser Studien zu unterschiedlichen Ergebnissen: während die einen für einen auf die Sprache der Prüflinge abgestimmten angemessenen Sprachgebrauch plädieren, zeigen die von Haladyna et al. untersuchten Studien, dass das Vokabular generell eher einfach gehalten werden soll. Haladyna et al. treten daher für eine simple Sprache immer dann ein, wenn etwas anderes (nicht die Sprache) geprüft werden soll. In diese Kerbe schlägt auch Abedi (2009: 393), der die Reliabilität und Validität einer MC-Prüfung durch unnötige linguistische Komplexität gefährdet sieht. Besonders Prüflinge mit einer anderen Muttersprache als der in der Prüfung verwendeten sind dadurch benachteiligt.

Richtlinie 10 – Vertikale Formatierung der Items (vgl. Haladyna et al. 2002: 316)

Hier liegen keine Studien vor. Haladyna et al. argumentieren mit der ihrer Meinung nach übersichtlicheren Darstellung bei der vertikalen Formatierung des Beispiels im Vergleich zu einer horizontalen Anordnung.

Richtlinie 15 – Zentrale Frage im Fragenstamm (vgl. Haladyna et al. 2002: 316)

Hier wurde von Haladyna et al. zwar eine Studie, die unterschiedliche Arten von Fragenstämmen untersucht hat (zentrale Frage im Fragenstamm versus unfokussierter Fragenstamm), analysiert, es konnten jedoch keine Unterschiede in der Leistung der Studierenden festgestellt werden.

Diese Richtlinie wird bei Brauns und Schubert mittels der „Antwortoptionen-Abdecken-Regel“ (engl. *Cover-The-Options-Rule*) genauer beschrieben. Diese besagt, dass eine gute MC-Frage auch nach dem Abdecken der Antwortoptionen als Freitextfrage beantwortbar sein muss. Dadurch ist sichergestellt, dass alle notwendigen Informationen zur korrekten Lösung der Aufgabe in der Frage und nicht in den Antwortoptionen enthalten sind und die Frage klar und verständlich formuliert wurde (vgl. Brauns/Schubert 2008: 98).

Richtlinie 17 – Positive Formulierung des Fragenstammes (vgl. Haladyna et al. 2002: 316 f.)

Hier waren die untersuchten Studienergebnisse nicht einheitlich, weshalb Haladyna et al. zu einem „vorsichtigen Umgang“ mit negativen Formulierungen raten. Abedi (2009: 387) rät von negativen Formulierungen generell ab, da diese Prüflinge mit anderer Muttersprache durch die damit verbundene höhere Sprachkomplexität benachteiligen könnten.

Richtlinie 18 – So viele sinnvolle Antwortoptionen als möglich (vgl. Haladyna et al. 2002: 317 f.)

Üblicherweise werden bei MC-Beispielen vier Antwortoptionen verwendet, was auch von den bei Haladyna et al. untersuchten Textbuchautor/innen empfohlen wird. Haladyna et al. bleiben aber bei ihrer Meinung, dass so viele Antwortoptionen als möglich eingesetzt werden sollten, wobei sie die Erstellung von mehr als zwei plausiblen Distraktoren aber als sehr unwahrscheinlich halten. Die zahlreichen Studien auf diesem Gebiet sind jedoch nicht einheitlich, was die Auswirkungen von mehr oder weniger Antwortoptionen betrifft (vgl. Rogers/Harley 1999: 246).

In der aktuellen und daher von Haladyna et al. noch nicht berücksichtigten Studie von Baghaei und Amrahi aus dem Jahr 2011 wurde diesem Umstand wiederum Aufmerksamkeit geschenkt. Verglichen wurden MC-Beispiele in einem Vokabeltest mit drei, vier und fünf Antwortoptionen. Als Ergebnis halten die beiden fest, dass es keine statistisch signifikanten Unterschiede sowohl im Schwierigkeitsgrad als auch bei der Reliabilität der Beispiele gibt. Einzig die Trennschärfe der Beispiele steigt leicht bei einer Reduktion der Anzahl der Antwortoptionen. Die gleichbleibende Reliabilität ist das wohl überraschendste Ergebnis dieser Studie und bezieht sich auf das Antwortverhalten der einzelnen Personen. Die oftmals geäußerte Vermutung, dass sich weniger Antwortoptionen in einem gesteigerten Rateverhalten der Prüflinge ausdrückt, kann hier widerlegt werden: die Studierenden ändern ihr Antwortverhalten nicht, wenn ihnen weniger Auswahloptionen zur Verfügung stehen. Baghaei und Amrahi (2011: 206 f.) schlussfolgern, dass Studierende immer dann raten, wenn die Beispiele zu schwierig sind oder aus Zeitmangel nicht alle Beispiele bearbeitet werden können. Sie empfehlen daher – zumindest für Vokabeltests – zur Minimierung

von Zeit und Mühen bei der Prüfungserstellung drei Antwortoptionen, da die Ratewahrscheinlichkeit, die Trennschärfe sowie die Reliabilität bei höherer Erstellungs- und Lesezeit gleich bleiben. Die Studie von Rogers und Harley aus dem Jahr 1999 untersuchte die Anzahl der optimalen Antwortoptionen im Fachbereich Mathematik. Verglichen wurden dabei MC-Beispiele mit drei sowie mit vier Antwortmöglichkeiten, wobei nicht auf Faktenwissen abgestellt wurde, sondern jede Frage mittels einer Berechnung gelöst werden musste. Es konnte festgestellt werden, dass die Testreliabilität bei drei Antwortoptionen mit jener bei vier vergleichbar ist und es keine großen Unterschiede bei den Bearbeitungszeiten der beiden Antwortformate gibt. Außerdem konnte gezeigt werden, dass der Einfluss von Lösungsstrategien bei drei Antwortoptionen abnimmt. Dies wird darauf zurückgeführt, dass zumindest der dritte Distraktor meist nicht mehr gut funktioniert, da dieser dem Prüfling mit Teilwissen oftmals nicht mehr plausibel erscheint.

Ganz anders argumentieren Lissmann und Jäger. Um die Ratewahrscheinlichkeit zu minimieren soll die Anzahl der Distraktoren auf fünf bis sieben erhöht werden (vgl. Lissmann/Jäger 2008).

Richtlinie 21 – Logische/Numerische Anordnung der Antwortoptionen (vgl. Haladyna et al. 2002: 318)

Hier konnte von den Autoren um Haladyna eine Studie gefunden werden, die jedoch keinerlei Unterschiede im Schwierigkeitsgrad zwischen logischer und zufälliger Anordnung der Antwortoptionen erkennen lässt. Es wird aber gefolgert, dass zufällig angeordnete Antwortoptionen Hindernisse für weniger talentierte Studierende sein könnten. Solange es aber keine Untersuchung zu dieser Richtlinie gibt, schlagen Impara und Foster (2009: 102) jedenfalls vor, die Antwortalternativen zufällig anzuordnen. Ihrer Meinung nach kann dies die Sicherheit der Prüfung erhöhen, indem die Wahrscheinlichkeit des Abschreibens und damit die Schummelgefahr während der Prüfung verringert werden.

Richtlinie 23 – Inhaltliche und grammatikalische Homogenität der Antwortoptionen (vgl. Haladyna et al. 2002: 318)

Auch für diese Richtlinie konnten bisher von den Autoren keine empirischen Beweise gefunden werden, die diesen Anspruch untermauern.

Richtlinie 25 – Antwortoption „Keine Antwort ist richtig“ vorsichtig verwenden (vgl. Haladyna et al. 2002: 319)

Die Autor/innen der untersuchten Textbücher waren in etwa zu 50 % der Meinung, diese Antwortoption könne und zu 50 % der Meinung, sie könne nicht verwendet werden. Alle fünf von Haladyna et al. untersuchten Studien kommen zu dem Ergebnis, dass der Einsatz dieser Antwortoption den Schwierigkeitsgrad des jeweiligen Beispiels erhöht. Was allerdings die Trennschärfe der Beispiele angeht, sind die Ergebnisse nicht einheitlich. Da es sich hierbei um eine komplexe Antwortoption handeln dürfte, deren Effekte noch nicht hinreichend belegt wurden, wird der Einsatz von Haladyna et al. nur für routinierte MC-Ersteller/innen befürwortet.

Macher merkt an, dass – wenn schon nicht auf dieses Format verzichtet werden kann – zumindest hie und da auch diese Antwortmöglichkeit die richtige sein muss (vgl. Macher 2005: 45).

Richtlinie 26 – Antwortoption „Alle der genannten“ vermeiden (vgl. Haladyna et al. 2002: 319)

In die gleiche Kerbe wie Richtlinie 25 schlägt die Antwortoption „Alle der genannten“. Die Autoren um Haladyna hatten lediglich eine Studie zu dieser Antwortoption zur Verfügung, die eine signifikant niedrigere Reliabilität des betreffenden Beispiels bei deren Einsatz zeigt.

Richtlinie 31 – Humor sparsam verwenden (vgl. Haladyna et al. 2002: 320)

Hier stand den Autoren um Haladyna wiederum nur eine empirische Untersuchung zur Verfügung, deren Ergebnisse diese zum Schluss kommen lässt, dass Humor im Klassenzimmer bei Prüfungen zum Einsatz kommen kann, wenn dieses Frageformat zum Lehrenden passt und die Vorteile die Nachteile überwiegen. Vom Gebrauch in formalen Testprogrammen würden die Autoren jedoch abraten.

2 Weitere aktuelle Forschungsergebnisse zur Erstellung von MC-Beispielen für höhere kognitive Leistungsmessungen

Als eine der wenigen erklärt auch Macher in ihrem Handbuch zur Erstellung guter MC-Aufgaben sehr ausführlich, wie Fähigkeiten wie Informationsinterpretation, -integration oder Problemlösefähigkeiten ihrer Meinung nach mittels geschlossenem Frageformat geprüft werden können. Sie folgert, dass immer dann zweiteilige Item-Stämme nötig sind, sobald den Studierenden höhere kognitive Leistungen abverlangt werden. Auf einen langen Fragenstamm mit allen zur Lösung benötigten Informationen folgen als Überleitung die Fragestellung und danach kurze, klare Antwortalternativen. Diese längeren Fallbeschreibungen bezeichnet Macher als „Vignetten“ (engl. *Short-cases*) (vgl. Macher 2005: 42 f.).

Als ganz wesentliche Kriterien bei der Erstellung von Fallvignetten sind ihrer Meinung nach die folgenden beiden Punkte zu berücksichtigen:

- Alle erforderlichen Informationen sollen im Stamm enthalten sein und nicht in den Antwortoptionen, was zum zweiten wesentlichen Kriterium führt, nämlich:
- Die Beantwortung der Fragen soll auch ohne Kenntnis der Antwortoptionen möglich sein (vgl. Macher 2005: 43).

Diese Überlegungen decken sich im Wesentlichen mit der oben vorgestellten Richtlinie 15 – Zentrale Frage im Fragenstamm – der Taxonomie von Haladyna et al.

Zimmaro (2010: 17 f.) schlägt weiters eine Interpretation von Diagrammen, Tabellen oder Ähnlichem vor und sieht die Prüfungsbeispiele in der realen Welt oder anderen praktischen Situationen angesiedelt.

Burton et al. nennen vor allem die Anwendung plausibler Distraktoren als Gütekriterium Nummer eins. Um gute Distraktoren in einem MC-Beispiel anführen zu können, sind deren Meinung nach vor allem Praxis und Erfahrung der Prüfer/innen von großer Bedeutung, wenngleich auch durch das Studium einschlägiger Literatur viel dazugelernt werden kann (vgl. Burton et al. 1991: 6). Zum selben Schluss kommen Torres et al.: “It [Anm.: die Erstellung guter MC-Beispiele] requires a certain amount of skills and knowledge. […] However, this skill can be increased through study, training, practice and experience.” (vgl. Torres et al. 2011: 2).

3 Schlussfolgerungen

Der Beitrag zeigt, dass man als MC-Prüfer/in auf zahlreiche Empfehlungen für die Erstellung von MC-Klausuren zugreifen kann, sich diese aber vorwiegend mit technischen sowie formalen Gestaltungsmöglichkeiten beschäftigen. Konkrete Regeln oder Hilfestellungen zu Maßnahmen, um höhere kognitive Leistungen inhaltlich testen zu können, bleiben weitgehend aus. Außerdem sind nach wie vor viele Handlungsempfehlungen empirisch nicht belegt. Haladyna et al. sind sich dessen selbst bewusst, da zu den Richtlinien ihrer Taxonomie bislang nur etwa die Hälfte ausreichend empirisch belegt wurden. Für die andere Hälfte gibt es lediglich theoretische Begründungen bzw. es handelt sich um Werturteile oder um einen Expertenkonsens (vgl. Brauns/Schubert 2008: 97).

Auch Torres et al. bleiben weiterführende Erklärungen für Erstellungsregeln schuldig. In ihrer Zusammenfassung merken sie zwar an, dass es im Fachbereich Mathematik möglich sei, MC-Beispiele auf höheren Lernzielniveaus zu stellen, dies aber in der konkreten Umsetzung Schwierigkeiten aufwirft (Torres et al. 2011: 10 f.).

Unumstritten ist, dass für die Konstruktion hochwertiger MC-Beispiele viel Zeitaufwand vonseiten der Ersteller/innen erforderlich ist. “Multiple-choice questions measuring in-depth understanding are difficult and time-consuming to write. That is why many introductory textbook test banks are dominated by recognition and understanding questions” (Buckles/Siegfried 2006: 56). Brown vermutet, dass die Erstellung einer MC-Prüfung typischerweise doppelt so lange dauert wie die Erstellung derselben Prüfung im offenen Format (vgl. Brown 2001: 17).

Problematisch ist auch, dass es für die Prüfer/innen keinerlei Anreize gibt, qualitativ hochwertige Prüfungsfragen zusammenzustellen. Auch an der WU Wien werden zwar Prüfungstaxen für die Arbeiten rund um die Klausuren bezahlt, diese werden jedoch rein an quantitativen Maßstäben (z. B. Anzahl der angetretenen Prüflinge, Anzahl der durchgeführten Prüfungstermine) bemessen. Personal, das viel Zeit für die Erstellung hochwertiger Prüfungsbeispiele aufwendet, handelt demnach nur aus intrinsischer Motivation heraus. Bible et al. diskutieren in ihrem Beitrag aus diesem Grund sogar eine Bestrafung der Prüfer/innen bei minderwertig erstellten Prüfungsfragen (vgl. Bibel et al.: 2008: 57).

Wenngleich es also viele Studien und noch mehr Handlungsempfehlungen zu dem Thema gibt, fühlt man sich als MC-Prüfer/in dennoch weitgehend alleine gelassen und muss nach wie vor vieles „aus dem Bauch heraus“ entscheiden. Es wäre daher schon dringend an der Zeit, dass sich die einzelnen Fachdisziplinen dem Thema annehmen.

Literatur

Abedi, J. (2009): Language Issues in Item Development. In: Downing, S./Haladyna, T. (Hrsg.): Handbook of Test Development, S. 377–398.

Bacon, D.R. (2003): Assessing learning outcomes: a comparison of multiple-choice and short answer questions in a marketing context. In: Journal of Marketing Education, 25 (1), S. 31–36.

Baghaei, P./Amrahi, N. (2011): The effects of the number of options on the psychometric characteristics of multiple choice items. In: Psychological Test and Assessment Modeling, 53 (2), S. 192–211.

Bible, L./Simkin, M./Kuechler, W. (2008): Using Multiple-choice Tests to Evaluate Students’ Understanding of Accounting. In: Accounting Education: An International Journal, Vol. 17, Supplement 1, S. 55–68.

Brauns, K./Schubert, S. (2008): Qualitätssicherung von Multiple-Choice-Prüfungen. In: Prüfungen auf die Agenda. Bielefeld, S. 92–102.

Brown, R. (2001): Multi-Choice versus descriptive examinations. Paper presented at the 31st ASEE/IEEE Frontiers in Education Conference, Oct. 2001, Reno.

Buckles, S./Siegfried, J. J. (2006): Using Multiple-Choice Questions to Evaluate In-Depth Learning of Economics. In: Journal of Economic Education, 37 (1), S. 48–57.

Burton, S./Sudweeks, R./Merrill, P./Wood, B. (1991): How to prepare better Multiple-Choice Test Items: Guidelines for University Faculty, bezogen unter: https://testing.byu.edu /handbooks/betteritems.pdf (03.12.2015).

Gatterer, B. (2013): Würfeln oder Prüfen – Wie zuverlässig und aussagekräftig sind Multiple Choice-Tests im Fach Kostenrechnung? Wien.

Haladyna, T./Downing, S. (1989): A Taxonomy of Multiple-Choice Item-Writing Rules. In: Applied Measurement in Education, 2 (1), S. 37–50.

Haladyna, T./Downing, S./Rodriguez, M. (2002): A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. In: Applied Measurement in Education, 15 (3), S. 309–334.

Impara, J./Foster, D. (2009): Item and Test Development Strategies to Minimize Test Fraud. In: Downing, S./Haladyna, T. (Hrsg.): Handbook of Test Development, S. 91–114.

Legg, S. M. (1991): Handbook on Testing and Grading, Gainsville, Florida, bezogen unter: https://teachingcenter.ufl.edu/files/materials/training/handbook_testing_grading.pdf (03.12.2015).

Lissman, U./Jäger, R.S. (2008): Multiple-Choice Aufgaben. In: Journal für Lehrerinnen und Lehrerbildung, 1/2008, S. 45–50.

Litzenberger, M./Punter, J. F./Gnambs, T./Jirasko, M./Spiel, C. (2007): Qualitätssicherung bei der Studierendenauswahl mittels lernpsychologisch fundierter Wissensprüfung. In: Kluge, A./Schüler, K. (Hrsg.): Qualitätssicherung und -entwicklung an Hochschulen: Methoden und Ergebnisse. Lengerich.

Macher, S. (2005): Standardisierte Prüfungsmethoden in der medizinischen Ausbildung – Handbuch zur Konstruktion von Prüfungsaufgaben. Graz: Medizinische Universität, bezogen unter: https://www.medunigraz.at/fileadmin/lehren/planen-organisieren/pdf/QM_SM_Handbuch_PruefungenAllgemein_20050404_01.pdf (03.12.2015).

Mandernach, B. J. (2003): *Quality True-False Items*, bezogen unter: http://www.park.edu /center-for-excellence-in-teaching-and-learning/resources/cetl-quick-tips/true-false.html (03.12.2015).

Roberts, D.M. (1993): An empirical study on the nature of trick test questions. In: Journal of Educational Measurement, Vol. 30, S. 331–344.

Rogers, W.T./Harley, D. (1999): An Empirical Comparison of Three- and Four-Choice Items and Tests: Susceptibility to Testwiseness and Internal Consistency Reliability. In: Educational and Psychological Measurement, Vol. 59, S. 234–247.

Torres, C./Lopes, A.P./Babo, L./Azevedo, J. (2011): Improving Multiple-Choice Questions. In: US-China Education Review B, Vol. 1, S. 1–11.

University of Oregon Teaching and Learning Center (2015): Practical Suggestions for Writing Multiple-Choice Questions, bezogen unter: http://tep.uoregon.edu/resources /assessment/multiplechoicequestions/practicalsuggestions.html (03.12.2015).

Zimmaro, D. (2010): Writing good Multiple-Choice Exams, bezogen unter: https://learning-sciences.utexas.edu/sites/default/files/writing-good-multiple-choice -exams-04-28-10_0.pdf (03.12.2015).