Bibliografische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnd.ddb.de abrufbar.

Die Reihe „Blickpunkt Hochschuldidaktik" ist die Wissenschaftliche Buchreihe der Arbeitsgemeinschaft für Hochschuldidaktik (AHD) und wird herausgegeben von Ulrich Welbers (Düsseldorf).

Die Verantwortung der Beiträge liegt bei den Autorinnen und Autoren.

Band Nr. 118

© W. Bertelsmann Verlag GmbH & Co. KG, Bielefeld, 2008

Gesamtherstellung:
W. Bertelsmann Verlag Bielefeld, www.wbv.de

Gestaltung: www.lokbase.de

ISBN 978-3-7639-3571-0
Bestell-Nr. 60.01.851

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Insbesondere darf kein Teil dieses Werkes ohne vorherige schriftliche Genehmigung des Verlages in irgendeiner Form (unter Verwendung elektronischer Systeme oder als Ausdruck, Fotokopie oder unter Nutzung eines anderen Vervielfältigungsverfahrens) über den persönlichen Gebrauch hinaus verarbeitet, vervielfältigt oder verbreitet werden.

Für alle in diesem Werk verwendeten Warennamen sowie Firmen- und Markenbezeichnungen können Schutzrechte bestehen, auch wenn diese nicht als solche gekennzeichnet sind. Deren Verwendung in diesem Werk berechtigt nicht zu der Annahme, dass diese frei verfügbar seien.

Sigrid Dany, Birgit Szczyrba, Johannes Wildt (Hg.)

Blickpunkt Hochschuldidaktik Band 118

Prüfungen auf die Agenda!

Hochschuldidaktische Perspektiven auf Reformen im Prüfungswesen

AHD Arbeitsgemeinschaft für Hochschuldidaktik

wbv

Katrin Brauns, Sebastian Schubert:

Qualitätssicherung von Multiple-Choice-Prüfungen

„Despite attacks on MC testing, it has thrived in recent years"
Haladyna (2004)

Zusammenfassung

Dieser Beitrag beschäftigt sich mit der Notwendigkeit und den Möglichkeiten zur Qualitätssicherung von Multiple-Choice-Prüfungen (MC-Prüfungen) im Spannungsfeld von testtheoretischen Anforderungen und praktischer Umsetzbarkeit. Der Beitrag berücksichtigt dabei sowohl die Qualitätssicherung auf der Ebene der einzelnen Multiple-Choice-Fragen (MC-Fragen), als auch auf der Ebene von Multiple-Choice-Klausuren.

In der Einleitung wird erläutert, warum Prüfungen als Messinstrumente bzw. diagnostische Tests grundsätzlich die gleichen Gütekriterien wie andere (psychologische) Messinstrumente (z.B. Intelligenz- oder Persönlichkeitstests) erfüllen sollten. Die Vor- und Nachteile von MC-Fragen und MC-Klausuren gegenüber anderen Prüfungsformaten und die Kriterien zur Auswahl eines bestimmten Prüfungsformats zur Überprüfung eines bestimmten Inhalts werden kurz dargestellt. Wir gehen davon aus, dass die Entscheidung zum Einsatz von MC-Fragen bereits gefallen ist, und es jetzt darum geht, die Qualität dieser Fragen zu sichern.

Im Hauptteil werden die Gütekriterien bzw. Standards mit praktischen Beispielen vorgestellt, und es wird erläutert, wie diese Gütekriterien bei MC-Prüfungen umgesetzt werden können. Dazu werden Möglichkeiten zur Messung der Gütekriterien und Methoden zur Verbesserung bei mangelhafter Güte vorgestellt. Am Ende werden Empfehlungen gegeben, mit welchen Methoden die Qualität von MC-Prüfungen kontinuierlich gesichert werden kann.

Gliederung

1. Einleitung
   1.1 Prüfung als Messinstrumente
   1.2 MC-Prüfungen

2. Qualitätssicherung von MC-Prüfungen
   2.1 Itemkennwerte als Qualitätskriterien auf Itemebene
   2.2 Testgütekriterien als Qualitätskriterien auf Testebene
   2.3 Regeln zur Erstellung von MC-Fragen
3. Praktische Umsetzbarkeit
4. Maßnahmen zur Sicherung der Qualität von MC-Prüfungen

1. Einleitung

1.1 Prüfungen als Messinstrumente

Hochschulprüfungen können als „diagnostische" Leistungsmessungen verstanden werden. Ihr Ziel ist die möglichst genaue und fehlerfreie Abbildung der Fähigkeiten von Studierenden in einem bestimmten Stoffgebiet zu einem bestimmten Zeitpunkt. Grundsätzlich gelten daher für sie die gleichen Anforderungen wie für andere (psychologische) Messinstrumente (zum Beispiel Intelligenz- oder Persönlichkeitstests). Das wird auch dadurch deutlich, dass die *American Educational Research Association (AERA), die American Psychological Association (APA) und der National Council on Measurement in Education (NCME)* gemeinsame *Standards for educational and psychological testing* (AERA, APA, NCME 1999) erarbeitet haben. Dort werden für die drei Bereiche „Testkonstruktion, -evaluation und -dokumentation", „Testfairness" und „Testanwendung" in 15 Kapiteln insgesamt 264 Standards definiert. Nicht alle Standards sind dabei für alle Tests relevant. Für die Bereiche „Testkonstruktion, -evaluation und -dokumentation" und „Testfairness" gelten die Standards gleichermaßen für psychologische Tests wie für Tests im Bildungsbereich, das heißt für Prüfungen. Im Bereich „Testanwendung" gibt es unterschiedliche Standards für die beiden Anwendungsgebiete, da sich hier zum Beispiel die Konsequenzen der Testergebnisse und damit auch der Anreiz zu Betrugsversuchen deutlich unterscheiden.

1.2 MC-Prüfungen

Die Begriffe „MC-Frage" und „MC-Prüfung" werden sowohl für die gesamte Gruppe schriftlicher Prüfungsformate mit Auswahlantworten (engl. selected response questions) als auch für bestimmte Unterformate dieser Gruppe verwendet. Sie unterscheiden sich von den Formaten mit Freitextantworten (engl. constructed response questions) dadurch, dass die Prüfungskandidaten

bei MC-Fragen die richtige(n) Antwort(en) aus mehreren vorgegebenen Antwortmöglichkeiten auswählen, während sie bei den Formaten mit Freitextantworten die Antworten selbst erstellen („konstruieren") müssen. In der Gruppe der Prüfungsformate mit Auswahlantworten gibt es mehrere verschiedene Formate. Am bekanntesten und häufigsten eingesetzt wird das so genannte A(pos)-Format. Dabei folgen auf eine Frage drei bis fünf (oder mehr) Auswahlantworten, von denen nur eine richtig ist. In Abbildung 1 ist ein Beispiel für eine A(pos)-MC-Frage dargestellt.

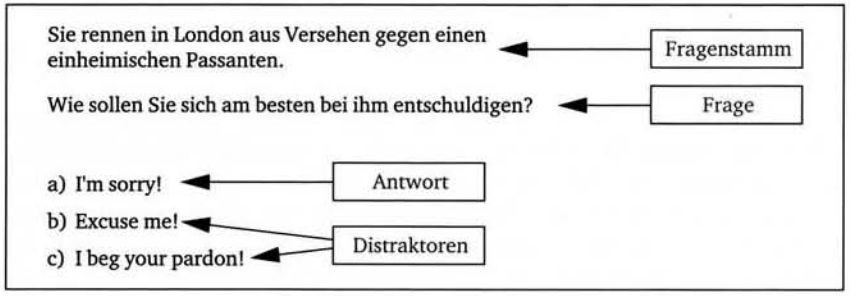

Abbildung 1: Beispiel für eine MC-Frage des A(pos)-Typs

Neben den A(pos)-MC-Fragen gibt es auch MC-Fragen mit vier oder fünf Auswahlantworten, von denen nur eine *falsch* ist (Aneg-MC-Fragen), Zuordnungsaufgaben, bei denen die Antworten auf mehrere Fragen aus einer langen Liste von Antwortmöglichkeiten ausgewählt werden müssen, MC-Fragen mit mehreren richtigen Antworten, Fragen, bei denen eine Aussage als richtig oder falsch beurteilt werden muss und eine Reihe anderer Formate. Eine ausführliche Darstellung mit den Vor- und Nachteilen der verschiedenen MC-Formate findet sich zum Beispiel bei IAWF (1999), Case & Swanson (2002) oder Haladyna (2004). Seit ihrer Einführung vor über 80 Jahren haben sich MC-Fragen zur Prüfung kognitiven Wissens im Hochschulbereich fest etabliert. Dies liegt vor allem an ihren gut erforschten und günstigen psychometrischen Eigenschaften (hohe Durchführungs- und Auswertungsobjektivität, günstiges Verhältnis von erreichbarer Reliabilität und Prüfungszeit), ihrer Kosteneffizienz bei der Prüfung großer Studierendenzahlen und ihrer vielseitigen Einsetzbarkeit zur Prüfung der meisten kognitiven Wissensbereiche. Die valide Prüfung dieser kognitiven Wissensbereiche verlangt allerdings sowohl qualitativ hochwertige einzelne MC-Fragen (Qualitätssicherung auf der Itemebene) als auch sorgfältig zusammengestellte MC-Prüfungen (Qualitätssicherung auf der Testebene). Die Erstellung von qualitativ hochwertigen MC-Fragen und MC-Prüfungen ist dabei aufwändiger und teurer als vielfach angenommen wird. Es muss auch beachtet werden, dass MC-Prüfungen nur für die Überprüfung der Fähigkeiten eingesetzt werden, für die sie geeignet sind. So lassen sich zum Beispiel praktische oder kreative Fertigkeiten auch mit sehr sorgfältig erstellten MC-Fragen nicht valide prüfen.

Grundsätzlich sollte die Wahl des Messinstruments von den zu messenden Fähigkeiten bestimmt sein und nicht umgekehrt. Leider ist es in der Praxis jedoch oft so, dass die tatsächlich gemessenen Fähigkeiten von Prüfungsteilnehmern von der Verfügbarkeit und der Kosteneffizienz des Messinstruments bestimmt werden.

2. Qualitätssicherung von MC-Prüfungen

Die Maßnahmen zur Qualitätssicherung von MC-Prüfungen beinhalten sowohl die Gütekriterien nach der klassischen Testtheorie auf der Item- und Testebene (KTT, siehe Lienert & Raatz, 1998) als auch die Regeln zur Erstellung guter MC-Fragen. Wir stellen zunächst die Gütekriterien nach der klassischen Testtheorie und anschließend die Regeln zur Erstellung guter MC-Fragen dar.

2.1 Itemkennwerte als Qualitätskriterien auf Itemebene

Zu den Kennwerten als Qualitätskriterien auf der Itemebene gehören Schwierigkeit, Trennschärfe und Homogenität.

Die *Schwierigkeit* einer Frage gibt an, wie viel Prüfungsteilnehmer diese richtig beantwortet haben. Der Begriff der Schwierigkeit ist in diesem Zusammenhang ungünstig gewählt, da eine hohe Schwierigkeit eine leichte Frage, eine niedrige Schwierigkeit eine schwere Frage anzeigt. Die Schwierigkeiten der Fragen sollten in der Prüfung näherungsweise normal verteilt sein. Das heißt, eine Prüfung sollte viele mittelschwere und wenige schwere bzw. leichte Fragen enthalten. Um diese Qualitätsanforderung praktisch umzusetzen, wird eine ausreichende Anzahl von Fragen bekannter und unterschiedlicher Schwierigkeit benötigt. Die Schwierigkeit kann zum Beispiel durch Vortests oder wiederholten Einsatz bekannt sein oder von Experten geschätzt werden.

Die *Trennschärfe* gibt an, wie gut eine Frage zwischen leistungsstärkeren und leistungsschwächeren Prüfungsteilnehmern diskriminieren kann. Eine Frage erreicht die maximale Trennschärfe, wenn sie von keinem der leistungsschwächeren Prüfungsteilnehmer aber von allen leistungsstärkeren Prüfungsteilnehmern richtig beantwortet wird. Üblicherweise werden Trennschärfen mit der korrigierten punktbiserialen oder der Produkt-Moment-Korrelation berech-

net. Trennschärfen können Werte zwischen –1 und 1 annehmen. Trennschärfen unter 0,3 werden als gering, zwischen 0,3 und 0,5 als mittel und über 0,5 als hoch bezeichnet (Itten & Krebs, 1997). Wenn eine Frage eine Trennschärfe von unter 0,3 aufweist, sollte kontrolliert werden, ob sich dies durch formale oder inhaltliche Fehler erklären lässt. So kann zum Beispiel eine Frage mit einem versteckten Lösungshinweis auch von Prüfungsteilnehmern, die das gefragte Wissen nicht besitzen, richtig beantwortet werden. Oder eine Frage kann so missverständlich oder verwirrend formuliert sein, dass sie auch von Prüfungsteilnehmern, die das gefragte Wissen eigentlich besitzen, falsch beantwortet wird. Auch wenn die einzelnen MC-Fragen einer MC-Prüfung möglichst hohe Trennschärfen haben sollten, sollten formal gute und inhaltlich sinnvolle MC-Fragen nicht allein auf Grund einer geringen Trennschärfe aus einer MC-Prüfung entfernt werden.

Die *Homogenität* ist als Gütekriterium für MC-Prüfungen nur von geringer praktischer Relevanz, soll der Vollständigkeit halber aber kurz erwähnt werden. Lienert und Raatz (1998) schlagen im Rahmen der Anwendung der KTT vor, zur Bestimmung der Homogenität den Homogenitätsindex zu errechnen. Dieser gibt das Ausmaß an, in dem ein Item zur Homogenität des Tests beiträgt und wird mit der mittleren Interkorrelation eines Items zu den übrigen Items berechnet.

2.2 Testgütekriterien als Qualitätskriterien auf Testebene

Zu den Testgütekriterien auf der Testebene gehören die so genannten Hauptgütekriterien Objektivität, Reliabilität und Validität.

Bei der *Objektivität* wird zwischen Durchführungs-, Auswertungs- und Interpretationsobjektivität unterschieden. Das bedeutet, dass jeder Prüfungsteilnehmer unter den gleichen Testbedingungen geprüft wird, die Auswertung bei allen Prüfungsteilnehmern standardisiert erfolgt und das Ergebnis der jeweiligen Prüfung anhand einheitlicher Kriterien bewertet wird. Dies ist bei MC-Prüfungen in der Regel gegeben.

Die *Reliabilität* zeigt, wie genau die Messung bzw. wie hoch die Reproduzierbarkeit der Ergebnisse ist. Es gibt drei Methoden zur Bestimmung der Reliabilität: die Retest-Reliabilität, die Paralleltest-Reliabilität und die Verfahren zur Ermittlung der inneren Konsistenz. Bei der Berechung der Reliabilität von Prüfungen werden hauptsächlich Verfahren der Konsistenzanalyse verwendet. Es werden dabei die mittleren Interkorrelationen aller Fragen berechnet. Hierzu zählen Cronbachs Alpha und die Kruder-Richardson-Formel (KR20). Da beide Verfahren davon ausgehen, dass in der Prüfung nur eine zugrunde liegende Fähigkeit gemessen wird, kann es sein, dass beide die wahre Reliabilität unterschätzen. Häufig wird für Prüfungen mit hohem Einsatz, zum Beispiel Hochschulabschlussprüfungen, eine Mindestreliabilität von 0,8 gefordert. Da die Reliabilität mit der Anzahl der MC-Fragen in einer Prüfung zusammenhängt, kann die Reliabilität in der Regel durch eine Erhöhung der Fragenanzahl gesteigert werden. Dabei muss man jedoch beachten, dass die Konzentrationsleistung der Prüflinge abnimmt, je länger die Prüfung dauert.

Die *Validität* gibt darüber Auskunft, ob eine Prüfung das misst, was sie zu messen vorgibt. Man unterscheidet hierbei Inhalts-, Konstrukt- und kriterienbezogene Validität. Inhaltsvalide Prüfungen repräsentieren den Inhalt der Lehre. Die Konstruktvalidität beschäftigt sich mit dem Konstrukt Lehre und untersucht, ob die Prüfung auch alle Aspekte der Lehre beinhaltet. Bei der Kriteriumsvalidität wird zwischen Übereinstimmungs- und Vorhersagevalidität unterschieden. Bei beiden wird ein bereits validiertes Außenkriterium herangezogen, welches entweder gleichzeitig oder später gemessen wird. Dieses Kriterium und die Prüfung, welche validiert werden soll, müssen einen starken Zusammenhang aufweisen.

2.3 Regeln zur Erstellung von MC-Fragen

Für die Erstellung von MC-Fragen haben verschiedene Autoren eine Reihe von Regeln aufgestellt. Haladyna & Downing (1989) haben die bestehenden Regeln in einer Taxonomie zusammengefasst und die wissenschaftlichen Beweise für die Gültigkeit jeder Regel untersucht. Während die Validität einiger Regeln in empirischen Studien nachgewiesen wurde, gibt es für mehr als die Hälfte der Regeln keine Studien. Diese Regeln werden meist theoretisch begründet, sind Werturteile oder Expertenkonsens. In Abbildung 2 sind die wichtigsten Regeln dargestellt.

Inhalt

1. Erstellen Sie jede Frage auf der Grundlage eines spezifischen Inhalts, den diese Frage prüfen soll.
2. Halten Sie die spezifischen Inhalte der Fragen unabhängig voneinander.
3. Vermeiden Sie übermäßig spezielle und übermäßig allgemeine Fragen.
4. Vermeiden Sie Fragen auf der Grundlage von Meinungen.
5. Vermeiden Sie absichtliche „Fallen" in Fragen.
6. Lassen Sie die Fragen von Kollegen inhaltlich kontrollieren.

Format

7. Legen Sie die zulässigen MC-Fragenformate (z.B. Einfachauswahl aus drei bis fünf Antwortoptionen) vorher fest.

Stil

8. Halten Sie das Vokabular und die Grammatik der Fragen so einfach wie möglich.
9. Kontrollieren Sie die Grammatik, Interpunktion und Rechtschreibung der Fragen.
10. Halten Sie den Leseaufwand der Fragen so gering wie möglich.

Fragenstamm

11. Formulieren Sie den Stamm als Frage, die auch ohne Kenntnis der Antwortoptionen beantwortet werden kann.
12. Stellen Sie sicher, dass die Frage klar formuliert ist.
13. Formulieren Sie den Stamm positiv, vermeiden Sie negative Formulierungen wie *nicht* oder *außer*.

Antwortoptionen

14. Verwenden Sie so viele Antwortoptionen wie sinnvoll möglich. Oft ist es nicht möglich, mehr als zwei plausible falsche Antwortoptionen zu finden.
15. Stellen Sie sicher, dass eine Antwortoption eindeutig die richtige oder beste Antwort ist.
16. Ordnen Sie die Antwortoptionen logisch an (z.B. aufsteigende Zahlenwerte oder alphabetisch).
17. Die Antwortoptionen sollten unabhängig voneinander sein und sich nicht überlappen.
18. Die Antwortoptionen sollten inhaltlich homogen und ungefähr gleich lang sein.
19. Formulieren Sie die Antwortalternativen positiv, vermeiden Sie negative Formulierungen wie *nicht* oder *außer*.
20. Vermeiden Sie Hinweise auf die richtige Antwort, wie
    - Worte wie *nie, immer* oder *vollkommen* in falschen Antwortoptionen,
    - die Wiederholung eines Wortes aus der Frage in der richtigen Antwort,
    - offensichtlich falsche, absurde oder lächerliche falsche Antwortoptionen.
21. Verwenden Sie häufige und typische Fehler oder Missverständnisse als Grundlage für die falschen Antwortoptionen.

Abbildung 2: Regeln zur Erstellung von MC-Fragen (modifiziert nach Haladyna, 2004)

Eine häufig genannte Grundregel der MC-Fragen-Erstellung ist die so genannte „Antwortoptionen-abdecken-Regel" (engl. cover-the-options-rule). Sie besagt, dass eine gute MC-Frage nach dem Abdecken der Antwortoptionen auch als Freitextfrage beantwortbar sein muss. Durch die Einhaltung dieser einfachen Regeln wird sichergestellt, dass die Frage auf einen Inhalt fokussiert ist, dass die zur Beantwortung notwendigen Informationen in der Frage und nicht in den Antwortoptionen enthalten sind und dass die Frage klar und verständlich formuliert ist. Ein häufiger Fehler bei der Formulierung der Antwortoptionen ist die Meinung, unbedingt drei oder besser vier falsche Antwortoptionen haben zu müssen, da die richtige Antwort sonst zu leicht zu erraten sei. Oft ist es auch für geübte Fragenautoren nicht möglich, mehr als zwei plausible und attraktive Distraktoren zu finden. Distraktorenanalysen haben gezeigt, dass der dritte und

vierte Distraktor häufig so offensichtlich falsch ist, dass er auch von den leistungsschwächsten Prüfungsteilnehmern nicht gewählt wird. Rodriguez (2005) zeigt, dass Fragen mit drei Antwortalternativen vergleichbare Trennschärfen und Reliabilitäten aufweisen wie Fragen mit vier oder fünf Antwortalternativen. Das Hinzufügen unplausibler Distraktoren verlängert damit nur die Erstellungs- und Lesezeit, während Ratewahrscheinlichkeit, Trennschärfe und Reliabilität gleich bleiben.

Besonders wichtig für die Sicherung der Fragenqualität ist der wiederholte Einsatz von Fragen. Häufig kann die Schwierigkeit und Trennschärfe einer Frage erst durch den wiederholten Einsatz zuverlässig beurteilt werden. Beim wiederholten Einsatz von Fragen ist die Geheimhaltung der Fragen für den Erhalt der bekannten Schwierigkeit und Trennschärfe unerlässlich. Werden Fragen veröffentlicht, werden sie in der Regel leichter und ihre Trennschärfe nimmt ab. Dies ist insbesondere dann ein Problem, wenn eine konkrete Prüfung aus einem kleinen Gesamtpool von Fragen zusammengestellt wird. Enthält ein Fragenpool mehrere tausend Fragen, wird es auch bei Veröffentlichung der Fragen für die Prüfungsteilnehmer irgendwann unökonomisch, die Fragen auswendig zu lernen.

Weiterführende Literatur zur Fragenerstellung findet sich unter anderem bei Case & Swanson (2002), Haladyna & Downing (1989), Haladyna et al. (2002) und Osterlind (1992).

3. Praktische Umsetzbarkeit

Ein wichtiges Kriterium für den dauerhaften „Erfolg" eines Prüfungsformats ist seine Akzeptanz bei Studierenden und Lehrenden. Diese Akzeptanz ist für MC-Prüfungen von Fach zu Fach unterschiedlich und hängt eng mit der Häufigkeit des Einsatzes und der Qualität der MC-Fragen zusammen. Während MC-Prüfungen bei ihrer Einführung häufig als objektive und faire Alternative zu den als subjektiv und unfair empfundenen mündlichen Prüfungen gelobt wurden, hat ihr übermäßiger Einsatz in einigen Fächern, wie zum Beispiel in der Medizin, zu weit verbreiteter Kritik sowohl der Studierenden als auch der Lehrenden an diesem Prüfungsformat geführt. Dabei ist der häufigste Kritikpunkt, dass mit MC-Prüfungen überwiegend (praxisirrelevantes) isoliertes Detailwissen geprüft würde, nicht durch das MC-Format bedingt. Er lässt sich durch die sorgfältige Gestaltung der Fragen vermeiden. Die Kritik, dass MC-Prüfungen einen ungünstigen Einfluss auf das Lernverhalten der Studierenden hätten, ist dann berechtigt, wenn ausschließlich mit MC-Prüfungen geprüft wird und dadurch das Lernen nicht „MC-geeigneter" Inhalte behindert wird. Auch das ist jedoch kein Problem der MC-Prüfung, sondern der Prüfungsverantwortlichen.

Ein weiteres wichtiges Kriterium für die Auswahl eines Prüfungsformats ist dessen Einfluss auf das Lernverhalten der Studierenden. Dabei wird durch das Prüfungsformat sowohl der Inhalt als auch die Art des Lernens beeinflusst. Wenn mit MC-Prüfungen überwiegend isoliertes Detailwissen geprüft wird und diese MC-Fragen veröffentlicht werden, wird das unweigerlich dazu führen, dass die Studierenden hauptsächlich mit alten Prüfungsfragen lernen. Wenn nur „MC-geeignete" Inhalte, also kognitives Wissen, geprüft wird, weil die Prüfung anderer Inhalte, zum Beispiel praktischer Fertigkeiten oder sozialer Kompetenzen, zu aufwändig und teuer ist, werden die geprüften Inhalte von den Studierenden unweigerlich als wichtiger wahrgenommen und intensiver gelernt als die nicht geprüften Inhalte.

Zwar sind MC-Fragen in der Durchführung und Auswertung kostengünstig, aber ihre Erstellung ist aufwändig und teuer. Neue MC-Fragen-Autoren müssen geschult werden (6-8 Stunden) und für die Erstellung einer neuen MC-Frage inklusive Korrektur und Überarbeitung muss mindestens eine Stunde veranschlagt werden. Auch die Zusammenstellung einer MC-Prüfung nach gewichteten Ausbildungszielen mit unterschiedlichen Schwierigkeiten und guten Trennschärfen kostet Zeit.

4. Maßnahmen zur Sicherung der Qualität von MC-Prüfungen

Die Validität einer MC-Prüfung kann durch eine Reihe von Maßnahmen gesichert bzw. erhöht werden. Jede MC-Prüfung sollte auf der Grundlage von verbindlichen, schriftlich veröffentlichten Ausbildungszielen erstellt werden. Die Gewichtung der einzelnen Ausbildungsziele in der MC-Prüfung (Anzahl der Fragen pro Ausbildungsziel) sollte vor der Zusammenstellung der Prüfung festgelegt und dokumentiert werden. In der MC-Prüfung sollten ausreichend Fragen enthalten sein, um alle Ausbildungsziele zu prüfen. Die Relevanz und Repräsentativität der Fragen für das betreffende Ausbildungsziel kann zusätzlich noch von Experten beurteilt werden. Weitere aufwändigere Möglichkeiten der Sicherung der Validität von MC-Prüfungen sind die Bestimmung der Vorhersagevalidität (zum Beispiel durch die Korrelation der Prüfungsergebnisse mit den Ergebnissen späterer Prüfungen oder mit Erfolgsparametern im späteren Berufsleben) oder der Übereinstimmungsvalidität (zum Beispiel durch die Korrelation der Prüfungsergebnisse mit den Ergebnissen anderer Prüfungen, die den gleichen Inhalt prüfen sollen).

Die Reliabilität einer MC-Prüfung kann dadurch gesichert bzw. erhöht werden, dass die inhaltsirrelevante Varianz möglichst gering gehalten wird. Dazu zählt die Einhaltung der formalen Richtlinien zur Erstellung von MC-Fragen, so können u.a. unbeabsichtigte Lösungshinweise verhindert werden. Wie schon erwähnt, kann die Reliabilität auch durch die Steigerung der Fragenanzahl erhöht werden. Dabei muss aber berücksichtigt werden, dass – wie bereits angemerkt – eine zu lange Prüfung zu Ermüdungseffekten und damit zu einer Erhöhung der inhaltsirrelevanten Varianz führt.

Die Objektivität einer MC-Prüfung wird durch eine standardisierte Durchführung (jeder Studierende wird unter den gleichen Bedingungen geprüft) und eine standardisierte Auswertung gewährleistet. Eine hohe Objektivität ist bei MC-Prüfungen im Gegensatz zu mündlichen Prüfungen relativ leicht zu erreichen.

Nach jeder Prüfung sollten die Itemstatistiken ermittelt werden, vor allem Schwierigkeit und Trennschärfe, und auffällige Fragen noch einmal betrachtet und ggf. herausgenommen oder umkodiert werden. Die Rückmeldung aus den Itemstatistiken ist auch wichtig für den nächsten Prüfungsprozess. So kann man z.B. gut erkennen, ob eine Frage tatsächlich die Lehre repräsentiert oder formale Richtlinien zur Erstellung von MC-Fragen eingehalten wurden.

Die Fragen sollten in einen kontinuierlichen Qualitätssicherungsprozess eingebunden werden. So kann man sie vor dem Test von Experten auf inhaltliche und formale Kriterien prüfen lassen. Nach der Ermittlung der Itemstatistiken werden auffällige Fragen wieder denselben Experten vorgelegt, die dann aus der Rückmeldung der Prüfung für die zukünftige Beurteilung von neuen und auch alten Fragen weiter geschult werden. Wenn bereits verwendete Fragen für eine Prüfung gezogen werden, können sie anhand der Itemstatistik noch geändert werden. So können Distraktoren ausgetauscht oder der Fragenstamm eindeutiger formuliert werden. Weiterhin sollte die Frage unbedingt auf Aktualität geprüft werden. Im Laufe der Zeit wird ein Itempool entstehen, der hochwertige Fragen enthält. Eine weitere positive Entwicklung ist die Schulung der Experten, so dass auch neue Fragen immer besser werden.

Zusammenfassend kann gesagt werden, dass die Erstellung qualitativ hochwertiger MC-Prüfungen ein aufwändiger, langwieriger, aber lohnender Prozess ist. In der Praxis haben sich hier zentrale Prüfungsbereiche an Fakultäten bewährt, deren Mitarbeiter sich auf die Vorbereitung, Durchführung und Auswertung von Prüfungen spezialisiert haben und diese Aufgaben für alle Institute einer Fakultät übernehmen.

Literatur

American Educational Research Association, American Psychological Association, National Council on Measurement in Education (1999): *Standards for educational and psychological testing*. Washington, DC.

CASE, S.M., SWANSON, D.B. (2002): *Constructing written test questions for the basic and clinical sciences*. Philadelphia.
HALADYNA, T.M., DOWNING, S.M. (1989): "A taxonomy of multiple-choice item-writing rules". In: *Applied Measurement in Education*, 2, S. 37-50.
HALADYNA, T.M., DOWNING, S.M., RODRIGUEZ, M.C. (2002): "A review of multiple-choice item-writing guidelines for classroom assessment". In: *Applied Measurement in Education*, 15, S. 309-333.
HALADYNA, T.M.(2004): *Developing and validating multiple-choice test items*. Mahwah NJ (3. Aufl.).
INSTITUT FÜR AUS-, WEITER- UND FORTBILDUNG DER MEDIZINISCHEN FAKULTÄT UNIVERSITÄT BERN (1999): *Kompetent prüfen*. Bern. Online verfügbar unter: http://www.iawf.unibe.ch/aae/redirect.asp?ID=74 [27.03.2007].
ITTEN, S., KREBS, R. (1997): *Messqualität der verschiedenen MC-Itemtypen in den beiden Vorprüfungen des Medizinstudiums an der Universität Bern*. Bern. Online verfügbar unter: http://www.iawf.unibe.ch/aae/documents/publikationen/Itemtypen_Vorklinik.pdf [26.02.2007].
LIENERT, G.A., RAATZ, U.(1998): *Testaufbau und Testanalyse*. Weinheim (6. Aufl.).
OSTERLIND, S.J.(1992): *Constructing test items*. London.
RODRIGUEZ, M.C.(2005): "Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research". In: *Educational Measurement issues and practice*, 24, S. 3-13.

Simone Scheffer, Annette Fröhmel, Waltraud Georg:

Prüfung praktischer Fertigkeiten: Performance-based Testing in der Medizin

Zusammenfassung

In der ärztlichen Ausbildung hat die Prüfung klinisch-praktischer Fertigkeiten in den letzten 30 Jahren an Bedeutung gewonnen. Entscheidend dafür war die Entwicklung eines neuen Prüfungsformats, das gegenüber der traditionellen Prüfung am Krankenbett viele Vorteile aufweist, die Objective Structured Clinical Examination (OSCE). Mittels dieser Methode kann ein vielfältiges Aufgabenspektrum standardisiert geprüft werden. Dabei kommen statt realen Patienten Simulationspatienten und Modelle zum Einsatz. Simulationspatienten sind Personen, die dazu ausgebildet werden, eine Patientenrolle möglichst realitätsnah darzustellen. Mit dem OSCE kann eine entsprechend den Testgütekriterien (Objektivität, Reliabilität, Validität) qualitativ hochwertige Prüfung durchgeführt werden, die bei den Studierenden auf hohe Akzeptanz trifft. Weiterentwicklungen des Formats als Beobachtung und Beurteilung des Verhaltens im klinischen Alltag finden vor allem in der Facharztausbildung Anwendung. Auch für andere Fächer könnte dieses Prüfungsformat zur Überprüfung praktischer Fertigkeiten interessant sein und das Methodenspektrum ergänzen.

Gliederung

1. Zur Relevanz praktischer Prüfungen in der Medizinerausbildung
2. Historischer und theoretischer Kontext
   2.1 Von der „Prüfung am Krankenbett" zur strukturierten Prüfung praktischer Fertigkeiten
   2.2 Theoretische Einordnung von performance-based testing
3. Objective Structured Clinical Examination (OSCE)
   3.1 Bewertungsinstrumente und Rollenverteilung
   3.2 Prüfung kommunikativer Kompetenzen
   3.3 Auswertung der Prüfung
4. Der Stellenwert von „Simulationspatienten" für die Prüfung praktischer klinischer Fertigkeiten
   4.1 Einsatz von Simulationspatienten