Format und Durchführung schriftlicher Prüfungen

Format and implementation of written assessments

• Johannes Schulze[1] • Stefan Drolshagen[1]

Zusammenfassung:
Alle Leistungsnachweise des klinischen Studienabschnittes nach neuer Ärztlicher Approbationsordnung müssen benotet werden; hierzu sind in der Regel schriftliche Prüfungen notwendig. Bisher erprobte Methoden beinhalten die Prüfung passiven Wissens (Einfachauswahlfragen, multiple choice-Fragen, progress test-Fragen) und aktiven Wissens (short essay questions, long essay questions). Vor- und Nachteile dieser Verfahren werden diskutiert, sowie die zur Erstellung, Durchführung und Auswertung schriftlicher Prüfungen notwendigen Ressourcen.
Schlüsselwörter:Einfachauswahlverfahren, MC-Fragen, short essay questions, long essey questions, Prüfungen, Validität, Reliabilität, Objektivität, Ressourcen

Abstract:
According to the new German licensing regulations all clinical certificates have to be graded; mostly this requires written tests. Methods tested to date include tests for passive knowledge (single choice questions, multiple choice questions, progress test) and active knowledge (short essay questions, long essay questions). We discuss advantages and disadvantages of these methods, as well as the resources required for test preparation, testing and grading of student answers.
Keywords:single choice questions, multiple choice questions, short essay questions, long essay questions, assessment, Validity, reliability, objecitvity, ressources

Einleitung

Mit der im Juni 2002 neu gefassten Ärztlichen Approbationsordnung (ÄAppO) wird ein beträchtlicher Teil der universitären Prüfungen in der Humanmedizin wieder auf die Fakultäten zurückverlagert. Während bisher zwar Leistungsnachweise verlangt wurden, stellte es die frühere ÄAppO den Fachbereichen frei, wie die erfolgreiche Teilnahme nachgewiesen wurde; nicht selten verzichteten einzelne Disziplinen hierauf. Nun müssen im klinischen Studienabschnitt benotete Leistungsnachweise in insgesamt 39 Fächern, Querschnittsbereichen und Blockpraktika erbracht werden. Auch jetzt ist den Fachbereichen die Form der Notenfindung freigestellt; die Notwendigkeit der Benotung macht es jedoch erforderlich, studentische Leistungen nachvollziehbar und überprüfbar zu messen. Die bisher während des klinischen Studienabschnittes durchgeführten Teile des Staatsexamens, frühestens nach dem 2. bzw. 6. klinischen Semester, entfallen; der neue 2. Abschnitt wird als fallbezogene schriftliche Abschlussprüfung nach dem Praktischen Jahr durchgeführt.

Prüfungsformate sollen nach den Lernzielen, der Ausrichtung des studentischen Lernens und den Ressourcen ausgewählt werden [22]; in der Regel wird vor allem im klinischen Studium nur eine Mischung verschiedener Formate zu einer guten Prüfung führen. Darüber hinaus eignen sich schriftliche Prüfungen nicht für alle Lernziele. Praktische Fähigkeiten und Routineanwendungen lassen sich nur mit anderen Methoden wie OSCE prüfen [10], [15], [23]. Auch wenn dabei die Ergebnisse beider Prüfungsmethoden stark korrelieren, so werden doch unterschiedliche Wissensbereiche geprüft [1], [25]. Die klinische Kompetenz wird dabei von der Summe beider Tests besser vorhergesagt als von den Einzeltests [1], [25]. Dieser Artikel befasst sich nur mit schriftlichen Prüfungen; für eine Bewertung anderer Formate wie OSCE [11], [14], [15], [24] oder globale Beurteilungen in Praktika [11], [22] sei auf die entsprechenden Quellen verwiesen [12].

Methoden

• Prüfungsziele

Jede Prüfung hat sich den Kriterien Objektivität, Reliabilität und Validität zu unterwerfen [6]. Die lange Diskussion um den Stellenwert von Objektivität, Reliabilität und Validität soll hier nicht wiederholt werden [3], [16], [17]. Eine gute Prüfung erreicht, dass die Studierenden sich den Lernstoff so aneignen, dass er später in einer den Anforderungen entsprechenden Weise präsent ist [22]. Die Prüfungsart soll sicherstellen, dass der Lernstoff in einer geeigneten Art - aktives oder passives Wissen - wiedergegeben werden kann; sie steuert dadurch das Lernverhalten der Studierenden [9]. Ideal erscheint auf dem ersten Blick damit eine Prüfungsart, die den späteren Erfordernissen möglichst nahe kommt (direct validity, [7]). Hierbei wird als Charakteristikum angesehen, dass eine gute Prüfung mehr durch den Inhalt charakterisiert wird als durch die Prüfungsergebnisse. Sollen diese Punkte erreicht werden, sind Chancen und Limitationen von schriftlichen Prüfungsverfahren anhand der angestrebten Lernziele gegeneinander abzuwägen.

Im Folgenden sollen die weiter verbreiteten schriftlichen Prüfungsformate kurz skizziert werden, für eine ausführliche Bewertung sei insbesondere auf das Handbuch "Kompetent prüfen" des IAWF [13] in Bern verwiesen.

• Einfachauswahlverfahren

Bei diesem Fragentyp ist unter fünf vorgegebenen Möglichkeiten eine und nur eine Möglichkeit zutreffend und muss markiert werden. Vom Prüfling wird verlangt, sich zwischen mehreren Mög-

[1] *Johann Wolfgang Goethe-Universität Frankfurt/Main, Dekanat des Fachbereichs Medizin, Frankfurt/Main, Deutschland*

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000263.shtml
Bitte zitieren als: Schulze J, Drolshagen S. Format und Durchführung schriftlicher Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc44.
gQp020060785 31

lichkeiten zu entscheiden; wenn dies das Lernziel ist, ist dieser Fragentyp sehr gut geeignet.

Dieser Fragentyp (im Folgenden als MC-Fragen bezeichnet) hat durch die Verwendung in den schriftlichen Staatsexamina über die letzten drei Jahrzehnte die medizinischen Prüfungen dominiert. Er ist für universitäre Prüfungen durch Gewohnheit, aber auch durch die verwaltungsgerichtliche Überprüfung der Fragen [26] sehr beliebt. Die neue ÄAppO schreibt diesen Fragentyp nicht explizit vor, sondern legt lediglich ein Antwort-Wahl-Verfahren und die Vorgabe möglicher Antworten fest (ÄAppO, §29,3).

MC-Fragen sind

- schnell und reproduzierbar auszuwerten (hohe Objektivität)

- und beinhalten durch viele Fragen viele Items des Stoffgebietes (hohe Reliabilität).

Auch wenn das Kriterium der rechtlichen Überprüfung allein kein Gütekriterium ist - wie die sehr geringe Zahl von Anfechtungsklagen beim prinzipiell anders strukturierten Abitur mit einem ähnlich hohen Selektionscharakter zeigt - wird dieses Argument doch immer wieder für MC-Fragen vorgebracht. Auch im Hinblick auf die nun von den Fachbereichen zu vertretenden eventuellen Klagen wird einer bisher guten juristischen Prüfungslage ein hoher Stellenwert zugemessen.

Gute MC-Fragen enthalten neben der richtigen Antwort gute Distraktoren. Während für die Prüfung des Faktenwissens geeignete Distraktoren relativ einfach zu finden sind, gilt dies bei Zusammenhangswissen nur eingeschränkt. Hier sind gute Fragen nur schwer zu konstruieren, durch das Einfügen von Schlüsselworten kann die richtige Antwort häufig auch ohne Wissen erraten werden. Sie werden durch ungenaue oder irreführende Formulierungen schnell angreifbar. Die schwierige Trennung zwischen "falsch" und "richtig" vor allem im klinischen Studienabschnitt (Lehrmeinungen in unterschiedlichen Büchern) kann durch MC-Fragen mit der geforderten Trennschärfe nur schwer nachgebildet werden.

Die Erstellung guter MC-Fragen ist sehr aufwendig. Hierbei ist weniger die Formulierung der "einen" anzukreuzenden Möglichkeit schwierig, sondern die Identifikation geeigneter Distraktoren ("medical education is awash with seriously flawed test questions" [5]). Ein guter Distraktor erscheint zwar plausibel und hat einen engen Zusammenhang zur Fragestellung, andererseits ist er aber eindeutig falsch, und vier geeignete Distraktoren müssen gefunden werden. Eine zu große Ferne von der richtigen Lösung wird die Frage sehr leicht machen, eine zu große Nähe macht die eindeutige Einordnung als "falsch" schwierig (construct irrelevant variations, [3]). Die Stellung von Fragen einer geeigneten Schwere wird damit in vielen Bereichen problematisch, ebenfalls schwierig wird die Zusammenstellung einer Prüfung, die sowohl anspruchsvoll ist, aber auch alltägliches medizinisches Wissen sowie Standardprozeduren beinhaltet (construct underrepresentation; [3]). Diese Faktoren führen dazu, dass der in geeigneter Weise zu prüfende Stoff vor allem in der Klinik limitiert ist, die Verwendung im vorklinischen Bereich erscheint dagegen weniger problematisch.

• True-False-Aufgaben

Auswahlfragen können auch so strukturiert werden, dass eine Frage eine oder mehrere Antwortmöglichkeiten enthält, die unabhängig bewertet werden müssen (z.B. in Progress Tests). Hierdurch entfallen Distraktoren, da der Zwang zur Entscheidung zwischen Alternativen nicht besteht. Ein Beispiel hierfür ist die Führerscheinprüfung. Neben den Möglichkeiten "richtig" und "falsch" wird in Progress-Test-Prüfungen oft auch die Möglichkeit "weiß ich nicht" angeboten; die Verwendung dieser Option ist umstritten [22]. True false Fragen können mit einer beliebigen Anzahl von Möglichkeiten gestellt werden; hierzu gehört auch, dass die Fragen bei fünf zu bewertenden Aussagen formal identisch zu den MC-Fragen aussehen, dass aber nichtsdestotrotz jede der Alternativen individuell beantwortet werden muss.

Eine Modifikation dieser Fragen sind "k aus n-Fragen" [13]. Hierbei werden n Antwortalternativen gegeben, von denen k richtig sind und markiert werden müssen. Hier bleibt ein größerer Bereich für Alternativentscheidungen durch den Zwang, sich auf k Antworten festzulegen.

Beide Fragetypen (MC-Fragen, true-false-Fragen) testen das Wiedererkennen der richtigen Lösung bzw. das Wiedererkennen, ob eine Aussage zutrifft. Sie eignen sich damit für den Bereich des passiven Wissens. Durch die zusätzliche Informationsgabe können auch Entscheidungsprozesse imitiert werden; für Wissen, welches später ohne Vorgaben präsentiert werden soll, sind sie nicht geeignet.

• Kurzantwort - Fragen

Kurzantwortfragen (short essay questions SEQ) erfordern das Niederschreiben der richtigen Antwort (oder eine der möglichen richtigen Antworten) in freier Form. Damit wird primär aktives Wissen geprüft, der "cueing"-Effekt durch Wiedererkennen einer Antwort entfällt. Kurzantwortfragen werden in einer Reihe gebräuchlicher Versionen verwendet, die als "short essay questions", "short answer questions", "modified essay questions" oder "short answer management problems" bezeichnet werden [12]. Short essay-Fragen wurden als Alternative zu MC-Fragen entwickelt und derzeit vor allem in angelsächsischen Ländern als Ersatz oder Ergänzung zu MC-Fragen eingesetzt [18]; in Deutschland hat diese Frageform bisher keine weite Verbreitung gefunden. Wie bei MC-Fragen erfordern gute short essay-Fragen eine sorgfältige Formulierung und Erstellung des Lösungsthesaurus. Das freie Niederschreiben erlaubt auch, vermeintlich einfache Sachverhalte wieder zu prüfen; sie sind vor allem für Kontext-reiche Fragen geeignet [22].

Als Vorteil von Kurzantwort-Fragen ist zu werten, dass durch die aktive Reproduktion in einigen Bereichen die klinische Realität besser dargestellt werden kann (Vervollständigung einer Anamnese, Anforderung weiterführender Diagnostik, Befundung). Dies erlaubt eine Ausweitung des abzufragenden Wissensbereiches. Bei der Prüfung von Routinewissen, bei dem auch im Alltag a priori keine Vorgaben für die Tätigkeiten gegeben werden, können Kurzantwort-Fragen das aktiv vorhandene Wissen überprüfen. Sie sind dementsprechend in modifizierter Form oft auch Bestandteile praktischer Prüfungen [11], [21].

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000263.shtml
Bitte zitieren als: Schulze J, Drolshagen S. Format und Durchführung schriftlicher Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc44.
gky(020060703.1)

Eine Alternative zu reinen SEQ-Fragen sind dabei "extended matching"-Fragen (EMQ), bei denen die richtige Lösung aus einem Thesaurus entnommen werden muss. Wenn dieser hinreichend groß ist und Schreibvarianten sowie Verwechslungen beinhaltet, eignen sich EMQ-Fragen in gleicher Weise wie SEQ [22]. Dies gilt insbesondere, wenn der Lösungsthesaurus für mehrere Fragen zusammengefasst ist.

Auch gute SEQ-Fragen sind schwer zu erstellen und erfordern einen hohen Aufwand. Nachteilig ist zudem der höhere Auswertungsaufwand, vor allem durch die Möglichkeit von Schreibvarianten oder mehreren richtigen Lösungen. Dies verringert auch deutlich die Reliabilität und Objektivität dieses Fragetyps.

• Long essay - Fragen

In Long essay-Fragen werden, analog zum Abitur, offene Fragen gestellt, die mit einem längeren Text beantwortet werden. Gefordert wird hierbei die schriftliche Darlegung von Zusammenhängen oder die Begründung einer Entscheidung; sie testen hierdurch nicht die Entscheidung selbst, sondern das dieser Entscheidung zugrunde liegende Wissen. Diese Fragen testen vor allem Zusammenhangswissen; eine Anwendung auf konkrete Probleme ist möglich, aber nicht unbedingt erforderlich.

Long essay-Fragen sollten ebenfalls sorgfältig formuliert werden; die längere Bearbeitungszeit reduziert die Anzahl der für eine Prüfung notwendigen Fragen (verminderte Reliabilität). Problematisch ist auch die Auswertung; Objektivität und Reliabilität werden durch Auswertung durch mehrere Personen sowie eine Aufgabengebundene Bewertung erhöht (jeder Auswerter bearbeitet eine Frage für alle Studierenden).

Bisher ungeklärt ist die Validität dieses Verfahrens in der Medizin. Zumindest bei problemorientierten Fragen erscheint prima vista ein größerer Zusammenhang mit dem späteren Patientenmanagement gegeben. Durch das Fehlen von Vorgaben berücksichtigen long essay-Fragen auch das aktive Wissen, und sprechen hierdurch einen anderen Wissensbereich an [25]. Die Auswertung von long essay-Prüfungen ist schwierig und nur durch vorgegebene Bewertungslisten objektivierbar. Hierdurch wird der im Vergleich zu anderen Fragenformaten niedrige Entwicklungsaufwand auf die Auswertung verlagert.

Für alle Fragenformate bestehen ähnliche Probleme, die einer jeden schriftlichen Prüfung inhärent sind. Einige wichtige Aspekte sollen im Folgenden dargestellt werden, die bei der Auswahl eines Formates berücksichtigt werden müssen:

Prozedurale Überlegungen

• Ratewahrscheinlichkeit

Studierende und Prüfer sind sich meist darüber im Klaren, dass ein Teil der richtigen Antworten durch geschicktes Raten erreicht werden kann. Dies kann eine erwünschte Eigenschaft sein, die den ärztlichen Alltag widerspiegelt; auch hier spielt die Unsicherheit und das "richtige" Raten der Diagnose und geeigneten Therapie eine große Rolle. Raten sollte aber begrenzt sein und bleiben, bei der Überprüfung des Verständnisses oder der Fähigkeit, Schlussfolgerungen zu begründen, sollte der Ratefaktor so klein wie möglich sein.

Ob Raten grundsätzlich schlecht ist, oder ob Raten als "begründete Entscheidungsfindung bei unsicherem Wissen" ein wesentliches Element der ärztlichen Tätigkeit gilt, ist umstritten [2], [4], [6]. Dabei sollte unterschieden werden zwischen "blindem Raten" bei völliger Unkenntnis des Lernstoffes, eine eher artifizielle Annahme. Für dieses Raten wird als Szenario typischerweise angenommen, ob ein Laie eine Prüfung durch Raten bestehen kann. Dieses Raten wird wesentlich durch die Formulierung beeinflusst (z.B. durch "immer", "nie", "ausschließlich" als cueing-Faktoren), d.h. versteckte Hinweise auf die richtige Antwort in der Fragenformulierung. Vor allem im klinischen Bereich gibt es kaum ein "immer" oder "nie"; die Verwendung dieser Wörter markiert eine Möglichkeit sofort als unzutreffend. Ein häufigerer cueing-Effekt besteht bei Widerspruch zwischen zwei Aussagen; in diesem Falle ist meist eine der beiden Aussagen richtig. Es gibt derzeit wenig belastbare Untersuchungen zur Stärke dieses Effektes [22]. Bei einem Vergleich von Prüfungsfragen, die als MC-Frage, alternativ dazu als Frage mit mehreren möglichen richtigen Antworten (zwischen 0 und 5 richtigen Möglichkeiten) präsentiert wurde, zeigte sich, dass ohne Vorgaben weniger als die Hälfte der Studierenden die vollständig richtige Antwort wissen, bei MC-Fragen die richtige Antwort aber zu >80% geraten wird [19]. Cueing ist jedoch nicht auf MC-Fragen beschränkt. Da dieser Effekt formulierungsbedingt ist, ist dieses Phänomen auch bei anderen Formen der passiven Wissensüberprüfung in wechselndem Ausmaß präsent.

Daneben besteht Raten auch darin, bei MC-Fragen von den 5 Alternativen drei auszuschließen und zwischen den beiden verbleibenden Variablen eine begründete Entscheidung zu treffen ("scientific guess"). Dieses zweite Raten kommt vielen Entscheidungssituationen nahe und entspricht einem häufig akzeptablen, wenn nicht erwünschten Lernziel; dementsprechend kann das begründete Entscheiden ein angestrebtes Lernziel sein. Begründetes Entscheiden wird wesentlich durch den Fragenkontext beeinflusst und wird durch die Angaben im Fragenstamm gesteuert [22].

Generell gilt, dass bei einer Prüfung aktiven Wissens die Ratewahrscheinlichkeit gering ist. Bei true false-Aufgaben (ohne die Option "weiß ich nicht") besteht eine Ratewahrscheinlichkeit von 50%. Bei der Bewertung besteht einerseits die Möglichkeit, Falschantworten durch Minuspunkte zu "bestrafen", was für eine Gesamtklausur zu einer Ratewahrscheinlichkeit von 0% führt (ausgeglichene Zusammensetzung aus richtigen und falschen Aussagen vorausgesetzt). Da die Vergabe der Minuspunkte kritisch gesehen und auch aus Akzeptanzgründen eher abgelehnt wird [22], beträgt ohne Minuspunkte die Ratewahrscheinlichkeit 50% richtige Antworten.

• Trainierbarkeit

Sowohl die reine Ratewahrscheinlichkeit als auch cueing-Effekte können trainiert werden und begründen den Erfolg spezifischer, Format-abhängiger Prüfungsvorbereitungen. Dieser Effekt allein führt dazu, dass nach einer längeren Beibehaltung eines dominierenden Formates dieses bereits deshalb schlecht wird, weil es seit langer Zeit unverändert verwendet wird, nicht weil das Format als solches schlecht ist. In gleicher Weise wird jedes Prüfungsformat, welches durch eine spezifische nichtmedizinische Vorbereitung besser bestanden wird, nach einer längeren Verwendung zur Anpassung der studentischen Vorbereitung führen. Geeignet ist dann diejenige Prüfungsform, die zu einem vom Lehrenden gewünschten Lernverhalten führt. Bei allen anderen Verfahren kann die Dauer

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000263.shtml
Bitte zitieren als: Schulze J, Drolshagen S. Format und Durchführung schriftlicher Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc44.
gQp020060705 31

der Verwendung allein ein Parameter für eine zunehmend schlechtere Prüfungsqualität sein.

Die jetzt seit etwa 30 Jahren dominierenden MC-Prüfungen sind auch deshalb in Verruf gekommen, weil sie Altfragenlernen und Wiedererkennen belohnen, aktives Wissen jedoch nicht. Eine "Abkühlphase" über etwa 10 Jahre würde sicherlich gut tun. Generell ist vorzuziehen, Prüfungen mit einer Mischung verschiedener Fragentypen durchzuführen.

• Bestehensgrenzen

Zwischen dem Fragenformat einer Prüfung, dem angestrebten Bereich des geprüften Wissens (werden häufige oder alltägliche Sachverhalte thematisiert?) und der Bestehensgrenze besteht ein enger Zusammenhang (siehe Abbildung). Vor allem Prüfungsfragen des passiven Wissens - Wiedererkennen vorgegebener Lösungsalternativen - werden auch durch schwächere Studierende oft richtig beantwortet. Wird in diesem Format Grundlagenwissen abgefragt (schlechte Distraktoren), muss die Bestehensgrenze hoch liegen, damit der Selektionscharakter erhalten bleibt. Dies zeigt sich exemplarisch in praktischen Prüfungen nach OSCE (objective structured clinical examinations), bei denen die Bestehensgrenze je nach Station zwischen 60 und 80% der maximal erreichbaren Punktzahl liegt, um dem rechnerischen Einfluss der hohen Anzahl richtiger Items in "leichten" Stationen Rechnung zu tragen. Auch bei der Führerscheinprüfung liegt die Bestehensgrenze mit etwa 90% der Fragen sehr hoch. Andererseits muss die Bestehensgrenze bei Prüfungen mit "schwerem" Stoff niedrig liegen, damit nicht die Kohorte insgesamt durchfällt.

Bei den Staatsexamensfragen des IMPP sind sowohl das Fragenformat (MC-Fragen) als auch die Bestehensgrenze (60%) gesetzlich vorgegeben. Um einen Selektionscharakter zu erhalten, muss durch eine entsprechende Fragenschwere der Prüfungsdurchschnitt "eingestellt" werden, was in den letzten Jahren zum zunehmenden Fragen von Spezialwissen geführt hat. Dennoch erscheint es derzeit möglich, das Staatsexamen auch dann zu bestehen, wenn objektiv nur wenig gefestigtes Wissen vorliegt [8].

• Einfluss nichtmedizinischer Faktoren auf die Fragenbeantwortung

Vor allem bei MC-Fragen wird immer wieder argumentiert, dass Studierende durch Testtraining das Ergebnis beeinflussen können. Hierzu tragen mehrere Faktoren bei:

- schlechte Formulierung von Fragen;

- "cueing"- und "priming"-Effekt;

- fehlende Vortestung der Fragen durch Veröffentlichung;

- weite Verbreitung von Altfragensammlungen.

Probleme der Fragenformulierung und des cueing/priming wurden oben angesprochen. Vortestung ist ein essentieller Faktor bei der Generierung guter Fragen, unabhängig vom Typus. Hierbei werden bei der Erststellung der Fragen Schweregrad und häufigere Falschantworten eruiert und die Qualität der Frage bestimmt. Je stärker der Einfluss der spezifischen Formulierung auf das Antwortverhalten ist, d.h. je genauer zwischen "Richtig" und "Falsch" unterschieden werden muss, desto wichtiger ist eine Vortestung. Dies erfordert aber, dass Fragen nicht veröffentlicht werden und in der exakten Formulierung den Prüflingen unbekannt bleiben. Bei Veröffentlichung der gestellten Fragen ist eine Vortestung wegen der hohen Abhängigkeit der Fragenqualität von der exakten Formulierung nicht möglich.

• Veröffentlichung gestellter Fragen

Das Ausmaß der Fragenveröffentlichung bleibt den Fachbereichen überlassen, wird aber nicht einheitlich gehandhabt. Aus testtheoretischen Gründen erscheint eine Nichtveröffentlichung zwingend geboten, so wie es auch bei den Staatsexamina in anderen Ländern gehandhabt wird. Andererseits haben die Studierenden ein Einsichtsrecht [26], damit im Klagefall eine rechtliche Überprüfung begründet werden kann. Zwischen der Geheimhaltung und der Überprüfung muss eine geeignete Abwägung getroffen werden.

Auch die Verbreitung von "Altfragen" wird nicht zu unterbinden sein. Das studentische Bestreben einer optimalen Prüfungsvorbereitung ist verständlich, Altfragensammlungen sind hierzu ein effizienter Weg. Im Idealfall wird das gesamte zu prüfende Wissen durch Altfragen abgedeckt; in diesem Fall bilden die Altfragen den Gegenstandskatalog des Faches.

Auch für einen solchen "Gegenstandskatalog" ist das Fragenformat problematisch, da nicht alle Lernziele durch ein Fragenformat sinnvoll geprüft werden können. Bei offenen Fragen oder mündlichen Prüfungen stellen vollständige Fragensammlungen ein Lehrbuch in unorthodoxer Struktur dar (z.B. Fall-basierte Lehrbücher für die Vorbereitung zum mündlichen Examen). Für formulierungssensitive Formate (MC-Fragen, true-false-Antworten) kann dies nicht zutreffen. Eine Veröffentlichung ist damit abhängig von der Vorstrukturierung der Antwort - problematisch bei Fragen zum Ankreuzen, weniger bei SEQ und unproblematisch bei long essay-Fragen oder mündlichen Prüfungen.

• Notwendige Ressourcen

Die für die Erstellung, Durchführung und Auswertung guter schriftlicher Prüfungen notwendigen Ressourcen werden weithin unterschätzt. Der Erstellungsaufwand steigt mit der Formulierungssensitivität stark an, er ist insbesondere bei MC-Fragen mit der Notwendigkeit von vier guten Distraktoren sehr hoch. Allgemein werden richtige Alternativen von vielen Lehrenden leichter formuliert als falsche, aber plausibel erscheinende Aussagen. Hierbei sollten immer mehrere Entwickler einbezogen werden, eine gute Frage sollte zudem in einer Klausur vorgetestet sein (blueprinting, [22], [21]). Auch bei true-false-Fragen, Kurzantwortfragen oder extended matching-Fragen ist der Erstellungsaufwand nur unwesentlich geringer; für eine objektive Bewertung ist die Erstellung eines möglichst vollständigen Lösungsthesaurus aller als richtig anerkannten Möglichkeiten vor Fragenstellung notwendig. Dieser Thesaurus sollte möglichst alle zu erwartenden Antworten einschließlich der Schreibvarianten enthalten, sowie die wesentlichen erwarteten Falschantworten. Da keine Antwortalternativen vorformuliert werden müssen, beschränkt sich der Formulierungsaufwand der Fragen selbst auf den Fragenkopf. Er ist beträchtlich, wenn - wie angestrebt werden sollte [22] - Kontext-reiche Fragen gestellt werden.

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000263.shtml
Bitte zitieren als: Schulze J, Drolshagen S. Format und Durchführung schriftlicher Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc44.
gkp020060703.1)

Die Prüfungsdurchführung ist bei allen Fragentypen ähnlich. Alle aufgeführten Fragemöglichkeiten lassen sich derzeit als Papier-basierte Prüfung, bei vorhandener Hardware auch als Computer-basierte Prüfung durchführen. Der Aufwand für die Vervielfältigung der fertigen Prüfungen und für die Aufsicht ist vergleichbar und hängt eher von der Zahl der Prüflinge und dem Umfang ab.

Auswahlfragen lassen sich schnell maschinell auswerten. Das Einscannen der Bögen erfordert eine manuelle Kontrolle auf Lesefehler; die Bewertung der Fragen anhand eines vorgegebenen Bewertungsschemas kann durch Software erfolgen. Eine Anpassung der Software an unterschiedliche Bewertungskriterien ist bei "Richtig-Falsch"-Markierungen unproblematisch, bei Freitextaufgaben hoch. Derzeitige Computerprogramme lassen eine Vorbewertung der Antworten mittels Thesaurus zu; alle Antworten, die nicht im Thesaurus enthalten sind, müssen manuell bewertet werden. Dies erfordert die Korrektur durch einen entsprechend ausgebildeten Mitarbeiter, auch ein erfahrener Bewerter benötigt hierfür viel Zeit.

Die Bewertung von Freitextaufgaben sollte immer Frage für Frage erfolgen (Item-spezifisch), nicht Klausur für Klausur (Studenten-spezifisch; [6]); sie wird erleichtert, wenn die Antworten elektronisch, z.B. in einem Spreadsheet erfasst sind. Noch wesentlich höher ist der Bewertungsaufwand für Long essay Fragen, für die neben einer Musterantwort ein ausführlicher Bewertungsbogen erstellt und für jede Klausur ausgefüllt werden muss, und sollte unabhängig von mehreren Bewertern vorgenommen werden. Angesichts der Studierendenzahlen ist dieser Aufwand wohl von keinem Institut zu leisten und wird dementsprechend in Deutschland auch nicht eingesetzt.

Zusammenfassung

Kein Prüfungsformat ist per se schlecht oder gut, die Verwendung einzelner Formate sollte sich ausrichten:

- an der gewünschten Reproduktion des geprüften Stoffes (aktives vs. passives Antworten);

- an der gewünschten studentischen Vorbereitungsweise (Steuerung des Lernverhaltens);

- an der Ähnlichkeit zur späteren Tätigkeit (Praxisvorbereitung, Routinefähigkeit);

- an der bisherigen Dauer der Verwendung (Erfahrung mit Vor- und Nachteilen);

- an den zur Verfügung stehenden Ressourcen (Personalausstattung).

Im Spagat zwischen optimaler Prüfungsdurchführung und begrenztem Personal muss eine Abwägung getroffen werden. Die Auswahl eines einzelnen Formates als alleiniges Prüfungsverfahren erscheint dabei als die schlechteste aller Möglichkeiten. In Abhängigkeit vom Lernstoff, dem Ausbildungsstand und den Prüfungszielen erscheint eine Mischung verschiedener Fragenformate und eine auf das Format und den Schwierigkeitsgrad angepasste Bestehensgrenze als die beste Möglichkeit, das studentische Lernen richtig zu steuern, den geprüften Wissensstoff an die Lernziele anzupassen und damit eine möglichst valide Prüfung durchzuführen (siehe Abbildung 1).

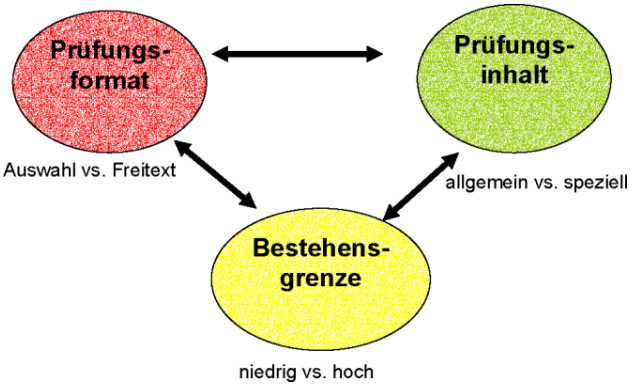

Abbildung 1: Zusammenhang zwischen Format, Inhalt und Bestehensgrenzen für "selektive" Prüfungen

Korrespondenzadresse:

• Prof. Dr. Johannes Schulze, Johann Wolfgang Goethe-Universität Frankfurt/Main, Dekanat des Fachbereichs Medizin, Theodor-Stern-Kai 7, 60590 Frankfurt/Main, Deutschland, Tel.: 069/6301-5681, Fax: 069/6301-5922
johannes.schulze@kgu.de

Literatur:

[1] Auewarakul C, Downing SM Jaturatamrong E, Praditsuwan R. Sources of validity evidence for an internal medicine student evaluation system: an evaluative study of assessment methods. Med Educ. 2005;39:276-283.

[2] Burton RF. Guessing in selected-response tests. Med Educ. 2004;37:112.

[3] Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ. 2004;38:327-333.

[4] Downing SM. Guessing on selected-response examinations. Med Educ. 2003;37:670-671.

[5] Downing SM. On guessing corrections. Med Educ. 2004;38(1):113.

[6] Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004:38;1006-1012.

[7] Ebel RL. The practical validation of tests of ability. Educ Meas: Issues Pract. 1983;2:7-10.

[8] Gebert G. Medizinstudium: Naturwissenschaftliche Grundkenntnisse nach der Vorklinik. Dt Arztebl. 2002;99:A252-253.

[9] Hakstian RA. The effects of type of examination anticipated on test preparation and performance. J Educ Res. 1971;64:319-324.

[10] Harden R, Stevenson M, Downie W, Wilson G. Assesment of clinical competence using objective structured examinations. Br Med J. 1975:,(5955):447-451.

[11] Hudson JN, Tonkin AL. Evaluating the impact of moving from discipline-based to integrated assessment. Med Educ. 2004;38:832-843.

[12] Institut für Medizinische Lehre (IML). Kompetent prüfen. Bern/Wien: IML. 1999:51-52. Zugänglich unter: http://www.iawf.unibe.ch/aae/aaecontent.asp?pg=72&ID=80.

[13] Institut für Medizinische und Pharmazeutische Prüfungsfragen (IMPP). Musteraufgaben für das Staatsexamen für Psychologische Psychotherapeuten. Mainz: IMPP: 2006. Zugänglich unter: www.impp.de/pdf/Musteraufgaben_PT.pdf.

[14] Jünger J, Schäfer S, Roth C, Schellberg D, Friedman BDM, Nikendei C. Effects of basic clinical skills training on objective structured clinical examination performance. Med Educ. 2005;39:1015-1020.

[15] Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65:63-67.

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000263.shtml
Bitte zitieren als: Schulze J, Drolshagen S. Format und Durchführung schriftlicher Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc44.

[16] Mosier CI. A critical examination of the concept of face validity. Educ Psychol Meas. 1947;7:191-205.

[17] Norcini JJ, Shea JA. The credibility and comparability of standards. Appl Meas Educ. 1997;10:39-59.

[18] Paul S. The job and the exam. Br Med J. 1999;319:Career focus.

[19] Schulze J, Drolshagen S, Nürnberger F, Ochsendorf F, Schäfer V, Brandt C. Einfluss des Fragenformates in Multiple-choice-Prüfungen auf die Antwortwahrscheinlichkeit. GMS Z Med Ausbild. 2005;22:Doc. 218. Zugänglich unter: http://www.egms.de/de/journals/zma/volume22.shtml.

[20] Schulze J, Drolshagen S, Ochsendorf F, Nürnberger F. Question format and knowledge presentation. Edinbourg: AMEE. 2004:Abstract 2G3.

[21] Schuwirth LWT, van der Vleuthen CPM. Changing education, changing assessment, changing research?. Med Educ. 2004;38:805-812. (A)

[22] Schuwirth LWT, van der Vleuthen CPM. Different written assessment methods: what can be said about their strengths and weaknesses?. Med Educ. 2004;38:974-979.

[23] Veloski JJ, Fields SK, Boex JR, Blank LL. Measuring Professionalism: A Review of Studies with Instruments Reported in the Literature between 1982 and 2002. Acad Med. 2005:80;366-370.

[24] Wass V, van der Vleuthen CPM, Schatzer J, Jones R. Assessment of clinical competence. Lancet. 2001;357:945-949.

[25] Wilkinson TJ, Frampton CM. Comprehensive undergraduate medical assessment improve prediction of clinical competence. Med Educ. 2004;38:1111-1116.

[26] Zimmerling W, Brehm RG. Der Prüfungsprozess: Überdenkungsverfahren, Klageverfahren, vorläufiges Rechtsschutzverfahren. Köln, Berlin, München: Carl Heymanns Verlag. 2004;2. Aufl.:54-74.

Artikel online frei zugänglich unter http://www.egms.de/en/journals/zma/2006-23/zma000263.shtml
Bitte zitieren als: Schulze J, Drolshagen S. Format und Durchführung schriftlicher Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc44.
gky(020060703.1)