Auswirkungen angeleiteter Itemanalysebesprechungen mit Dozierenden auf die Qualität von Multiple Choice Prüfungen

Effect of structured feedback to teachers on the quality of multiple choice examinations

Roger Kropf[1]
René Krebs[2]
Anja Rogausch[2]
Christine Beyeler[2]

1 Universität Zürich, Medizinische Fakultät, Studiendekanat, Zürich, Schweiz

2 Universität Bern, Institut für Medizinische Lehre, Abteilung für Assessment und Evaluation, Bern, Schweiz

Abstract

Background: In order to allow a meaningful interpretation of multiple choice (MC) assessment data, MC examinations have to be highly valid on the one hand and represent a reliable measure on the other. The aim of this study was to evaluate the effect of structured feedback given to teachers by assessment experts on the validity and reliability of subsequent MC examinations.
Methods: This feedback was introduced in the 3rd year of undergraduate medical training at the University of Zurich in 2007 and applied to nearly all teachers. Various validity and reliability criteria (relevance of content, taxonomic level, psychometric characteristics) of all end-of-term examinations one year before and one year after this intervention were compared. Other factors such as objectivity and representativeness were kept constant.
Results: After the introduction of structured feedback the multiple choice questions revealed a trend toward higher relevance. Taxonomic levels remained unchanged. However, selectivity and reliability coefficients increased significantly and the number of items eliminated from examination scoring due to insufficient psychometric properties decreased.
Conclusion: Structured feedback by assessment experts to teachers is a valuable tool for quality improvement of MC examinations, in particular regarding reliability.

Keywords: educational measurement, faculty development, continuous quality management, MCQ, assessment

Zusammenfassung

Hintergrund: Damit Multiple Choice Prüfungen über die reinen Prüfungsergebnisse hinausgehende Schlussfolgerungen zulassen, müssen sie für die jeweiligen Interpretationsabsichten inhaltlich gültig sein und hinreichend zuverlässig messen. Die vorliegende Studie geht der Frage nach, ob von Prüfungsexperten mit Dozierenden geführte Itemanalysebesprechungen eine Steigerung der Validität und Reliabilität nachfolgender Prüfungen bewirken.
Methoden: Diese Itemanalysebesprechungen wurden im dritten Studienjahr Humanmedizin an der Universität Zürich 2007 erstmalig flächendeckend eingeführt. Um deren Einfluss auf spätere Prüfungen zu untersuchen, wurden die Semesterabschlussprüfungen vor und nach dieser Intervention hinsichtlich verschiedener Validitäts- und Reliabilitätskriterien (inhaltliche Relevanz, taxonomische Stufe, psychometrische Kennwerte) miteinander verglichen. Andere Bedingungen wie beispielsweise Objektivität und inhaltliche Repräsentativität wurden konstant gehalten.
Ergebnisse: Nach Einführung der Itemanalysebesprechungen wiesen die Prüfungsfragen einen Trend zu höherer Relevanz auf. Die taxonomi-

sche Einstufung blieb unverändert. Hingegen stiegen sowohl die Trennschärfen als auch die Reliabilitätskoeffizienten signifikant an und es mussten weniger Prüfungsfragen wegen ungünstiger psychometrischer Eigenschaften aus der Prüfungsbewertung eliminiert werden.
Schlussfolgerung: Von Prüfungsexperten angeleitete Itemanalysebesprechungen mit Dozierenden stellen ein wertvolles Instrument zur Qualitätsverbesserung von Multiple Choice Prüfungen insbesondere hinsichtlich der Reliabilität dar.

Schlüsselwörter: Messmethoden in der Lehre, Fakultätsentwicklung in der Medizin, kontinuierliches Qualitätsmanagement, MCQ, Assessment

Einleitung

Multiple Choice Prüfungen sind nur dann von Nutzen, wenn sie Interpretationen zulassen, die über die Angabe des Anteils richtig beantworteter Fragen hinausgehen [5]. Zumindest sollte aus den Prüfungsergebnissen auf den Wissensstand in den Themenbereichen geschlossen werden können, aus denen die Prüfungsfragen ja lediglich eine mehr oder weniger repräsentative Stichprobe darstellen. Auch sollten sie eine Prognose ermöglichen, inwiefern sich die Prüfungsabsolventen das erforderliche Wissen und Verständnis angeeignet haben, um die nächsten Lernschritte, den nächsten Studienabschnitt erfolgreich bewältigen zu können. Solche Schlussfolgerungen sind nur zulässig, wenn die Prüfungen für die jeweiligen Interpretationsabsichten inhaltlich gültig sind und hinreichend zuverlässig messen.
Damit Prüfungsfragen (im nachfolgenden Text wird der Einfachheit halber der in der Prüfungsliteratur übliche Ausdruck “Item“ verwendet) zur inhaltlichen Gültigkeit (Validität) beitragen [1], [4], [2], [7], [15], müssen sie

- relevant sein hinsichtlich der Anforderungen in der weiteren Ausbildung und letztlich im Beruf,
- auf den gewünschten kognitiven Stufen prüfen (Wissen, Verstehen, Wissensanwendung zur Lösung von Problemen),
- repräsentativ zusammengestellt sein.

Bezüglich der Messzuverlässigkeit (Reliabilität) ist bei Multiple Choice Prüfungen, die hoch standardisiert sind, überwacht durchgeführt und nach vordefiniertem Schlüssel automatisiert ausgewertet werden, fast nur die Messgenauigkeit des Prüfungsinstrumentes von Belang. Diese wird heute praktisch ausschliesslich durch den alpha-Koeffizienten von Cronbach (innere Konsistenz der Items) erfasst, der für Prüfungen mit einschneidender Konsequenz für die Kandidaten nicht unter 0.8 liegen sollte [1], [3]. Damit Items zu einer konsistenten zuverlässigen Differenzierung beitragen, müssen sie [1], [3], [4]

- eindeutig lösbar sein,
- klar und verständlich formuliert sein,
- eine angemessene Schwierigkeit aufweisen,
- keine ungewollten Lösungshinweise enthalten.

Unter messtechnischer Betrachtung tragen Items dann zur Reliabilität der Prüfung bei, wenn ihre Beantwortung (richtig/falsch) positiv mit dem in der Prüfung erzielten Punktwert korreliert, sie also einen positiven Trennschärfekoeffizienten aufweisen [1]. Erwünscht sind dabei Werte ≥ 0.2. Neben der Itemqualität wird die Reliabilität wesentlich von der Prüfungslänge beeinflusst, da sich mit steigender Anzahl Items unerwünschte und zufällige Einflüsse gegenseitig ausnivellieren.
Bei der Prüfungsauswertung wird neben den gängigen Testgütekriterien meist auch eine detaillierte statistische Auswertung jedes Items (Itemanalyse) vorgenommen [12]. Die durch die Itemanalyse erzeugten Daten stellen eine wertvolle Form der Rückmeldung an Dozierende dar [6], [13], und tragen über die verbesserte Konstruktion von Neufragen dazu bei, Validität und Reliabilität zukünftiger Prüfungen zu verbessern [16]. Die vorliegende Arbeit geht der Frage nach, welchen Einfluss von Prüfungsexperten mit Dozierenden geführte Itemanalysebesprechungen auf die Qualität (Relevanz, taxonomisches Niveau, psychometrische Kennwerte) nachfolgender Prüfungen haben.

Methoden

Hintergrund

Im Zuge der 2002 initiierten Gesamtreform des Humanmedizinstudiums an der Medizinischen Fakultät der Universität Zürich, wurde das Curriculum 2005/06 im klinischen Abschnitt des dritten Studienjahres von einem fächerorientierten Unterricht auf einen organzentrierten Unterricht umgestellt. Dessen Organisationseinheit bildet der Themenblock. Verbunden mit dieser Umstellung war auch eine Neuausrichtung der Prüfungen. Die traditionellen Multiple Choice (MC) Prüfungen der Fächer Pathophysiologie, Pharmakologie und Mikrobiologie am Ende des dritten Studienjahres wurden überführt in insgesamt vier MC Semesterprüfungen (nachfolgend Prüfung genannt). Deren Prüfungsinhalte orientierten sich jeweils an den im Semester unterrichteten Themenblöcken. Grösstenteils mussten die Items neu erstellt werden, da ehemalige Items nur in Einzelfällen wieder verwendet werden konnten und in vielen Fächern gar keine Items existierten. Zur Vorbereitung wurden, in enger Zusammenarbeit mit dem Institut für Medizinische Lehre (IML) der Universität Bern, Schulungen für die sachgerechte Item-Konstruktion angeboten. Dozierende nahmen dieses Angebot in der Regel dankbar an und empfanden es als hilfreich für die

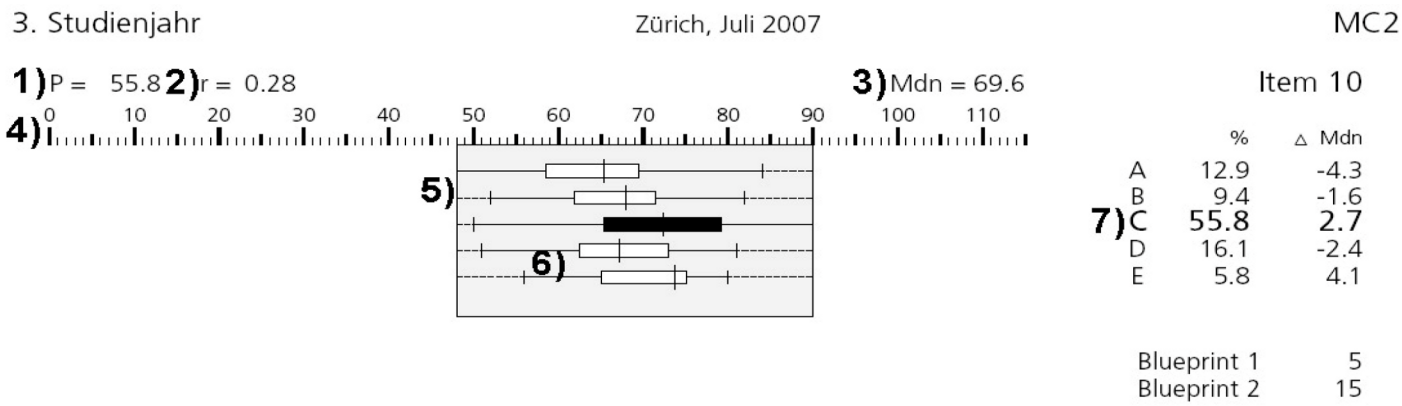

Abbildung 1: Beispiel einer Itemanalyse

Legende

1) Schwierigkeitskoeffizient (P): prozentualer Anteil der von der Kandidatengruppe in diesem Item erreichten zur maximal möglichen Punktzahl.

2) Trennschärfekoeffizient (r): Korrelation zwischen der erreichten Punktzahl in diesem Item und der Punktzahl in der Gesamtprüfung ohne dieses Item.

3) Medianwert (Mdn), den die gesamte Kandidatengruppe in der Prüfung ohne dieses Item erzielt hat.

4) Skala: zeigt die theoretisch in der Prüfung möglichen Punktzahlen

5) graues Feld: zeigt die Spannbreite der in der Prüfung (ohne dieses Item) tatsächlich erreichten Punktzahlen. Im Feld sind die Boxplots der verschiedenen Wählergruppen enthalten. Die richtige Antwort ist in der Box schwarz gefärbt. Die falschen Antworten sind weiss dargestellt.

6) Stabdiagramme: von Tukey vorgeschlagene prägnante Darstellung von Häufigkeitsverteilungen zur Analyse der Verteilungseigenschaften aller Wahlantworten [14]

7) Wahlantworten A-E: Prozentwert der Kandidaten, welche die betreffende Antwort gewählt haben, sowie Abweichung ihres Medianwerts vom Medianwert der Gesamtgruppe. Die richtige Antwort ist fett markiert.

8) Fragetext mit Wahlantworten

Erstellung eigener Items. Zur Gewährleistung inhaltlicher Repräsentativität und Konstanz erfolgte die Zusammenstellung der Prüfungen durchgehend auf Grundlage eines zweidimensionalen (Themenblöcke / Fächer) gewichteten Inhaltsrasters (sog. Blueprint). 2006 wurden die beiden Prüfungen (MC1 und MC2) des dritten Studienjahres erstmals mit einer Kohorte von 223 Kandidaten durchgeführt. Kandidatenkommentare zu den Items sowie allgemeine Rückmeldungen zur Prüfung wurden elektronisch erfasst und zusammen mit den Antwortbögen dem IML zur Auswertung zugeführt. Neben den gängigen Testgütekriterien wurde zusätzlich eine detaillierte statistische Auswertung jedes Items vorgenommen. Aus testpsychologischer Sicht auffällige Items [12] wurden identifiziert und nach Rücksprache mit den jeweiligen Themenblockverantwortlichen, wo angebracht, aus der Prüfungsauswertung entfernt. In Abbildung 1 ist exemplarisch eine Itemanalyse dargestellt [12].

Die Itemanalysen der Prüfungen 2006 wurden den sieben Themenblockverantwortlichen, vor dem Aufgebot Prüfungsfragen für die Prüfungen 2007 zu formulieren, schriftlich zur Verfügung gestellt. Drei Themenblockverantwortliche machten zudem Gebrauch vom zusätzlichen Angebot einer angeleiteten Besprechung der Itemanalyse mit einem Prüfungsexperten, die sie als überaus erkenntnisreich und wertvoll für die Überarbeitung des eigenen Unterrichtes und der zugehörigen Items empfanden. Deshalb wurde beschlossen, die aus den Prüfungsauswertungen von 2007 resultierenden Itemanalysen mit allen Themenblockverantwortlichen des dritten Studienjahres in Form angeleiteter Besprechungen zu diskutieren. Einzig mit einem Themenblockverantwortlichen konnte dies im vorgesehenen Zeitfenster nicht durchgeführt werden. Je nach Anzahl zu besprechender Items dauerten die Gespräche zwischen 45 und 90 Minuten. Die Besprechungsergebnisse wurden von den Themenblockverantwortlichen an die am Themenblock beteiligten Dozierenden weitergemeldet. 2008 wurden wiederum beide Prüfungen mit 220 (MC1) resp. 219 (MC2) Kandidaten durchgeführt und auf gleiche Art ausgewertet.

Um den Einfluss der angeleiteten Itemanalysebesprechungen von 2007 auf die Qualität nachfolgender Prüfungen zu untersuchen, wurden die Prüfungen von 2006 und 2008 anhand verschiedener Testgütekriterien miteinander verglichen.

Relevanzeinschätzung

Um mögliche Veränderungen in der Relevanz der Items zu prüfen, wurden aus jeder der vier Prüfungen je 15 Items nach dem Zufallsprinzip ausgewählt und 22 internistischen Oberärzten des Universitätsspitals Zürich (nachfolgend Rater genannt) unabhängig von einander vorgelegt. Die Anzahl von 60 Items erschien als adäquater Kompromiss zwischen der Verwendung einer hinreichend grossen Stichprobe einerseits und der zeitlichen Zumutbarkeit gegenüber den Ratern andererseits. Die Rater waren bezüglich Fragestellung und Itemherkunft verblindet. Um eine möglichst einheitliche Vorstellung des Relevanzbegriffs sicherzustellen, wurden die Rater gebeten, die Items jeweils unter Berücksichtigung folgender drei Teilaspekte einzuschätzen:

- Exemplarität: Wie gut prüft die Frage Wissen ab, das grundlegende medizinische Prinzipien erläutert und auf dem im weiteren Studium aufgebaut werden kann?
- Frequenz: Wie häufig kommt das in der Frage geprüfte Wissen im klinischen Alltag vor?
- Effektstärke: Wie hoch ist die Bedeutsamkeit des geprüften Wissens in Bezug auf eventuelle Nachteile / Schäden bei Patienten, die durch dessen Nichtwissen entstehen können?

Für die Einschätzung wurde eine 4-stufige Skala mit den Kategorien (1=relevant, 2=eher relevant, 3=eher irrelevant, 4=irrelevant und NB=nicht beurteilbar) verwendet.

Taxonomieinstufung

Zur Untersuchung der taxonomischen Stufen, wurde die gleiche Itemstichprobe von fünf Medizindidaktikern auf einer 3-stufigen Skala (1= Kennen, 2= Verstehen und 3= Anwenden und Beurteilen) eingeschätzt [7]. Auch hier waren die Rater bezüglich Fragestellung und Itemherkunft verblindet.

Psychometrische Kennwerte

Für alle vier Prüfungen wurden folgende psychometrische Kennwerte ermittelt und miteinander verglichen: Anzahl der Items, Mittelwert der Itemschwierigkeiten, Mittelwert der Itemtrennschärfen, Reliabilität und Standardmessfehler (auf 100 Items gerechnet), Mittelwert der Kandidatenscores mit Standardabweichung, Anzahl zur Elimination vorgeschlagene Items und effektiv eliminierte Items.

Statistische Auswertung

Die Prüfungsauswertung wurde durch das IML Bern mit einer eigens dafür entwickelten Testauswertungssoftware vorgenommen. Bei der Relevanzeinschätzung wurden die relativen Häufigkeiten der 4 Relevanzkategorien und der Median der Bewertungen pro Item ermittelt und Unterschiede in der zentralen Tendenz mittels Mann-Whitney-U Test verglichen. Bei der taxonomischen Einstufung wurde pro Item der am häufigsten gewählte Wert (Modalwert) ermittelt. Unterschiede bezüglich der Streuung der Kandidatenscores wurden mittels F-Test überprüft, Unterschiede bezüglich deren prozentualem Mittelwert mittels t-Test für unabhängige Stichproben. Mit dem gleichen Test wurden auch Unterschiede bezüglich der mittleren Schwierigkeitskoeffizienten und Trennschärfekoeffizienten der Items überprüft (Trennschärfen nach Fisher-Z-Transformation). Dies erschien angezeigt, da weniger als 15% aller Items der Prüfungen von 2006 zu Verankerungszwecken (Konstanthalten der Bestehensanforderungen) in den Prüfungen 2008 wieder verwendet wurden. Die statistische Auswertung erfolgte mit SPSS, Version 15.0.

Ergebnisse

Bei der Auswertung der Relevanzeinschätzung wurde ein Item ausgeschlossen, da es von 12 (55%) der Rater als nicht beurteilbar eingeschätzt wurde und somit für dieses Item nur wenige Bewertungen vorlagen. Tabelle 1 zeigt die relativen Häufigkeiten der Relevanzeinschätzungen der 59 Items von 2006 und 2008, die sich in der zentralen Tendenz signifikant unterscheiden (Mann-Whitney-U Test, p<0.01).

Tabelle 1: Vergleich der Relevanzeinschätzung (relative Häufigkeit der Einstufungen von 59 Items durch 22 Rater)

Semesterprüfungen	2006		2008	
Relevanzkategorie	Anzahl	Prozent	Anzahl	Prozent
1 (= relevant)	299	46.9	358	54.2
2 (= eher relevant)	189	29.6	194	29.4
3 (= eher irrelevant)	94	14.7	67	10.2
4 (= irrelevant)	46	7.2	31	4.7
NB (=nicht beurteilbar)	10	1.6	10	1.5

In Tabelle 2 sind die Mediane der Relevanzeinschätzungen dargestellt, wie sie sich aus den Bewertungen der 59 Items (29 Items von 2006; 30 Items von 2008) durch die 22 Rater ergeben. Der Anteil an Items, die im Median als „relevant“ (=1) eingeschätzt wurden, ist in den Prüfungen von 2008 höher als in denjenigen von 2006. Dieser Unterschied ist jedoch statistisch nicht signifikant.

Tabelle 2: Vergleich der Relevanzeinschätzung (Median der Einstufungen durch 22 Rater pro Item)

Semesterprüfungen	2006		2008	
Median	Anzahl	Prozent	Anzahl	Prozent
1.0 (= relevant)	12	41.4	16	53.3
1.5	3	10.3	4	13.3
2.0	12	41.4	7	23.3
2.5	0	0.0	0	0.0
3.0	1	3.4	3	10.0
3.5	1	3.4	0	0.0
4.0 (= irrelevant)	0	0.0	0	0.0

Die Ergebnisse der taxonomischen Einstufung (Modalwerte) der 60 Items durch die fünf Medizindidaktiker sind in Tabelle 3 dargestellt. Sowohl in den Prüfungen von 2006 als auch von 2008 wurden kumuliert 40% der Items den höheren Taxonomiestufen „Verstehen“ sowie „Anwenden und Beurteilen“ zugeordnet. Hier ist keine nennenswerte Veränderung erkennbar, weshalb auf eine statistische Überprüfung verzichtet wurde.

Tabelle 3: Vergleich der taxonomischen Einstufung (Modalwert der Einstufungen durch fünf Rater pro Item)

Semesterprüfungen	2006		2008	
Taxonomiestufen	Anzahl	Prozent	Anzahl	Prozent
Wissen	18	60.0	18	60.0
Verstehen	7	23.0	9	30.0
Anwenden und Beurteilen	5	17.0	3	10.0

Die für die vier Prüfungen 2006 und 2008 ermittelten psychometrischen Kennwerte finden sich in Tabelle 4. Bei der Prüfung MC2 wurden die Fragen 2008 nicht besser beantwortet als 2006. Bei der Prüfung MC1 resultierte 2008 ein tendenziell höherer mittlerer P-Wert (P-Differenz von 3.4%) und damit verbunden ein signifikant höherer prozentualer Mittelwert der Kandidatenscores (t=-3.64; p<.001). Die zur Konstanthaltung der Bestehensanforderung durchgeführte Verankerung nach dem Rasch-Modell zeigte aber auf, dass diese Differenz weitgehend durch eine etwas leistungsstärkere Kandidatenkohorte erklärt werden kann. Die Streubreite der Kandidatenscores (Standardabweichung in %) nahm in beiden Prüfungen bei gleichbleibendem Standardmessfehler signifikant zu um 1.81% resp. 1.61% (F=10.09; p=.002 resp. F=6.48; p=.01). Ebenfalls stiegen in beiden Prüfungen die mittleren Trennschärfen hoch signifikant an, in der MC1 von 0.17 auf 0.22 (t=-3.71; p<.001), in der MC2 von 0.15 auf 0.21 (t=-3.90; p<001). Damit verbunden stiegen auch die Reliabilitätskoeffizienten standardisiert auf eine Länge von 100 Items sehr deutlich von 0.79 auf 0.86 resp. von 0.77 auf 0.85 an. Parallel dazu verringerte sich die Anzahl eliminierter Items von 14 auf 10 (MC1) resp. von 16 auf 7 (MC2). Die Berechnungen mittels Spearman-Brown Formel [11] ergaben, dass die Prüfung MC1 von 2006 um 63 und die MC2 um 57 Items hätte verlängert werden müssen, um die Reliabilitäten der Prüfungen von 2008 zu erzielen. Dies entspricht einem Verlängerungsfaktor von 1.6.

Diskussion und Schlussfolgerung

Mit Bezug auf die Fragestellung, welchen Einfluss die angeleiteten Itemanalysebesprechungen mit einem Prüfungsexperten auf nachfolgende Prüfungen haben, sind folgende Beobachtungen von Bedeutung. Wie eingangs beschrieben, setzt sich die Validität aus verschiedenen Aspekten zusammen. Es konnte zum einen gezeigt werden, dass der Anteil als „relevant“ eingeschätzter Items in den Prüfungen 2008 höher war als in den Prüfungen 2006. Allerdings fiel die Veränderung der Relevanzeinschätzung weniger deutlich aus als erwartet. Dies könnte damit zusammenhängen, dass unter dem Begriff „Relevanz“ relativ heterogene Aspekte wie Exemplarität, Frequenz und Effektstärke zusammengefasst wurden. So könnte theoretisch ein Item zwar hoch-exemplarisch sein (d.h. auf grundlegende medizinische Prinzipien abzielen), zugleich aber eine weniger deutliche Effektstärke aufweisen (in dem Sinne, dass Nicht-Wissen Schäden an Patienten nach sich ziehen) und umgekehrt. Möglicherweise wären Relevanzunterschiede deutlicher zu Tage getreten, wenn die genannten Aspekte getrennt bewertet worden wären.

Die taxonomische Einstufung der Items als weiteres Kriterium der Validität ergab keinen nennenswerten Unterschied zwischen den Prüfungen 2006 und 2008. Der Anteil an Wissensfragen überwog mit 60% in beiden untersuchten Prüfungen. Dies ist als durchaus adäquat zu betrachten unter der Überlegung, dass im dritten Studienjahr die Vermittlung von Grundlagen der klinischen Medizin im Vordergrund steht. 40% der untersuchten Fragen wurden den höheren Taxonomiestufen „Verstehen“ sowie „Anwenden und Beurteilen“ zugeordnet. Dies belegt, dass

Tabelle 4: Psychometrische Kennwerte der vier Semesterprüfungen

Semesterprüfung	MC1 2006		MC1 2008		MC2 2006		MC2 2008	
Psychometrische Kennwerte	1. Run*	2. Run**	1. Run	2. Run	1. Run	2. Run	1. Run	2. Run
Anzahl Items	116	102	117	107	115	99	113	106
Anzahl wieder verwendete Items von 2006	-	-	16	15	-	-	13	13
Mittelwert Itemschwierigkeit (P)	58.75	63.60	63.57	66.98	64.17	68.97	66.96	69.36
Mittelwert Trennschärfe (r)	0.15	0.17	0.20	0.22	0.13	0.15	0.20	0.21
Reliabilität (Cronbach-alpha) auf 100 Items gerechnet	0.75	0.79	0.82	0.86	0.71	0.77	0.83	0.84
Standardmessfehler (s_e)	±4.0	±4.0	±3.9	±4.0	±3.9	±4.0	±3.9	±3.9
Anzahl zur Elimination vorgeschlagene Items	15	-	16	-	13	-	5	-
Anzahl eliminierte Items	-	14	-	10	-	16	-	7
Mittelwert der Kandidatenscores ($\bar{x}$)	68.42	65.13	72.58	71.91	73.99	68.47	75.20	73.77
Mittelwert der Kandidatenscores in % ($\bar{x}_{\%}$)	58.98	63.85	62.03	67.21	64.34	69.16	66.55	69.59
Standardabweichung der Kandidatenscores (s)	9.14	9.00	10.89	11.37	8.31	8.14	10.66	10.42
Standardabweichung der Kandidatenscores in % ($s_{\%}$)	7.88	8.82	9.31	10.63	7.23	8.22	9.43	9.83

*1. Run = vorläufige Erstauswertung; **2. Run = definitive Endauswertung

es sich bei den untersuchten Prüfungen aber nicht um reine Wissensprüfungen handelte.

Eine zuverlässige Leistungsdifferenzierung ist dann möglich, wenn die Kandidatenscores bei hoher Reliabilität der Prüfung breit streuen. Unter diesem Gesichtspunkt sind die von 2006 zu 2008 erfolgten Zunahmen der Standardabweichungen, Trennschärfen und Reliabilitätskoeffizienten allesamt erwünschte Effekte. Sowohl bei der Trennschärfe als auch bei der Reliabilität wurde 2008 der international geltende Zielwertbereich von ≥ 0.2 (Trennschärfe) und ≥ 0.8 (Reliabilität) erreicht. Dass zur Erzielung der beobachteten Reliabilitätsanstiege eine 1.6-fache Prüfungsverlängerung erforderlich gewesen wäre, verdeutlicht deren praktische, ökonomische Bedeutung. Ebenfalls sehr positiv zu werten ist der Rückgang der Zahl eliminierter Items, der vor allem in der MC2 mit neun Items sehr deutlich ausfiel, bedeutet doch der Ausschluss jedes Items potentiell einen Verlust hinsichtlich inhaltlicher Validität der Prüfung [11].

Zusammengefasst wurde eine deutliche positive Veränderung der Reliabilität festgestellt, sowie ein tendenzieller Anstieg bei einem Teilaspekt der inhaltlichen Validität. Diese Befunde sollten bezüglich ihrer Nachhaltigkeit weiter untersucht werden.

Die vorliegende Untersuchung wurde als offene Studie ohne Kontrollgruppe angelegt, weil die Dozierenden mit grosser Wahrscheinlichkeit von den angeleiteten Itemanalysebesprechungen erfahren hätten, da sie meist in mehreren Themenblöcken unterrichteten. Da sich die Rahmenbedingungen des Curriculums und der Prüfungen im beobachteten Zeitraum nicht geändert haben, liegt nahe, dass die festgestellten Validitäts- und Reliabilitätsverbesserungen wesentlich auf die 2007 flächendeckend durchgeführten Itemanalysebesprechungen zurückzuführen sind. Dass die Rückmeldungen an Dozierende vorzugsweise in Form angeleiteter Itemanalysebesprechungen mit Prüfungsexperten geschehen sollten, wird durch die Feststellung untermauert, dass statistische Prüfungsdaten für Dozierende oftmals schwierig zu verstehen und richtig zu interpretieren sind [8].

Im Einklang mit den Ergebnissen anderer Studien zur Qualitätsverbesserung von MC Prüfungen [9], [13], [16] lässt sich festhalten: Von Prüfungsexperten angeleitete Itemanalysebesprechungen mit Dozierenden stellen ein wertvolles Instrument zur Qualitätssteigerung von Prüfungen dar. Auch liefern sie wertvolle Hinweise zum Erfolg des Unterrichtes und helfen Dozierenden und Curriculumsentwicklern bei der Evaluation und Optimierung des Curriculums, indem sie die kontinuierliche Überprüfung von Lern- und Prüfungszielen in Bezug auf deren Angemessenheit und Realisierbarkeit gewährleisten. Sie sind trotz des zu Beginn notwendigen Aufwandes lohnenswert, da sie Einsparungen bei der Anzahl zu entwickelnder Items bringen und dadurch in der Folge die Dozierenden wieder entlasten.

Danksagung

Die Autoren danken den Oberärztinnen und Oberärzten der Klinik und Poliklinik für Innere Medizin am Universitätsspital Zürich sowie den fünf Medizindidaktikern in Bern und Zürich für ihre Beiträge zum Gelingen dieses Artikels.

Literatur

1. Bortz J, Döring N. Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler. 3. Auflage. Berlin: Springer-Verlag; 2002
2. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ. 2004;38(3):327-333. DOI:10.1046/j.1365-2923.2004.01777.x
3. Downing SM. Reliability: on the reproducibility of assessment data. Med Educ. 2004;38(9):1006-1012. DOI:10.1111/j.1365-2929.2004.01932.x
4. Downing SM. Threats to the validity of clinical teaching assessments: what about rater error? Med Educ. 2005;39(4):353-355. DOI:10.1111/j.1365-2929.2005.02138.x
5. Downing SM. Validity: On the meaningful interpretation of assessment data. Med Educ. 2003;37(9):830-837. DOI:10.1046/j.1365-2923.2003.01594.x
6. Georg W, Schubert S, Scheffner D, Burger W. Fünf Jahre Prüfungen im Reformstudiengang Medizin an der Charité - Universitätsmedizin Berlin. GMS Z Med Ausbild 2006;23(3):Doc48. Zugänglich unter: http://www.egms.de/static/en/journals/zma/2006-23/zma000267.shtml

7. Guilbert JJ. Educational handbook for health personnel. WHO Offset Publication No. 35. 6. Edition. Geneva: World Health Organization; 1998

8. Joseph MR. Practices, issues, and trends in student test score reporting. In: Downing SM, Haladyna TM, editors. Handbook of test development. 1. Edition, New York: Routledge; 2006. 677-710

9. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med. 2002;77(2):156-161. DOI:10.1097/00001888-200202000-00016

10. Krebs R. Anleitung zur Erstellung von MC-Fragen und MC-Prüfungen für die Ärztliche Ausbildung. Bern: IML/AAE; 2004. Zugänglich unter: http://www.fnl.ch/LOBs/LOs_Public/MC_Anleitung.pdf

11. Lienert GA, Raatz U. Testaufbau und Testanalyse. 5. Auflage. Weinheim: Beltz; 1994

12. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. GMS Z Med Ausbild. 2006;23(3):Doc53. Zugänglich unter: http://www.egms.de/static/en/journals/zma/2006-23/zma000272.shtml

13. Rotthoff T, Soboll S. Qualitätsverbesserung von MC Fragen: Ein exemplarischer Weg für eine medizinische Fakultät. GMS Z Med Ausbild. 2006;23(3):Doc45. Zugänglich unter: http://www.egms.de/static/en/journals/zma/2006-23/zma000264.shtml

14. Tukey JW. Exploratory Data Analysis. Reading, MA: Addison-Wesley, 1977

15. Ware J, Vik T. Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. Med Teach. 2009;31(3):238-243. DOI:10.1080/01421590802155597

16. Weih M, Harms D, Rauch C, Segarra L, Reulbach U, Degirmenci U, de Zwaan M, Schwab S, Kornhuber J. Qualitätsverbesserung von Multiple-Choice-Prüfungen in Psychiatrie, Psychosomatik, Psychotherapie und Neurologie. Nervenarzt. 2009;80(3):324-328. DOI:10.1007/s00115-008-2618-8

Korrespondenzadresse:
Dr. med. Roger Kropf
Universität Zürich, Medizinische Fakultät, Studiendekanat, Pestalozzistraße 3/5, CH-8091 Zürich, Schweiz, Tel.: +41 44 634 1099, Fax: +41 44 634 1088
roger.kropf@dekmed.uzh.ch

Bitte zitieren als
Kropf R, Krebs R, Rogausch A, Beyeler C. Auswirkungen angeleiteter Itemanalysebesprechungen mit Dozierenden auf die Qualität von Multiple Choice Prüfungen. GMS Z Med Ausbild. 2010;27(3):Doc46. DOI: 10.3205/zma000683, URN: urn:nbn:de:0183-zma0006834

Artikel online frei zugänglich unter
http://www.egms.de/en/journals/zma/2010-27/zma000683.shtml

Eingereicht: *17.07.2009*
Überarbeitet: *10.12.2009*
Angenommen: *14.12.2009*
Veröffentlicht: *17.05.2010*