

Vergleich kollegialer Einzel- mit Gruppen-Reviews allgemeinmedizinischer Multiple-Choice-Fragen

Zusammenfassung

Zielsetzung: Im Fach Allgemeinmedizin werden die obligat zu benotenden Leistungsnachweise an vielen Hochschulstandorten über Klausuren mit Multiple-Choice-Fragen (MCF) erbracht. Zur Qualitätssicherung bietet sich u.a. ein Peer-Review-Verfahren der eingesetzten MCF an. Für die optimale Effektivität und Effizienz solcher zeit- und personalintensiven Peer-Reviews ist nicht zuletzt die Verfahrensweise von Bedeutung. Ziel der Studie war es zu untersuchen, ob sich Einzel- von Gruppen-Reviews hinsichtlich definierter Parameter unterscheiden.

Methodik: In einer kontrollierten Studie mit cross-over Design, durchgeführt mit je drei allgemeinmedizinischen Reviewern vier verschiedener deutscher Hochschulstandorte, wurden die Beurteilungen der Reviewer von 80 MCF (für jeden Teilnehmer jeweils 40 im Einzel-, 40 im Gruppen-Review) mit externen Beurteilungen durch ein Expertengremium und untereinander verglichen. Daneben wurden über Fragebögen subjektive Einschätzungen der Studienteilnehmer zum Review und der Zeitaufwand erfasst.

Ergebnisse: Statistisch signifikante Unterschiede in der Validität und Reliabilität fanden sich zwischen Einzel- und Gruppen-Review nicht. Der Zeitaufwand für die Gruppen-Reviews lag im Mittel etwas höher als für die Einzel-Reviews. Die subjektiven Einschätzungen der Studienteilnehmer zur Zufriedenheit mit dem Review-Prozess, der Effektivität und Wichtigkeit der Reviews lassen auf eine Präferenz für den Gruppen-Review schließen.

Schlussfolgerungen: Eindeutige Empfehlungen für oder gegen die Durchführung eines der beiden Review-Verfahren lassen sich aufgrund der Studienergebnisse nicht abgeben. Die spezifische Arbeitsstruktur und -organisation sowie die Präferenzen der Mitarbeiter an den einzelnen Hochschulstandorten sollten bei der Wahl des Verfahrens berücksichtigt werden.

Schlüsselwörter: Medizinische Ausbildung, Prüfung, Multiple-Choice-Fragen, Review

Klaus Böhme¹

Jörg Schelling²

Irmgard

Streitlein-Böhme³

Katharina Glassen⁴

Jeannine Schübel⁵

Jana Jünger⁶

1 Universitätsklinik Freiburg,
Lehrbereich
Allgemeinmedizin, Freiburg,
Deutschland

2 LMU München, Lehrbereich
Allgemeinmedizin, München,
Deutschland

3 Universität Freiburg,
Medizinische Fakultät,
Studiendekanat, Freiburg,
Deutschland

4 Universitätsklinikum
Heidelberg, Abteilung
Allgemeinmedizin und
Versorgungsforschung,
Heidelberg, Deutschland

5 Uniklinikum Dresden, Carus
Hausarztpraxis, Dresden,
Deutschland

6 Universität Heidelberg,
Kompetenzzentrum für
Prüfungen in der Medizin,
Heidelberg, Deutschland

Einleitung und Fragestellung

Die Approbationsordnung von 2002 [1] brachte für jedes Fach im klinischen Studienabschnitt Humanmedizin die Notwendigkeit mit sich, benotete Leistungsnachweise für die Studierenden einzuführen. Aus Gründen der Praktikabilität geschieht dies vielfach in Form von schriftlichen Prüfungen mit Multiple-Choice-Fragen (MCF), die sich durch eine zufriedenstellende Reliabilität und Objektivität auszeichnen [2]. In der Literatur finden sich Übersichten, die Regeln für die Erstellung „guter“ MCF sowohl auf formaler wie auch auf inhaltlicher Ebene beschreiben [3],

[4], [5]. Bis heute sind an deutschen medizinischen Fakultäten viele Autoren von MC-Fragen nicht in der Anwendung dieser Regeln geschult, auch ein standardisierter Review-Prozess für die zum Einsatz kommenden Fragen existiert vielfach nicht [6]. Dementsprechend ist die Qualität der sich im Einsatz befindlichen MCF zumindest nicht gesichert.

Zur Gewährleistung eines angemessenen Niveaus der eingesetzten MCF bietet sich im Vorfeld neben Prüferschulungen ein standardisiertes Peer-Review-Verfahren an. Das Spektrum denkbarer Review-Verfahren reicht von einem Einzel-Review zu beurteilender Fragen über mehrere Einzel-Reviews bis hin zu moderierten oder nicht

moderierten Gruppen-Reviews, „face-to-face“ oder virtuell [7].

Optimale Effektivität und Effizienz der recht zeit- und personalintensiven Peer-Reviews hängen von verschiedenen Faktoren ab. Die Frage der Validität und Reliabilität der Beurteilungen spielt dabei eine zentrale Rolle. Will man die Motivation von Reviewern stärken, so sollten diese mit dem Review-Prozess zufrieden sein, ferner sollten sie von der Effektivität wie auch von der Wichtigkeit desselben überzeugt sein. Einen weiteren bedeutenden Faktor stellt der Zeitaufwand dar.

Seit November 2008 greift der Lehrbereich Allgemeinmedizin der Universität Freiburg bei der Erstellung von MC-Klausuren auf ein web-basiertes elektronisches Prüfungssystem, entwickelt im „Kompetenzzentrum für Prüfungen in der Medizin Baden-Württemberg“ der Universität Heidelberg, zurück. Der Fragen-Pool des Lehrbereiches Allgemeinmedizin ist in diesem „Item-Management-System“ (IMS) hinterlegt und dient als Grundlage für die Erstellung der Freiburger Klausuren.

Bundesweit greifen mittlerweile 15 Fakultäten auf das IMS als elektronische Hilfe bei der Erstellung und Auswertung von Klausuren zurück. Vereinfacht das System die organisatorischen Abläufe schon deutlich, liegt ein wohl noch bedeutenderer Mehrwert darin, theoretisch auf die Prüfungsfragen anderer Fakultäten zugreifen und für die eigenen Klausuren verwenden zu können. Ein solcher Zugriff setzt einerseits die Bereitschaft der einzelnen Standorte voraus, anderen Fakultäten ihre Fragen zur Verfügung zu stellen. Andererseits war mit allen Nutzern des Systems konsentiert, dass nur Fragen in einen „öffentlichen“, also anderen Nutzern zugänglichen Pool gestellt werden können, die einen definierten Review-Prozess durchlaufen haben.

Im Rahmen der hier vorgestellten Studie sollten anhand folgender konkreter Fragestellungen Einzel-Reviews mit nicht moderierten „face-to-face“-Gruppen-Reviews verglichen werden:

- Gibt es Unterschiede von Einzel- und Gruppenreviews bei der (a) Häufigkeit festgestellter Fehler (Übereinstimmung der Reviewformen: „Reliabilität“) und (b) im Vergleich zu einem Standard-Review ausgewiesener Experten (Übereinstimmung mit Standard: „Validität“)?
- Beeinflusst das Review-Verfahren die Zufriedenheit der Reviewer mit dem Prozess, ihre Einschätzung der Effektivität und der Wichtigkeit des durchgeführten Reviews?
- Unterscheidet sich der Zeitaufwand für die Durchführung von Einzel- bzw. Gruppen-Reviews?

Methoden

Stichprobe

Für die Studie ausgewählt wurden vier allgemeinmedizinische Abteilungen deutscher Hochschulen, die ihre Bereitschaft zur Studienteilnahme erklärten: Dresden, Frei-

burg, Heidelberg und LMU München, im Folgenden aus Gründen der Anonymisierung in willkürlicher Reihung als Uni 1-4 bezeichnet. Jeder Standort stellte drei Mitarbeiter, die sich abteilungsintern mit dem Erstellen sowie dem Review von MCF befassen.

Item-Stichprobe

Aus dem Fragenpool des Lehrbereiches Allgemeinmedizin der Universität Freiburg wurden für die Studie zufällig 2 x 40 MCF (Gruppe A und B) ausgewählt. Es handelte sich hierbei ausnahmslos um sog. Typ A-Fragen (positive oder negative Einfachauswahl aus fünf Wahlantworten).

Materialien und technische Voraussetzungen

Es ist eine der Funktionalitäten des IMS, alle erfassten Prüfungsfragen online mittels eines zehnkriterien umfassenden Bewertungsbogens (siehe Tabelle 1) einem Review unterziehen zu können¹. Ein Eingabefeld für Freitextkommentare ermöglicht es, Kritikpunkte zu konkretisieren und gezielte Korrekturvorschläge zu unterbreiten.

Beurteilungsgrundlage für sämtliche Reviews der Studie stellte eine „Kurz-Anleitung zum Review von MC-Fragen“ des Kompetenzzentrums dar, die wiederum auf einschlägiger Literatur zu dieser Thematik beruht [3], [4], [5].

Für die Erfassung der Zufriedenheit mit dem Review-Prozess, der Effektivität des Prozesses sowie der subjektiven Einschätzung der Wichtigkeit des Reviews wurde ein Kurzfragebogen erstellt. Die Antworten waren anzugeben auf einer 6-stufigen Likert-Skala (siehe Abbildung 1). Ferner wurden die offenen Fragen „Was hat mir bei dem Review am besten gefallen?“ und „Womit hatte ich beim Review am meisten Probleme?“ gestellt.

Durchführung

In einem ersten Schritt wurde über ein Experten-Review aller 80 MCF ein Vergleichsstandard für die Beurteilungen der Studienteilnehmer geschaffen. Das vierköpfige Gremium mit entsprechender Expertise (MME, bzw. Mitarbeiter des Kompetenzzentrums), besetzt mit drei Fachvertretern und einem fachfremden Kollegen, unterzog im „Kompetenzzentrum für Prüfungen in der Medizin Baden-Württemberg“ alle 80 MCF, die in der Studie zur Anwendung kommen sollten, in einer Sitzung einem Gruppen-Review.

Allen Studienteilnehmern wurde die „Kurz-Anleitung zum Review von MC-Fragen“ zur Verfügung gestellt, in der die Kriterien der Checkliste für den Review (siehe Tabelle 1) erläutert wurden. Eine darüber hinausgehende Schulung der Reviewer fand nicht statt. Entsprechend dem Studiendesign (siehe Tabelle 2) waren dann an jedem Standort von jedem der drei Reviewer 40 MCF im Einzel- und 40 MCF im Gruppen-Review zu beurteilen.

Für die Gruppen-Reviews vereinbarten die Studienteilnehmer Termine, an denen alle 40 MCF in einer Sitzung beurteilt wurden. Bei den Einzel-Reviews hatten die Review-

Tabelle 1: Checkliste für den Review von MC-Fragen

		Keine Mängel	Geringe Mängel	Gravierende Mängel
	Formal			
1.	Der Fragentyp ist zulässig/geeignet			
2.	Die Frage enthält keine versteckten Antworthinweise (Cueing)			
3.	Die Frage ist auch ohne Kenntnis der Antwortalternativen verständlich und beantwortbar (z.B. keine doppelten Verneinungen, Verhältnis Fragenstamm zu Antwortoptionen)			
4.	Die Antwortalternativen sind homogen (alle Antwortalternativen sind ungefähr gleich lang, in der Regel max. zwei Zeilen; Antwortalternativen, die nur aus Zahlen oder einem Begriff bestehen, sind aufsteigend sortiert)			
	Inhaltlich			
5.	Die Frage passt zur Klassierung (z.B. Fach)			
6.	Die Frage ist inhaltlich korrekt			
7.	Die Frage ist dem Ausbildungsniveau angemessen			
8.	Die Frage hat einen Anwendungsbezug/eine Patientenvignette			
9.	Fragenstamm und Frage sind verständlich formuliert und plausibel			
10.	Alle Antwortoptionen sind verständlich formuliert und plausibel			

Fragen zum Review-Prozess						
1. Einzel-Review:	1= stimme gar nicht zu			6= stimme voll und ganz zu		
	1	2	3	4	6	6
1.1 Mit dem Review-Prozess war ich zufrieden	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1.2 der Review-Prozess war effektiv	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1.3 ich halte den Review der Fragen für wichtig	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Zeitaufwand für alle 40 Fragen im **Einzel-Review**: Minuten

Fragen zum Review-Prozess						
2. Gruppen-Review:	1= stimme gar nicht zu			6= stimme voll und ganz zu		
	1	2	3	4	5	6
2.1 Mit dem Review-Prozess war ich zufrieden	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2.2 der Review-Prozess war effektiv	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2.3 ich halte den Review der Fragen für wichtig	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

Zeitaufwand für alle 40 Fragen im **Gruppen-Review**: Minuten

Abbildung 1: Kurzfragebogen zu den Reviews

er die Möglichkeit, sich die Zeit für die Beurteilungen frei einzuteilen und beliebig zu fraktionieren.

Jeweils im Anschluss an den Einzel- bzw. Gruppen-Review waren die Studienteilnehmer angehalten, den o.a. Kurzfragebogen auszufüllen.

Statistik

Über ein Experten-Review der für diese Studie zum Einsatz kommenden MCF wurde für jede Frage eine Referenz-Beurteilung, ein „Gold-Standard“ geschaffen. Mit dieser externen Referenz wurden die Beurteilungen innerhalb der Studie verglichen und so ihre Validität (Gültigkeit)

Tabelle 2: Ablaufschema zum Review allgemeinmedizinischer Prüfungsfragen (Gruppe A=MCF 1-40, Gruppe B=MCF 41-80)

	Uni 1	Uni 2	Uni 3	Uni 4
1. Durchgang	1 x Gruppenreview (Gruppe B)	3 x Einzelreview (Gruppe A)	3 x Einzelreview (Gruppe B)	1 x Gruppenreview (Gruppe A)
2. Durchgang	3 x Einzelreview (Gruppe A)	1 x Gruppenreview (Gruppe B)	1 x Gruppenreview (Gruppe A)	3 x Einzelreview (Gruppe B)

überprüft. Die Reliabilität (Zuverlässigkeit) wurde durch den Vergleich der festgestellten Mängel an den vier verschiedenen Standorten überprüft.

Die insgesamt 3200 Einzel-Beurteilungen der innerhalb der Studie durchgeführten Reviews (80 MCF x 10 Beurteilungskriterien x 4 Standorte) wurden dichotomisiert nach 0=keine Abweichung vom Experten-Review und 1=Abweichung vom Experten-Review, bzw. nach 0=kein Mangel und 1=geringer oder gravierender Mangel und so einer statistischen Analyse unterzogen. Bei den Einzel-Reviews wurde der gerundete Mittelwert der drei einzelnen Beurteilungen zum Vergleich herangezogen.

Die Variablen „Zahl der Abweichungen vom Vergleichsstandard“ und „Anzahl der gefundenen Fehler“ wurden varianzanalytisch (lineares gemischtes Modell, „Linear Mixed Model“ [8]) mit den festen Faktoren („fixed factor“) „Review-Form“ (Einzel/Gruppe), „Standort“ (Uni 1- Uni 4), „Fragengruppe“ (MCF-Gruppe A / MCF-Gruppe B) und „Durchgang“ sowie dem Zufallsfaktor („random factor“) „MCF“ (d. h. es wird angenommen, dass die MCF eine Zufallsauswahl aus dem MCF-Pool darstellen) ausgewertet. Von primärem Interesse waren dabei die Faktoren „Review-Form“ und „Standort“, die anderen Faktoren dienten als Kontrollvariablen.

Die Variablen für die Einzelkategorien sind binär (Wertebereich 0 und 1) und wurden daher mit einem nichtlinearen gemischten Modell mit logistischer Linkfunktion analog analysiert (logistisch-normales Modell, [9], [10]). Die statistische Auswertung der die Studie begleitenden Fragebögen erfolgte deskriptiv.

Ergebnisse

Zusammensetzung der Stichprobe

An der Studie nahmen zwölf Allgemeinmediziner von vier deutschen Hochschulstandorten teil, acht Frauen und vier Männer. Das Alter der Teilnehmer reichte von 24 bis 61 Jahren, im Mittel betrug es 41 Jahre (SD=14,6).

Validität

Die Ergebnisse der Abweichungen der Studien-Reviews vom Experten-Review sind in Tabelle 3 dargestellt. Statistisch signifikante Unterschiede fanden sich weder zwischen der Art des Reviews (Einzel- vs. Gruppenreview), noch zwischen Durchgang 1 und Durchgang 2, noch

zwischen Fragengruppe A und Fragengruppe B. Ein signifikanter Unterschied fand sich lediglich zwischen den einzelnen Reviewer-Gruppen der verschiedenen Standorte.

Reliabilität

Die Anzahl der im Rahmen der Studien-Reviews festgestellten Mängel sind in Tabelle 4 dargestellt. Bei der Gesamtbetrachtung über alle 3200 Kriterien fanden sich auch hier keine signifikanten Unterschiede zwischen der Art des Reviews (Einzel- vs. Gruppen-Review), dem Durchgang sowie der Fragengruppe. Ein solcher fand sich auch hier nur zwischen den Standorten.

Fragebögen

Die Frage nach der Zufriedenheit mit dem Review-Prozess beantworteten die Studienteilnehmer nach dem Einzel-Review im Mittel mit 4,92 (SD=0,69), nach dem Gruppen-Review mit 5,17 (SD=0,83). Die Frage nach der Effektivität wurde wie folgt beantwortet: nach dem Einzel-Review im Mittel mit 4,92 (SD=0,69), nach dem Gruppen-Review mit 5,58 (SD=0,67). Die Wichtigkeit des Reviews wurde nach dem Einzel-Review im Mittel mit 5,75 (SD=0,45) bewertet, nach dem Gruppen-Review mit 6,00 (SD=0). Die Freitextkommentare zu den Fragen „Was hat mir bei dem Review am besten gefallen?“ und „Womit hatte ich beim Review am meisten Probleme“ lassen sich wie folgt zusammenfassen:

Aus Sicht der Reviewer war der positivste Aspekt beim Einzel-Review die freie Zeiteinteilung, beim Gruppen-Review wurden der kollegiale Gedankenaustausch und der damit verbundene Lerneffekt hervorgehoben. Beim Einzel-Review wurden häufiger Probleme mit unklaren Bewertungskriterien bei fehlenden Rücksprachemöglichkeiten beklagt, im Gruppen-Review die zeitlichen Abstimmungsprobleme, der häufig nicht leicht herzustellende Konsens bei den Bewertungen sowie die mit dieser Review-Art verbundene lange Sitzungsdauer mit entsprechenden Konzentrationsproblemen.

Zeitaufwand

Die durchschnittliche Bearbeitungszeit für die 40 Fragen im Einzel-Review betrug 113 Minuten (SD=44), im Gruppen-Review 139 Minuten (SD=48).

Tabelle 3: Anzahl der Abweichungen vom Experten-Review in den Reviews der 4 Universitätsstandorte im Einzel- (ER) bzw. Gruppenreview (GR) in allen Einzel-Kriterien der Review-Checkliste einschließlich Signifikanzniveaus

keine Abweichung	Uni 1		Uni 2		Uni 3		Uni 4		Total
	ER	GR	ER	GR	ER	GR	ER	GR	
Abweichungen (in Prozent)	330 (17,50%)	347 (13,25%)	322 (19,50%)	321 (19,75%)	339 (15,25%)	334 (16,50%)	279 (30,25%)	262 (34,50%)	2534 (20,81%)
Total	400	400	400	400	400	400	400	400	3200

Unterschiede: Uni – Review-Art - $F(79,3) = 45.82; p < 0.0001$
 $F(79,1) = 0.10; p = 0.7499$

Tabelle 4: Anzahl der mit Mängeln bewerteten Einzel-Kriterien der 4 Universitätsstandorte im Einzel- (ER) und Gruppenreview (GR) einschließlich Signifikanzniveaus

keine Mängel (in Prozent)	Uni 1		Uni 2		Uni 3		Uni 4		Total
	ER	GR	ER	GR	ER	GR	ER	GR	
Mängel (in Prozent)	356 (11,00%)	353 (11,75%)	364 (9,00%)	352 (12,00%)	328 (18,00%)	349 (12,75%)	282 (29,50%)	266 (33,50%)	2650 (17,18%)
Total	400	400	400	400	400	400	400	400	3200

Unterschiede: Uni – Review-Art - $F(79,3) = 67.59; p < 0.0001$
 $F(79,1) = 0.28; p = 0.6006$

Diskussion

Dass der systematische kollegiale Review von MCQs die Qualität der Prüfungsfragen verbessert, konnte für den deutschen Sprachraum bereits hinreichend dargestellt werden [11], [12], [13]. Was die Vor- und Nachteile verschiedener Review-Verfahrens angeht, finden sich wenige Hinweise [7], die hier vorgestellte Studie sollte helfen, die für die Praxis relevante Frage nach den Vor- und Nachteilen von Einzel- im Vergleich zu Gruppen-Reviews zu klären:

Validität

Die Ergebnisse dieser Studie zeigen, dass beide Review-Verfahren in gleicher Weise valide sind. Es ergibt sich somit keine Entscheidungsgrundlage, eine der beiden Varianten vorrangig zu empfehlen.

Statistisch signifikante Unterschiede in den Beurteilungen der MCF fanden sich allerdings zwischen den Reviewern der vier einzelnen Standorte. Die „Kurz-Anleitung zum Review von MC-Fragen“ allein, obwohl auf den gleichen Grundlagen wie das Wissen der „Experten“ beruhend, war offensichtlich nicht geeignet, über die vier Universitätsstandorte hinweg zu homogenen Beurteilungen zu kommen. Bei den Abweichungen vom Experten-Review gab es über alle Checklisten-Items hinweg betrachtet einen signifikanten Unterschied. Das Ergebnis legt den Schluss nahe, neben andernorts geforderten Autorenschulungen [4] bei breitflächiger Einführung eines standardisierten Review-Verfahrens eine Schulung der Reviewer zu empfehlen [13].

Reliabilität

Auch die an den verschiedenen Standorten festgestellten Mängel waren über alle Checklisten-Items betrachtet nicht davon abhängig, ob die Fragen im Einzel- oder im

Gruppen-Review beurteilt worden waren. Somit ergibt sich auch hier keine Entscheidungsgrundlage für eine der beiden Varianten.

Anders verhielt es sich wieder beim Vergleich der Hochschulstandorte untereinander: hier fanden sich statistisch signifikante Unterschiede bei allen Items, die einer statistischen Auswertung zugeführt werden konnten. Die Forderung nach einer Reviewer-Schulung lässt sich somit bekräftigen.

Fragebögen

Möchte man sicherstellen, dass sich bundesweit möglichst viele Lehrende an einem wechselseitigen Review von MCF nachhaltig beteiligen, sollte sichergestellt sein, dass sie von der Notwendigkeit eines solchen Vorgehens überzeugt sind und die Verfahrensweise aus ihrer Sicht effektiv und subjektiv zufriedenstellend ist. Aus diesem Grund wurden die einzelnen Reviews von Fragebögen begleitet, die diese Parameter abfragten.

Über alle vier Standorte gemittelt erhielten sowohl die Fragen nach der Zufriedenheit, der Effektivität des Review-Prozesses, als auch nach der Einschätzung der Wichtigkeit des Reviews eine geringfügig höhere Zustimmung nach dem Gruppen-Review als nach dem Einzel-Review. Betrachtet man die Fragen für die einzelnen Standorte getrennt, findet sich die größere Zustimmung nach dem Gruppen-Review nahezu durchgehend wieder. Lediglich an einem Standort (Uni 1) fand die Frage nach der Zufriedenheit mit dem Review-Prozess mehr Zustimmung nach dem Einzel-Review. Auch hier wären gruppendynamische Prozesse zu diskutieren: der Standort Uni 1 hatte mit 210 Minuten Bearbeitungszeit im Gruppen-Review die längste Zeit von allen Review-Sitzungen benötigt. Dies könnte auf Probleme schließen lassen, zu gemeinsam verantworteten Entscheidungen zu finden.

Die Freitextkommentare zu den Einzel- und Gruppen-Reviews spiegeln die zu erwartenden Vor- und Nachteile der

beiden Verfahren wieder: die freie Zeiteinteilung steht beim Einzel-Review den fehlenden Rückkopplungsmöglichkeiten mit Kollegen gegenüber. Beim Gruppen-Review erkennt man, dass die zeitlichen Abstimmungsprobleme mit dem häufig angeführten und als positiv erachteten kollegialen Gedankenaustausch und dem damit verbundenen Lerneffekt im Widerspruch stehen.

Zeitaufwand

Betrachtet man lediglich den Fragenpool des Lehrbereiches Allgemeinmedizin in Freiburg, der ca. 280 MCF umfasst (andere Standorte mögen über deutlich umfangreichere Pools verfügen), so lässt sich der enorme Zeitaufwand erahnen, der für den Review des Bestandes aller allgemeinmedizinischen Abteilungen und Lehrbereiche erforderlich ist. Insofern sind die im Rahmen der Studie ermittelten Zeiten bei der Diskussion um das praxistauglichere Review-Verfahren zu berücksichtigen: für die Gruppen-Reviews von 40 MCF wurden im Durchschnitt 26 Minuten mehr benötigt als für die Einzel-Reviews. Folgte man allein diesem Kriterium, müsste eine Empfehlung in Richtung Einzel-Reviews ausgesprochen werden. Interessanterweise wird dieser etwas höhere Zeitaufwand in den Freitextkommentaren nicht thematisiert.

Schwächen

Die relativ kleine Stichprobe der Reviewer, die zudem aus Gründen der Machbarkeit keine Zufallsstichprobe darstellte, schränkt die Verallgemeinerbarkeit der Studienergebnisse ein. Da bewusst die Mitarbeiter in den allgemeinmedizinischen Abteilungen und Lehrbereichen ausgewählt worden waren, die sich auch in der Alltagsroutine mit dem Erstellen und/oder dem Review von Prüfungsfragen befassen ohne über eine spezifische Zusatzausbildung zu verfügen, wurden zwangsläufig Diskrepanzen bei medizinischer Erfahrung und testtheoretischem Vorwissen in Kauf genommen. Beides schränkt die Vergleichbarkeit ein.

Die zehn Items umfassende Checkliste, anhand derer die MCF zu beurteilen waren, bildet nicht alle in der Literatur [3], [4], [5] aufgeführten Kriterien zur Erstellung „guter“ MCF ab. Einige Items müssen als redundant angesehen werden, andere Bewertungskriterien, wie die wichtige Frage nach der Relevanz einer MCF [11], fehlen.

Schlussfolgerungen

Bei der Suche nach einer Entscheidungshilfe für die Empfehlung des besser geeigneten Review-Verfahrens können die Ergebnisse dieser Studie durchaus hilfreich sein.

Wie im Ergebnisteil dargestellt und in der Diskussion ausgeführt, lässt sich ein statistisch signifikanter Unterschied in der Validität und Reliabilität beider Vorgehensweisen nicht darstellen, hierüber allein ist eine Entscheidung für oder gegen eines der Verfahren nicht abzuleiten.

Vor diesem Hintergrund bekommen die subjektiven Einschätzungen der Studienteilnehmer zu den Reviews umso mehr Gewicht. Diese tendieren mehr in Richtung Gruppen-Review. Der Zeitfaktor wird dabei scheinbar von den Studienteilnehmern der Zufriedenheit sowie der subjektiven Einschätzung zur Effektivität des Prozesses untergeordnet.

Die spezifische Situation vor Ort kann bei der Auswahl des Verfahrens nachvollziehbar eine entscheidende Rolle spielen: sind die Reviewer wissenschaftliche oder klinisch tätige Mitarbeiter, die für gewöhnlich in einer Abteilung räumlich zusammenarbeiten, wird die Wahl eher auf relativ einfach zu terminierende Gruppen-Reviews fallen können. Werden die Reviews in der Regel von niedergelassenen Lehrbeauftragten durchgeführt (wie in der Allgemeinmedizin häufig üblich), mag der Einzel-Review vom eigenen Arbeitsplatz aus sinnvoller und praktikabler erscheinen.

Aus Gründen der Praktikabilität kamen in der Studie nur Typ A-Fragen zur Anwendung. Grundsätzlich sollten die Ergebnisse jedoch auch auf andere Fragenformate übertragbar sein.

Anmerkung

¹ Im IMS ist diesem Bewertungsbogen ein hier nicht näher zu definierender Algorithmus hinterlegt, der die Fragen entweder für den öffentlichen Ordner freigibt oder zur Korrektur an die Autoren verweist.

Danksagung

Für die statistische Beratung zu diesem Projekt gilt mein besonderer Dank Dr. Andreas Möltner, Kompetenzzentrum für Prüfungen in der Medizin - Baden Württemberg, Universität Heidelberg

Interessenkonflikt

Die Autoren erklären, dass sie keine Interessenkonflikte im Zusammenhang mit diesem Artikel haben.

Literatur

1. Bundesministerium für Gesundheit. Approbationsordnung für Ärzte vom 27.06.2002. BGBl. 2002:2405-2435.
2. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. *GMS Z Med Ausbild.* 2006;23(3):Doc53. Zugänglich unter/available from: <http://www.egms.de/static/de/journals/zma/2006-23/zma000272.shtml>
3. Haladyna TM, Downing SM, Rodrigues MC. A review of multiple-choice item-writing guidelines for a classroom assessment. *Appl Meas Educ.* 2002;15:309-344. DOI: 10.1207/S15324818AME1503_5

4. Krebs R. Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung. Bern: Institut für Medizinische Lehre IMS, Abteilung für Ausbildungs- und Examensforschung AAE; 2004.
5. AG Progress Test Medizin. Progress Test Medizin. Leitfaden für Fragenautorinnen und -autoren des Progress Test Medizin. Berlin: Charité Universitätsmedizin; 2003.
6. Jünger J, Möltner A, Lammerding-Köppel M, Rau T, Obertacke U, Biller S, Narciß E. Durchführung der universitären Prüfungen im klinischen Abschnitt des Medizinstudiums nach den Leitlinien des GMA-Ausschusses Prüfungen: Eine Bestandsaufnahme der medizinischen Fakultäten in Baden-Württemberg. *GMS Z Med Ausbild.* 2010;27(4):Doc57. DOI: 10.3205/zma000694
7. Kazubke E, Schüttpelz-Brauns K. Gruppenleistungen beim Review von Multiple-Choice-Fragen – ein Vergleich von face-to-face und virtuellen Gruppen, mit und ohne Moderation. *GMS Z Med Ausbild.* 2010;27(5):Doc68. DOI: 10.3205/zma000705
8. Brown H, Prescott R. *Applied Mixed Models in Medicine*, Second Edition. Oxford/UK: John Wiley & Sons, Ltd; 2006. DOI: 10.1002/0470023589
9. Beitler PJ, Landis JR. A Mixed-effects Model for Categorical Data. *Biometr.* 1985;41:991-1000. DOI: 10.2307/2530970
10. Wolfinger R. *SUGI Proceedings. Fitting Nonlinear Models with the New NLMIXED Procedure.* Cary/NC: SAS Institute Inc; 1999.
11. 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
12. 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
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/available from: <http://www.egms.de/static/de/journals/zma/2006-23/zma000264.shtml>

Korrespondenzadresse:

Dr. Klaus Böhme
 Universitätsklinik Freiburg, Lehrbereich Allgemeinmedizin,
 Elsässerstraße 2m, 79110 Freiburg, Deutschland, Tel.:
 +49 (0)761/270-27460, Fax.: +49 (0)761/270-27480
klaus.boehme@uniklinik-freiburg.de

Bitte zitieren als

*Böhme K, Schelling J, Streitlein-Böhme I, Glassen K, Schübel J, Jünger J. Vergleich kollegialer Einzel- mit Gruppen-Reviews allgemeinmedizinischer Multiple-Choice-Fragen. *GMS Z Med Ausbild.* 2012;29(4):Doc57. DOI: 10.3205/zma000827, URN: urn:nbn:de:0183-zma000827*

Artikel online frei zugänglich unter

<http://www.egms.de/en/journals/zma/2012-29/zma000827.shtml>

Eingereicht: 15.07.2011

Überarbeitet: 30.03.2012

Angenommen: 03.04.2012

Veröffentlicht: 08.08.2012

Copyright

©2012 Böhme et al. Dieser Artikel ist ein Open Access-Artikel und steht unter den Creative Commons Lizenzbedingungen (<http://creativecommons.org/licenses/by-nc-nd/3.0/deed.de>). Er darf vervielfältigt, verbreitet und öffentlich zugänglich gemacht werden, vorausgesetzt dass Autor und Quelle genannt werden.

Comparison of Collegial Individual and Group Reviews of General Practice Multiple Choice Questions

Abstract

Aims: In most German medical faculties, credits in general practice can be earned via exams using multiple-choice questions (MCQ). Measures such as peer-reviews may help assure the quality of these exams. In order to use time and personnel intensive peer reviews effectively and efficiently, the procedures used are key. Therefore, we wanted to find out whether there are differences between group and individual reviews regarding defined parameters.

Methods: We conducted a controlled cross-over study with three GP reviewers from four different German universities. Each reviewed 80 MCQs, 40 individually and 40 within a group, including external assessments by a panel of experts. Furthermore all reviewers were asked to evaluate the review process and the time spent carrying out these reviews.

Outcomes: We found no significant differences between the reliability and the validity of individual reviews versus group reviews. On average slightly more time was spent on group reviews compared with the individual reviews. The subjective assessments of the study participants regarding their satisfaction with the process and the efficiency and effectiveness of the reviews suggest a preference for group reviews.

Conclusions: Based on this study, there are no definite recommendations for or against either approach. When choosing between the two, the specific work structures and organisation at the local faculty should be taken into account.

Keywords: Medical Education, assessment, Multiple-Choice-Questions, Review

Klaus Böhme¹

Jörg Schelling²

Irmgard

Streitlein-Böhme³

Katharina Glassen⁴

Jeannine Schübel⁵

Jana Jünger⁶

1 University Hospital Freiburg, School of General Practice, Freiburg, Germany

2 LMU München, School of General Practice, München, Germany

3 University of Freiburg, Medical Faculty, Dean of Studies Office, Freiburg, Germany

4 University Hospital Heidelberg, Department for General Practice and Health Services' Research, Heidelberg, Germany

5 University Hospital Dresden, Carus GP Surgery, Dresden, Germany

6 University of Heidelberg, Competence Centre for Medical Exams, Heidelberg, Germany

Introduction and objectives

The Medical Licensure Act of 2002 [1] introduced the requirement for graded student performance records in each subject of the clinical study section in human medicine. For practical reasons, this is usually done in the form of written exams using multiple choice questions (MCQ), which are considered to be of satisfactory reliability and objectivity [2]. The literature contains general rules for creating "good" MCQs, both regarding form and content [3], [4], [5]. To date, many authors of MCQs have not been trained in the application of these rules at Ger-

man medical schools and, in many cases, no standardised review process for the questions used exists either [6]. Accordingly, there is no quality assurance of the MCQs in use.

To ensure an adequate standard of MCQs in use, a standardised peer-review process in addition to examiner training would appear suitable. The range of possible review process is wide, including individually reviewing questions or multiple individuals reviewing questions, moderated, un-moderated, face to face and virtual reviews [7].

Optimal effectiveness and efficiency of the peer reviews which are rather time and labour intensive depend on several factors. In this, the question of validity and reliab-

ility of the assessments play a central role. If one wants to strengthen the motivation of reviewers, they should be satisfied with the review process and they should be convinced of the effectiveness as well as the importance of it. Another important factor is the time required.

Since November 2008 the general medicine subject area at the University of Freiburg has been drawing upon a web-based digital exam system in the preparation of MC exams, developed by the “Centre of Excellence for Medical Exams in Baden-Württemberg” at the University of Heidelberg. The pool of questions for the general medicine subject area is held in this item management system (IMS) and used as the basis for creating exams at Freiburg.

In Germany, 15 departments now rely on the IMS as a digital tool in the preparation and evaluation of exams. While the system already significantly simplifies the organisational processes, a much more vital added benefit is that, in theory, a faculty can draw upon the exam questions of other faculties and use them for their own exams. Such access on the one hand requires the willingness of each faculty to share their questions with other faculties. On the other hand all users of the system had agreed that questions can only be placed in the public pool which can be accessed by other users if they have gone through a defined review process.

In the context of this study, individual reviews were compared with un-moderated face-to-face group reviews to investigate the following:

- Are there differences between individual and group reviews in (a) the frequency of errors found (consistency of review forms: reliability) and (b) compared to a standard review by designated experts (consistency with a standard: validity)?
- Does the review process influence the satisfaction of the reviewers with the process, their assessment of the effectiveness and importance of the reviews carried out?
- Is there a difference in the time required for implementing individual or group reviews?

Methods

Sample

Four general medical departments at German universities who declared their willingness to participate in the study were selected for the study: Dresden, Freiburg, Heidelberg and LMU Munich, hereinafter referred to in random order as Uni 1-4 for the sake of anonymity. Each site provided three employees who deal with the design and review of MCQs within their departments.

Item Sample

For the study, 2 x 40 MCQs (Group A and B) were randomly selected from the question pool of the general

medicine subject area at the University of Freiburg. They were all so-called Type A questions (single positive or negative selection from five answer choices).

Materials and technical prerequisites

One of the functions of the IMS is that all recorded exam questions can be reviewed online using a ten-criteria evaluation form (see Table 1)¹. An input field for free text comments makes it possible to record criticisms and to make specific suggestions for corrections.

The basis for assessing all reviews in this study was the “Short Guide to Reviewing MC Questions” by the competence centre, which in turn is based on the relevant literature on this subject [3], [4], [5].

For evaluating satisfaction with the review process, its effectiveness and the subjective assessment of its importance, a short questionnaire was created. Answers were given using a 6-point Likert scale (see Figure 1). In addition, there were two open questions: “Which part of the review did I like best?” and “Which part of the review caused me the most problems?”.

Conduct

The first step was creating a comparative standard for the assessment of study participants through a review of all 80 MCQs by experts. The four-member committee, all of whom had appropriate experience (MME or employee of the Centre of Excellence), was composed of three specialist representatives and a non-specialist colleague. The “Centre of Excellence for Exams in Medicine in Baden-Württemberg” subjected all 80 MCQs which were going to be part of the study to a group review.

The “Brief Guide to Reviewing MC Questions”, which explains the checklist criteria for the review (see Table 1), was made available to all study participants. No further training of the reviewers took place. According to the study design (see Table 2), each site the three reviewers then assessed 40 MCQs 40 individually and 40 MCQs in review groups.

For the group reviews, the study participants agreed dates when all 40 MCQs were assessed in a single session. In dealing with the individual reviews, the reviewers had the opportunity to work to their own timetable.

Following both the individual and the group reviews, the study participants were required to fill in the short questionnaire.

Statistics

Through a review of all MCQs used in this study by a panel of experts, a reference assessment, a gold standard, was set for each question. The assessments of this study were compared to this external reference to check its validity. The reliability was tested by comparing the identified deficiencies at the four different sites.

All of the 3200 individual assessments carried within this study’s reviews (80 MCQs × 10 assessment criteria × 4

Table 1: Checklist for the review of MC questions

		No deficiencies	Small deficiencies	Severe deficiencies
	Formal			
1.	The question type is acceptable/suitable			
2.	The question does not hidden answer clues (cueing)			
3.	The question can be understood, even without knowledge of the alternative answers and is answerable (e.g. no double negatives, ratio of question stem to answer options)			
4.	The response alternatives are homogeneous (all response options are about the same Length, usually max. of two lines; response alternatives which consist only of numbers or one term are sorted ascending)			
	Content			
5.	The question fits the classification (for example, subject)			
6.	The question is factually correct			
7.	The question is appropriate to the level of training			
8.	The question has a reference to practical application/patient vignette			
9.	The question stem and the question are comprehensible and plausible			
10.	All response options are comprehensible and plausible			

Questions on the review process						
1. Individual Review: 1= do not agree at all 6= fully agree						
	1	2	3	4	6	6
1.1 I was satisfied with the review process	<input type="checkbox"/>					
1.2 The review process was effective	<input type="checkbox"/>					
1.3 I think the review of questions if important	<input type="checkbox"/>					

Time needed for all 40 questions in **individual review:** minutes

Questions on the review process						
2. Group Review: 1= do not agree at all 6= fully agree						
	1	2	3	4	5	6
2.1 I was satisfied with the review process	<input type="checkbox"/>					
2.2 The review process was effective	<input type="checkbox"/>					
2.3 I think the review of questions if important	<input type="checkbox"/>					

Time needed for all 40 questions in **group review:** minutes

Figure 1: Short questionnaire on the reviews

sites) were dichotomized as follows: 0 = no deviation from the expert review and 1 = deviation from the expert review and 0 = no defect and 1 = slight or severe defect and thus subjected to statistical analysis. For the individual reviews, the rounded average of the three individual assessments was used for comparison.

The variables “Number of deviations from the gold standard” and “Number of errors found” were analysed using variance analysis (linear mixed model [8]) with the fixed factors “review form” (single/group), “location” (Uni 1 - Uni 4), “question group” (MCQ Group A/MCQ Group B) and “pass” as well as the random factor “MCQ” (i.e. it is assumed that the MCQs represent a random selection

Table 2: Flow chart for review of general medical exam questions (Group A=MCQs 1-40, Group B=MCQs 41-80)

	Uni 1	Uni 2	Uni 3	Uni 4
1. pass	1 x Group review (Group B)	3 x Individual review (Group A)	3 x Individual review (Group B)	1 x Group review (Group A)
2. pass	3 x Individual review (Group A)	1 x Group review (Group B)	1 x Group review (Group A)	3 x Individual review (Group B)

from the MCQ pool). Of primary interest were the factors “review form” and “location”, the other factors were used as control variables.

The variables for the individual categories are binary (values 0 and 1) and were therefore analysed with a similar non-linear mixed model with a logistic link function (logistic-normal model [9], [10]).

The statistical analysis of the questionnaires accompanying the study was done descriptively.

Results

Sample Composition

Twelve GPs from four German universities, eight women and four men, took part in the study. The participants' ages ranged from 24 to 61 years, the average age was 41 years (SD=14.6).

Validity

The results of the deviations of the study reviews from the expert reviews are presented in Table 3. There were no statistically significant differences, either between the type of review (single vs. group review), or between Pass 1 and Pass 2, or between Question Group A and Question Group B. A significant difference was found only between the individual reviewer groups of the different locations.

Reliability

The number of deficiencies identified as part of the review study are shown in Table 4. Overall, no significant differences between the nature of the review (single vs. group review), the pass and the question group were found for all the 3200 criteria. The only significant difference was also found to be between the sites.

Questionnaires

In response to the question regarding satisfaction with the review process, following the individual reviews the study participants responded with an average of 4.92 (SD=0.69), following the review group with 5.17

(SD=0.83). The question on effectiveness was answered as follows: following individual reviews with an average of 4.92 (SD=0.69), following group review with 5.58 (SD=0.67). The importance of the review following individual reviews was rated 5.75 (SD=0.45) on average, following group review 6.00 (SD=0).

The free text comments on the questions “Which part of the review did I like best?” and “Which part of the review caused me the most problems?” can be summarised as follows:

From the perspective of the reviewer, the most positive aspect of the individual reviews was the freedom in time-planning, where for the group reviews it was the collegial exchange of ideas and the associated learning effects. Commonly cited problems in individual review were complaints of vague evaluation criteria in the absence of opportunities for asking questions, whereas in the group reviews it was timing problems, as it is often not easy to produce consensus in carrying out assessments and the long sessions associated with this type of review and the resulting concentration problems.

Time Expenditure

The average processing time for the 40 questions in individual reviews was 113 minutes (SD=44), 139 minutes (SD=48) in group reviews.

Discussion

There is already sufficient evidence in German-speaking countries that the systematic collegial review of MCQs improves the quality of exam questions [11], [12], [13]. There are only a few indications as to the advantages and disadvantages of different review process [7]. This study was intended to help clarify the practical questions about the advantages and disadvantages of individual vs group reviews.

Validity

The results of this study show that are review processes are equally valid. There is therefore no basis upon which we could recommend one over the other.

Table 3: Number of deviations from the expert review in the reviews of the four university sites in the individual (IR) and group reviews (GR) in all individual criteria of the review checklist, including levels of significance

no deviations	Uni 1		Uni 2		Uni 3		Uni 4		Total
	IR	GR	IR	GR	IR	GR	IR	GR	
Deviations (in percent)	330 (17.50%)	347 (13.25%)	322 (19.50%)	321 (19.75%)	339 (15.25%)	334 (16.50%)	279 (30.25%)	262 (34.50%)	2534 (20.81%)
Total	400	400	400	400	400	400	400	400	3200

Differences: Uni – F(79.3) = 45.82; p<0.0001
 Review type - F(79.1) = 0.10; p=0.7499

Table 4: Number of separate criteria assessed as deficient of the 4 university sites in individual (IR) and group reviews (GR), including levels of significance

no deficiencies	Uni 1		Uni 2		Uni 3		Uni 4		Total
	IR	GR	IR	GR	IR	GR	IR	GR	
Deficiencies (in percent)	356 (11.00%)	353 (11.75%)	364 (9.00%)	352 (12.00%)	328 (18.00%)	349 (12.75%)	282 (29.50%)	266 (33.50)	2650 (17.18%)
Total	400	400	400	400	400	400	400	400	3200

Differences: Uni – F(79.3) = 67.59; p<0.0001
 Review type - F(79.1) = 0.28; p=0.6006

Statistically significant differences in the assessments of MCQs were found, however, between the reviewers of the four individual sites. The “Short Guide to Reviewing MC Questions” alone, although based on the same basis as the knowledge of the “experts”, apparently is not suitable for ensuring homogeneous assessments across the four university sites. In terms of the deviations from expert’s review, there was one significant difference across all checklist items. The result suggests that in addition to author training required elsewhere [4] and wide-scale introduction of a standardised review process, reviewer training should be recommended [13].

Reliability

The deficiencies observed at the four locations not dependent on whether the questions were assessed individually or in groups were viewed across all checklist items. There is therefore, again, no basis for deciding for a particular version.

In contrast, the comparison of university sites with each other revealed statistically significant differences for all items which could be statistically analysed. The demand for reviewer training can thus be affirmed.

Questionnaires

If the goal is to ensure that throughout Germany as many tutors participate in a mutual and sustainable review of MCQs, it must be ensured that they are convinced of the necessity of such an approach and that the methodology is effective and subjectively satisfactory from their point of view. For this reason, the individual reviews were accompanied by questionnaires which queried these parameters.

Averaged across all four sites, for both the question regarding satisfaction, the effectiveness of the review process and the importance of the review, slightly higher

approval was awarded following group reviews compared to individual reviews. When analysing the questions for each site separately, the higher approval following group reviews is replicated almost consistently. Only at one site (Uni 1), was the question regarding satisfaction with the review process ranked higher following individual reviews. In this case too, group dynamics should be considered, as the Uni 1 site took the longest of all review sessions, with 210 minutes of processing time in group reviews. This could suggest problems when trying to come to decisions with joint responsibility.

The free text comments on individual and group reviews reflect the expected advantages and disadvantages of both methods once again, juxtaposing the freedom to plan in individual reviews with of the lack of feedback opportunities from colleagues. It becomes obvious that in group reviews the timing issues are in contradiction with the often mentioned (and positively seen) collegial exchange of ideas and the associated learning effects.

Time Expenditure

If we consider only the pool of questions of general medicine in Freiburg which includes approximately 280 MCQs (though other sites may have much larger pools), we can guess the enormous amount of time necessary for reviewing the entire set for all general medical departments and subject areas. In this respect, the timings found by this study may be considered when discussing the more practical review process. Group reviews of 40 MCQs on average took 26 minutes longer than individual reviews. If one followed this criterion alone, there would have to be a pronounced recommendation of individual reviews. Interestingly, this extra need for time is not raised in the free text answers.

Weaknesses

The relatively small sample of reviewers, which for feasibility reasons was not a random sample also limits the generalisability of the study results. As employees of general medical departments and subject areas were consciously selected (i.e. people who in their daily routine are tasked with creating and/or reviewing exam questions without special additional training) this meant that inevitable discrepancies in medical experience and prior theoretical knowledge had to be accepted. Both factors limit the comparability.

The ten-item checklist against which the MCQs were judged did not reflect all criteria for the creation of "good" MCQs given in the literature [3], [4], [5]. Some items must be considered redundant while other criteria, such as the important question of the relevance of MCQs [11], are missing.

Conclusions

The results of this study do offer some help when trying to reach a decision about which review process may be recommended as the better.

As was shown in the results section and explained in the discussion, no statistically significant difference regarding validity and reliability of both procedures could be found which means that it will not be possible to decide for one or the other solely based on these factors.

Against this background, the subjective assessments of the study participants on the reviews gain in weight. These tended more towards group reviews. It would appear that the time factor is subordinated to the satisfaction and the subjective assessment of the effectiveness of the process by the study participants.

The specifics of the situation on the ground can understandably play a crucial role when selecting a method. If the reviewers are scientific or clinically-based staff who usually work together in a department, the selection will tend towards group reviews which are relatively easy to terminate. If the reviews will tend to be carried out by established lecturers (as is common in general medicine), individual reviews at their own workplace could seem more reasonable and practicable.

For practical reasons this study only looked at Type A questions. But essentially, the results should be transferable to other question formats.

Note

¹ In the IMS this evaluation form is based on an algorithm for which no further details are available which either makes the questions available to the public folder or refers them to the authors for correction.

Acknowledgement

My special thanks for support with the statistical aspects of this project goes towards Dr. Andreas Möltner, at the Competence Centre for Exams in Medicine - Baden Württemberg, University of Heidelberg

Competing interests

The authors declare that they have no competing interests.

References

1. Bundesministerium für Gesundheit. Approbationsordnung für Ärzte vom 27.06.2002. BGBl. 2002:2405-2435.
2. Möltner A, Schellberg D, Jünger J. Grundlegende quantitative Analysen medizinischer Prüfungen. *GMS Z Med Ausbild.* 2006;23(3):Doc53. Zugänglich unter/available from: <http://www.egms.de/static/de/journals/zma/2006-23/zma000272.shtml>
3. Haladyna TM, Downing SM, Rodrigues MC. A review of multiple-choice item-writing guidelines for a classroom assessment. *Appl Meas Educ.* 2002;15:309-344. DOI: 10.1207/S15324818AME1503_5
4. Krebs R. Anleitung zur Herstellung von MC-Fragen und MC-Prüfungen für die ärztliche Ausbildung. Bern: Institut für Medizinische Lehre IMS, Abteilung für Ausbildungs- und Examensforschung AAE; 2004.
5. AG Progress Test Medizin. Progress Test Medizin. Leitfaden für Fragenautorinnen und -autoren des Progress Test Medizin. Berlin: Charité Universitätsmedizin; 2003.
6. Jünger J, Möltner A, Lammerding-Köppel M, Rau T, Obertacke U, Biller S, Narciß E. Durchführung der universitären Prüfungen im klinischen Abschnitt des Medizinstudiums nach den Leitlinien des GMA-Ausschusses Prüfungen: Eine Bestandsaufnahme der medizinischen Fakultäten in Baden-Württemberg. *GMS Z Med Ausbild.* 2010;27(4):Doc57. DOI: 10.3205/zma000694
7. Kazubke E, Schüttpelz-Brauns K. Gruppenleistungen beim Review von Multiple-Choice-Fragen – ein Vergleich von face-to-face und virtuellen Gruppen, mit und ohne Moderation. *GMS Z Med Ausbild.* 2010;27(5):Doc68. DOI: 10.3205/zma000705
8. Brown H, Prescott R. *Applied Mixed Models in Medicine*, Second Edition. Oxford/UK: John Wiley & Sons, Ltd; 2006. DOI: 10.1002/0470023589
9. Beitel PJ, Landis JR. A Mixed-effects Model for Categorical Data. *Biometr.* 1985;41:991-1000. DOI: 10.2307/2530970
10. Wolfinger R, SUGI Proceedings. Fitting Nonlinear Models with the New NLMIXED Procedure. Cary/NC: SAS Institute Inc; 1999.
11. 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
12. 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

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/available from: <http://www.egms.de/static/de/journals/zma/2006-23/zma000264.shtml>

Please cite as

Böhme K, Schelling J, Streitlein-Böhme I, Glassen K, Schübel J, Jünger J. Vergleich kollegialer Einzel- mit Gruppen-Reviews allgemeinmedizinischer Multiple-Choice-Fragen. GMS Z Med Ausbild. 2012;29(4):Doc57.
DOI: 10.3205/zma000827, URN: urn:nbn:de:0183-zma0008277

This article is freely available from

<http://www.egms.de/en/journals/zma/2012-29/zma000827.shtml>

Corresponding author:

Dr. Klaus Böhme
University Hospital Freiburg, School of General Practice,
Elsässerstraße 2m, 79110 Freiburg, Germany, Phone:
+49 (0)761/270-27460, Fax: +49 (0)761/270-27480
klaus.boehme@uniklinik-freiburg.de

Received: 2011-07-15

Revised: 2012-03-30

Accepted: 2012-04-03

Published: 2012-08-08

Copyright

©2012 Böhme et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by-nc-nd/3.0/deed.en>). You are free: to Share – to copy, distribute and transmit the work, provided the original author and source are credited.