2010; 32: e300–e307 MEDICAL TEACHER

WEB PAPER

The modified essay question: Its exit from the exit examination?

EDWARD J. PALMER[1,2], PAUL DUGGAN[3], PETER G. DEVITT[1] & ROHAN RUSSELL[4]

[1]Department of Surgery, University of Adelaide, Adelaide, Australia, [2]Center for Learning and Professional Education, University of Adelaide, Adelaide, Australia, [3]Department of Obstetrics and Gynecology, University of Adelaide, Adelaide, Australia, [4]Royal Adelaide Hospital, Adelaide, Australia

Abstract

Background: Exit examinations in medicine are 'high stakes' examinations and as such must satisfy a number of criteria including psychometric robustness, fairness and reliability in the face of legal or other challenges.

Aims: We have undertaken a critical review of the exit examination from the University of Adelaide focussing on the written components. This examination consisted of an objective structure clinical examination (OSCE), a multiple choice question (MCQ) paper and a modified essay question (MEQ) paper.

Methods: The two written papers were assessed for item writing flaws and taxonomic level using modified Bloom's criteria. Curriculum experts independently assessed adequacy of the examination for validity and fidelity.

Results: The overall examination had good fidelity and validity. The results of the MEQ and MCQ were strongly and positively correlated and there was a weak negative correlation between these papers and the OSCE. The MEQ had a higher proportion of questions focussed on recall of knowledge and the questions were more structurally flawed compared with the MCQs. The MEQ re-marking process resulted in lower scores than were awarded by the original, discipline-based expert markers. The MEQ paper failed to achieve its primary purpose of assessing higher cognitive skills.

Conclusion: The University of Adelaide's MBBS programme has since dropped the MEQ paper from its exit examination and is evaluating in its place the Script Concordance test.

Introduction

The exit examinations of a medical programme are the means by which the graduate is assessed to be sufficiently and suitably competent and knowledgeable to allow safe practice in a supervised capacity as an intern, where students will continue to learn and refine their skills.

To make such an assessment the examination must be structured to ensure that students have adequate core knowledge and understanding of each discipline of the course and show proficiency in the key competencies associated with those disciplines. Students will 'learn' what they think they will be assessed on.

Traditionally, these exit examinations contain both a written and a clinical component. The written component tests and measures knowledge and understanding whilst the clinical component assesses practical skills and competence. In efforts to ensure fairness, objectivity and to minimise the logistical burden of setting and running these examinations, current examination processes often involve multiple choice questions (MCQs), some form of short answer question and a structured circuit of short clinical stations. The MCQ (and its many variants) has often been considered to test factual recall and the short answer question an opportunity to test logical thinking, judgement and application (Stratford & Pierce-Fenn 1985; Wass et al. 2001).

Practice points

- MEQs are often included in examinations to test higher order cognitive skills. MCQs are often regarded as testing knowledge recall only.
- When measured against Bloom's taxonomy, MEQs developed for a higher education assessment primarily tested lower order cognitive skills such as recall of knowledge.
- MCQs in the same assessment were found to be more likely to test higher order cognitive skills such as analysis and management.
- Many items in the MEQ and MCQ assessment suffered from IWFs and those in the MEQ were deemed more significant.
- Marking guides developed for the MEQ assessment were inadequate and lead to inconsistencies in marking.
- The MEQ was deemed to be an unsuitable assessment instrument and has been replaced by material in the script concordance format.

Students are well aware that the exit examination is a high stakes process and failure may lead to exclusion from the programme, financial hardship and emotional trauma. Apart from setting a target for study and achievement, this exit

Correspondence: P. G. Devitt, Department of Surgery, University of Adelaide, Royal Adelaide Hospital, Adelaide 5000, Australia. Tel: +61 8 82225144; fax: +61 8 82225876; email: peter.devitt@adelaide.edu.au

ISSN 0142–159X print/ISSN 1466–187X online/10/070300–8 © 2010 Informa Healthcare Ltd.
DOI: 10.3109/0142159X.2010.488705

examination is itself a target for those who fail and seek to challenge the assessment process. It behooves the examining institution to have an assessment process which has sufficient psychometric and statistical robustness to withstand these potential challenges and be confident that fairness is seen to prevail.

All assessments involve considerable time, effort and expense in their development, organisation, running and marking. A MCQ paper can be laborious to construct, but its marking is usually relatively easy. The short answer or modified essay question (MEQ) paper on the other hand can be easier to develop but is always much more onerous and difficult to mark (Lockie et al. 1990). Both formats have been criticised for their inability to consistently and rigorously test clinically relevant material to the high standards of a medical graduate where the emphasis is expected to be on assessment of the higher cognitive skills (Ferguson 2006; Epstein 2007; Palmer & Devitt 2007).

We have undertaken a study to assess the robustness of an exit examination with a critical analysis of the quality of its two written components (MCQ and MEQ) and an analysis of the validity and fidelity of the examination as a whole and of its three components.

Methodology

The exit examination for students at the University of Adelaide is held at the end of the fifth year of study, which coincides with the end of the second year of intense clinical exposure. At the end of 2007, the examination consisted of one MCQ paper of 180 questions undertaken over 3 h, a 15-question, 62-part MEQ paper over 3 h and an objective structure clinical examination (OSCE) of 18 questions over 2.1 h. The quality of the written (MCQ and MEQ) components of the 2007 examination, taken by 146 students, was analysed with regards to flaws in the questions or marking schemes and the level of cognitive ability tested. The validity and fidelity of the MEQ, MEQ and OSCE were also analysed.

The three components (MEQ, MCQ and OSCE) of the exit examination were set using a blueprint developed by the MBBS Curriculum Committee that broadly covered the clinical and basic science material in the first 5 years of the MBBS Programme. The OSCE examination was designed to assess history-taking, examination, test interpretation, management, counselling, simple clinical procedures, clerking and interactional skills. The two written papers were intended to assess knowledge of clinical and basic science with an emphasis on application to clinical problem-solving. The content of the examination was determined by representatives from the Disciplines of Medicine, Surgery, Paediatrics, Obstetrics and Gynaecology (O&G), Psychiatry, Pathology and General Practice (rural and urban).

MCQ question selection

Seventy-six percent of the MCQs were made available from the Australian Medical Council (AMC) (Australian Medical Council 2009) and were chosen from a bank of previously used questions on the basis of content and satisfactory point biserial analysis (>0.20). These questions had undergone a comprehensive review process by the AMC, were selected for possible inclusion by a multi-disciplinary panel, and were finally assessed by a representative from each discipline group of the university for appropriateness to the curriculum before acceptance for the exit examination. The remaining questions had been developed by staff of the university and had been used in previous years' examinations. These questions also had satisfactory point biserial data and had all been evaluated by multi-disciplinary panels within our faculty.

MEQ question selection

The MEQs were submitted by Disciplines and reviewed in a workshop attended by between 4 and 11 discipline experts per question. Some groups included senior representatives from a sister medical school and/or from clinical staff employed in South Australian Teaching Hospitals who were involved in the teaching of medical students. In construction of the MEQs, a request was made that each question contain at least one aspect of basic science. Questions were reviewed by members of the broad Discipline group (i.e. physicians reviewed internal medicine questions, obstetricians reviewed obstetrics questions, etc.). Questions were first edited for clarity of language, appropriateness to the curriculum, and appropriateness of weighting and accuracy of the marking template. Changes to the questions and marking template were made as required and a modified Angoff method was then employed to determine the borderline mark for each question. This involved a moderate discussion of the expected performance of the borderline candidate, confidential and independent recording of the expected borderline score by each member of the group for each part of the question, discussion of the range of de-identified scores, and concluding with a second round of confidential and independent scoring to be used as the final borderline score. Marking was arranged using whatever internal process the relevant disciplines deemed appropriate. For 14 of the 15 questions, this involved a single, senior examiner marking a question and for one discipline multiple examiners were used. All markers were experts in the relevant discipline.

Analysis of the MCQ and MEQ papers

The study began after the examination results were notified to the candidates. The research team consisted of three university academics, each with at least 8 years of experience in marking examinations and one Year 6 student who had recently passed the examinations (RR). Two of the academics were clinical experts in the field of Surgery (PGD) and Obstetrics and Gynaecology (PD). The other (EP) was from a non-clinical background. Each marker remarked the entire MEQ examination using the provided templates. Each marker was given instructions to follow the marking template as closely as possible. The margins of the MEQ answer sheets were covered to obscure the original marker's score.

Each question in the written papers (MCQ and MEQ) was analysed by each of the markers to test for item writing flaws (IWFs) (Tables 1 and 2) (Palmer & Devitt 2006) and using a

Table 1. Rating scale used to judge the rigour of the MCQs according to the presence of any IWFs.

Rating	Conditions required to achieve rating
1	Pass the cover test and no IWFs
2	Pass the cover test and 1–2 IWFs
3	Cover test dubious and no IWFs
4	Fail the cover test and 1–2 IWFs
5	Fail the cover test and more than 2 IWFs

Table 2. Rating scale used to judge the rigour of the MEQs according to the presence of any IWFs.

Rating	Conditions required to achieve rating
0	No identified IWFs
1	Question is ambiguous or open to misinterpretation (writer expects a response which differs from what the question asks)
2	Clarity of question: grammatically unsound/use of vague language/imprecise terms/unexplained abbreviations/ miscuing
3	Question does not promote synthesis by testing a logical progression of thought. Failure of continuity between the sections of a question
4	Failure at one point in the question may lead to failure in other parts
5	Double negatives in question
6	Unnecessary information in question
7	Cueing to other parts of question
8	Question covers content, which is not necessarily accepted or is contentious
9	Question is too broad in scope

Table 3. Modified Bloom's taxonomy.

Level I	Knowledge –recall of information
Level II	Comprehension and application –understanding and being able to interpret data
Level III	Problem-solving –use of knowledge and understanding in new circumstances

modified Bloom's taxonomy to test the level of cognitive ability being measured by the question (Table 3) (Bloom 1956). A schema was developed to address flaws in the marking template for the MEQ (Table 4). The analysis required that initially each marker independently evaluated every question and this was followed by a group analysis where differences were discussed and a group consensus was reached.

The MEQ paper was scrutinised for inter-marker reliability and in particular focussed on those students at the pass–fail boundary, where marking or assessment issues are likely to have the greatest impact. A repeated measures ANOVA test was used with the Tukey *post hoc* test to look for differences between markers.

To obtain an understanding of the validity and fidelity of the examination, a total of 13 experts in the curriculum covering the range of clinical disciplines were approached. One declined to participate, and two who completed the survey were unable to comment in relation to the curriculum as a whole and were excluded from the analysis. The results are reported for the remaining 10 experts.

Table 4. Rating scale used to judge the rigour of the marking scheme for MEQs according to the presence of any IWFs.

Rating	Conditions required to achieve rating
0	No identified IWFs
1	Scheme is incomplete (i.e. does not include important or correct alternative answers)
2	Scheme is highly specific/does not allow for minor variability from 'model response'
3	Scheme is difficult to apply (i.e. awards difficult fractions of marks or is unclear where marks are allocated)
4	Scheme is poorly weighted. Inconsistent, dubious 'relevance'/application/difficulty. Inappropriate to level of student. Answer length is not proportional to the marks allocated
5	Marks are awarded for the same answer at multiple points in the question

The experts were provided with a summary of each component of the exit examination.

- For the MCQ, the summary identified the discipline and described the theme and topic for the 180 questions;
- for the MEQ, the topic description for the 15 questions and a short summary of the 62 component parts to the questions was provided;
- for the OSCE, the discipline was identified and a short summary of the tasks for each station was provided.

The experts were invited to provide free responses in relation to the assessments and asked to rate the validity and fidelity of the components of the exit examination and the examination as a whole, as follows.

Validity

(1) How adequately does the (MEQ, MCQ and OSCE) examination sample the breadth of the curriculum?
(2) Overall, how adequately do the combined exit exam components sample the breadth of the curriculum?

Fidelity

(1) How adequately does the exit (MEQ, MCQ and OSCE) examination reproduce the challenges of clinical medicine?
(2) Overall, how adequately do the combined exit exam components reproduce the challenges of clinical medicine?

The possible responses were: very adequately; adequately; neither adequately nor inadequately; inadequately; very inadequately. Responses of 'very adequately' or 'adequately' were combined for the analysis and regarded as being in agreement that the validity or fidelity was satisfactory. All other responses were regarded as unsatisfactory.

Results

The research team each required approximately 3 h per question to mark the 15-question MEQ paper. The results of

Table 5. Marking results for all markers compared with official marks (maximum 12 per question).

Question number (discipline)	Original marker	Marker 1 student	Marker 2 non-clinical	Marker 3 clinical	Marker 4 clinical
1 (Surgery)	9.8 ± 0.2	9.2 ± 0.1[1]	9.5 ± 0.2	8.7 ± 0.1[1]	8.2 ± 0.1[1]
2 (Psychiatry)	8.5 ± 0.2	8.2 ± 0.2[1]	7.5 ± 0.2[1]	8.1 ± 0.2[1]	8.3 ± 0.2[1]
3 (Obstetrics)	6.3 ± 0.1	5.8 ± 0.1	6.2 ± 0.2	5.6 ± 0.2[1]	6.1 ± 0.1
4 (Surgery)	7.5 ± 0.2	7.3 ± 0.5	6.7 ± 0.1	6.7 ± 0.2	7.0 ± 0.1
5 (General Practice)	5.0 ± 0.1	5.7 ± 0.1[1]	6.2 ± 0.2[1]	6.4 ± 0.1[1]	6.1 ± 0.1[1]
6 (Psychiatry)	8.0 ± 0.2	6.9 ± 0.2[1]	6.8 ± 0.1[1]	6.6 ± 0.1[1]	7.6 ± 0.2[1]
7 (Paediatrics)	7.3 ± 0.2	7.0 ± 0.2[1]	6.7 ± 0.2[1]	6.6 ± 0.2[1]	7.2 ±v0.2
8 (Gynaecology)	8.4 ± 0.1	8.5 ± 0.1	8.5 ± 0.1	8.9 ± 0.2[1]	8.3 ± 0.1
9 (Paediatrics)	6.7 ± 0.1	6.8 ± 0.1	6.8 ± 0.1	5.4 ± 0.1[1]	6.0 ± 0.1[1]
10 (General Practice)	5.7 ± 0.2	5.7 ± 0.2	5.5 ± 0.2	5.5 ± 0.2[1]	5.9 ± 0.2[1]
11 (Surgery)	6.5 ± 0.2	4.1 ± 0.1[1]	4.4 ± 0.2[1]	3.4 ± 0.1[1]	5.2 ± 0.2[1]
12 (Medicine)	9.7 ± 0.1	9.3 ± 0.1[1]	9.5 ± 0.1[1]	9.5 ± 0.1[1]	9.5 ± 0.1
13 (Medicine)	7.1 ± 0.1	6.3 ± 0.1[1]	7.2 ± 0.1	6.4 ± 0.1[1]	7.0 ± 0.1
14 (Pathology)	5.9 ± 0.2	5.8 ± 0.2	5.8 ± 0.2	5.7 ± 0.2	6.7 ± 0.2[1]
15 (Medicine)	7.9 ± 0.1	7.3 ± 0.1[1]	6.8 ± 0.1[1]	7.7 ± 0.1	6.9 ± 0.1[1]
Total	110 ± 1	104 ± 1[1]	104 ± 1[1]	101 ± 1[1]	106 ± 1[1]
number of questions significantly different from original marker	n/a	10	7	12	9

Notes: Results show the mean and standard error.
1: Significant difference with official results as indicated by Tukey *post hoc* test on repeated ANOVA.

Table 6. Modified Bloom's categorisation for MCQ and MEQ questions.

Question type	MCQ (total 180)	MEQ (total 65 stages)
Bloom's level 1	84 (47%)	50 (77%)
Bloom's level 2	70 (39%)	10 (15.3%)
Bloom's level 3	26 (14%)	5 (7.7%)

Table 7. Contingency table for Fisher's exact test.

	Bloom's level 1	Bloom's level 2, 3	Item flawed	Item not flawed
MCQ	84	96	65	115
MEQ	50	15	41	24
p		<0.0001		=0.0002

this marking compared with the original marking are shown in Table 5. All markers showed significant differences between their marking and the original marks allocated in seven or more of the 15 questions, and all markers provided a median mark significantly different from that provided by the original marker. (Table 5) The mean difference over the entire examination ranged from 4.0 to 8.8 marks below the original mark out of a total possible 180 marks. The expert in Surgery marked significantly differently to the original expert marker in two of three surgical questions. The expert in O&G did not mark significantly differently to the original expert marker in the two O&G questions.

The modified Bloom's categorisation of the MCQs and MEQs (Table 6) shows that there was a greater proportion of MEQs testing lower level cognitive skills than MCQs ($p<0.0001$, Fisher's exact test) (Table 7). Over half of the MCQs were judged to be assessing level 2 or higher cognitive skills, whereas more than three quarters of the MEQ stages were deemed to be testing knowledge only (Table 7).

Table 8. IWFs for MCQs.

Question type	MCQ (total 180)
No IWFs (level 1)	115
Pass the cover test and 1–2 IWFs (level 2)	27
Cover test dubious and no IWFs (level 3)	21
Fail the cover test and 1–2 IWFs (level 4)	15
Fail the cover test and more than 2 IWFs (level 5)	2
Not phrased as a question	58

The MEQ paper had a significantly greater proportion of IWFs ($p=0.0002$, Fisher's exact test) (Table 7). Sixty-four percent of the MCQs were without flaw (Table 8) compared with 51% of the MEQ stages (Table 10). Sixty percent of the MEQs had at least one marking scheme flaw (Table 9).

Table 10 shows the effect on pass/fail decisions of variations in marking of the MEQ paper between the four examiners (Column 1). Column 2 records the number of candidates whose official score was ≥50% and who scored <50% in the remark. The University of Adelaide will allow candidates who score 45–49% in this examination to proceed to Year 6 provided they pass the other exit assessments. 'Critical' means that the student moves from a D to an E grade and would thus fail the year purely on this result. 'Could be critical' means the student moves from a C to a D grade and could fail the year but only if a D grade was received in one of the other exit examination assessments. 'Not critical' means a reduction in grade that could have no effect on pass/fail decisions.

Results: Validity and fidelity of the exit examination

Seven of the 10 experts felt that overall the exit examination had satisfactory validity and six experts felt that overall the exit examination had satisfactory fidelity (Table 11). For the components of the exit examination, it was notable that only

Table 9. Item writing and marking scheme flaws for MEQs.

Question type	MEQ (total 65)
No identified IWFs (level 0)	33
Question is ambiguous or open to misinterpretation (writer expects a response which differs from what the question asks). (level 1)	11
Clarity of question: grammatically unsound/use of vague language/imprecise terms/unexplained abbreviations/miscuing (level 2)	3
Question does not promote synthesis by testing a logical progression of thought. Failure of continuity between the sections of a question. (level 3)	8
Failure at one point in the question may lead to failure in other parts (level 4)	3
Double negatives in question (level 5)	0
Unnecessary information in question (level 6)	1
Cueing to other parts of question (level 7)	10
Question covers content, which is not necessarily accepted or is contentious (level 8)	1
Question is too broad in scope (level 9)	3
Marking scheme flaws	
No identified flaws (level 0)	26
Scheme is incomplete (i.e. does not include important or correct alternative answers) (level 1)	28
Scheme is highly specific/does not allow for minor variability from 'model response' (level 2)	3
Scheme is difficult to apply (i.e. awards difficult fractions of marks unclear where marks are allocated) (level 3)	14
Scheme is poorly weighted. Inconsistent, Dubious 'relevance'/application/difficulty. Inappropriate to level of student. Answer length is not proportional to the marks allocated (level 4)	3 (always with other flaws)
Marks are awarded for the same answer at multiple points in the question (level 5)	2 (always with other flaws)

Table 10. The effect of marking variations on determination of student grades.

	Extra candidates <50% (*n*)	Critical (*n*)	Could be critical (*n*)	Not critical (*n*)	Unaltered (*n*)
Student	13	6	13	41	86
Non-clinician	9	5	9	42	90
Clinician A	15	8	13	61	64
Clinician B	7	2	7	38	99

Note: *n*, number of students in each category.

Table 11. Percentage of experts who strongly agreed or agreed with statements regarding the validity and fidelity of the assessments.

	MEQ	MCQ	OSCE	Overall
Validity	50	60	60	70
Fidelity	40	30	60	60

four experts felt that the MEQ had satisfactory fidelity and only three experts felt that the MCQ had satisfactory fidelity. Experts were evenly divided regarding the validity of the MEQ.

The free responses are summarised below:

- The assessments have not sampled adequately knowledge and understanding of public health and health systems and evidence-based practice.
- There is an under-representation of topics from certain subspecialties and an over-representation from others.
- Discipline-based examinations would be better.
- The exit examination should be considered as a whole.
- The availability of good questions is dependent on the availability of staff to write them.
- Core areas and goals of the curriculum need to be redefined in relation to the initial career path (hospital intern) of the graduates.
- Fidelity would be improved by additional clinical examinations held throughout the year.

The reliability and correlation coefficients for the three components of the Year 5 exit examination are shown in Table 12.

Discussion

The combination of the MCQ, MEQ and OSCE examinations showed an assessment with high validity (sampling the breadth of the curriculum) and moderate fidelity (reproducing the challenges of clinical medicine) as adjudged by experts in the curriculum. This is a reflection both on the efforts of the contributors to the examination and the rigorous organisational efforts involved in setting a critical examination. The reliability of each of the examination components was satisfactory. It is quite reasonable to expect high reliability from 3 h written papers and this was observed with the Cronbach alpha reliability coefficient of 0.84 for both the MCQ and MEQ. The Cronbach alpha for the 2 h OSCE paper was 0.58 and just below bounds described in the literature for that length of assessment (Schwartz et al. 1998).

In terms of validity the individual examinations scored less well than the combined assessment. This is expected because the exit examination was structured so that its three component examinations complemented each other. Our data indicate that this combined approach was successful in the exit assessment that we have evaluated.

The curriculum experts were less generous in their views regarding the fidelity of the overall assessment and its components. This is possibly a reflection on the rules of the assessment that requires inclusion of basic sciences, which may not necessarily be directly clinically relevant, in the MEQ and MCQ papers. Alternatively, this could reflect an opinion that the number of questions was insufficient to achieve reasonable fidelity, but that is a less plausible explanation as the validity was acceptable and reliability statistics for the assessments are good. Another interpretation is that the experts were unreasonably harsh in their judgement of the written papers, particularly the MCQ paper. The poor rating for fidelity of the MCQ paper was unexpected because the majority of the questions in that paper were clinical questions that had been obtained from the AMC. These questions had undergone a rigorous quality control process in the construction of the questions, and had been shown to perform well in the AMC examination cohort (overseas trained doctors seeking registration to practice medicine in Australia) before being selected for our MCQ assessment. In addition to the AMC quality process, the MCQ questions were reviewed by our experts in the

Table 12. Reliability and correlation coefficients for the three components of the Year 5 exit examination.

Exam (146 candidates)	Number of questions	Duration (h)	Cronbach alpha	Pearson correlation coefficient (referenced to OSCE)	Pearson correlation coefficient (referenced to MEQ)	Pearson correlation coefficient (referenced to MCQ)
OSCE	18	2.1	0.58	1	−0.2	−0.2
MEQ	15	3.0	0.84	−0.22	1	0.77
MCQ	180	3.0	0.84	−0.22	0.77	1

appropriate disciplines before inclusion in the exit examination and were considered to be appropriate to the curriculum and clinically relevant. Our data show that the MCQ questions performed well in the modified Bloom's analysis. We did not provide our psychometric experts with copies of the MCQ questions (forbidden by confidentiality agreements with the AMC) and it is possible that our description of the MCQ questions failed to convey the quality and content of the questions adequately. It is also possible that there was a belief in our pool of experts that MCQs are not appropriate ways of testing, perhaps because there is a strong school of thought that MCQs can only test knowledge.

The curriculum experts identified gaps in the assessment that are most likely related to the broader strengths and weaknesses of the MBBS programme (which is dependent on the interest and availability of academic and clinical staff to undertake teaching and assessment). The value of an exit examination versus discipline-based assessments was raised. Whether disciplines would mount better examinations is debateable, particularly as the effort required to generate reliable assessments requires assistance that may not be afforded within individual discipline budgets. It is very doubtful in a programme of this size that all clinical disciplines have enough experts to undertake satisfactory question preparation and analysis or enough academic and clinical staff to run independently the number of questions required for clinical examinations. In contrast, there are a number of strengths of the multi-disciplinary exit examination, including cross-disciplinary checks on the quality of questions, having a larger pool of interested people from which to derive input on to assessments, and enhanced professional development of staff to participate in the assessment workshops.

The re-marking of the MEQ examination, which strictly followed the templates provided, resulted in a consistent grading some 3% below the official grading by discipline experts. Two of our markers were discipline experts (in Surgery and O&G). One of these experts marked consistently lower in his own discipline's questions than the original marker whereas the other expert marked the same as the original marker. In both cases our experts marked strictly to the marking template. The extent to which this represents individual differences (hawk–dove effect) and differences in the quality of the marking templates for the two disciplines is uncertain. What is certain, however, is that the marking templates overall were not sufficiently comprehensive to permit the use of markers who are not experts in the discipline. This observation has important ramifications in a resource-constrained environment where it is becoming increasingly difficult to identify discipline-based experts who are available to mark papers in a timely fashion.

Ambiguous wording and incomplete marking schemes accounted for 28% and 56% of the item-writing and marking scheme flaws, respectively, for the MEQ. This was despite the fact that the MEQ examination had undergone a discipline-based modified Angoff standard-setting process earlier in the same year of the examination, part of which included specific analysis of the proposed questions for the flaws mentioned. Thus, despite the best efforts of our faculty, some flaws in the MEQ remained undetected until this study. We cannot determine whether this observation is generalisable or specific to our faculty. We speculate on the basis that a significant number of flaws were also detected in MCQ questions derived from the AMC (and which underwent an extensive quality control process independent of our faculty) that this is a generalisable observation.

In terms of feedback to compilers of these types of examination questions, once these types of flaws can be identified and defined, more effective questions can be produced. It would be possible, for example, to use the IWF and modified Bloom's criteria that we have used in this study in the evaluation of proposed questions. Clearly, questions or marking templates with IWFs would need to be rewritten. It would be possible to categorise questions as types I–III using the modified Bloom's criteria and to use such a categorisation to select a desired taxonomic mix for a particular examination. One difficulty with this approach is the resource implications. Our analysis was laborious and time-consuming and we are sceptical that we could summon sufficient enthusiasm from all the clinical disciplines that need to be involved in this approach if it was to become a routine part of the process of developing questions.

The effects of IWFs on the credibility of an examination have previously been highlighted and suggested as acting in favour of the poorly performing student (Tarrant & Ware 2008). In situations where there is uncertainty in the marking due to IWFs, examiners may mark in favour of the candidate in the absence of precise guidelines or in the presence of ambiguity. This raises the philosophical issue of the role of expert or discipline-specific markers in high stakes assessment. Such individuals may compensate for marking templates that might not always cover the range of possible and sometimes unanticipated but appropriate responses to written questions, but may err in favour of the borderline candidate. Perhaps a more appropriate observation would be that with such subjectivity, the MEQ should not be used in the high stakes assessment process, but kept for formative assessment.

Higher order cognitive skills are an important component of clinical competence and the graduating student should be able to show the appropriate skills of interpretation, analysis and judgement in handling clinical problems. This is the crux of 'expert performance' where individuals are able to display coherent knowledge and an ability to solve problems using their understanding of principles and concepts (Gijbels et al. 2005). One of the weaknesses of the assessment process in this study was the undue focus given to recall of knowledge at a level in the course where more attention should have been given to measurement of applied knowledge. The MEQ is purported to test higher order cognitive skills (Stratford & Pierce-Fenn 1985) but this study and those of others (Feletti & Smith 1986) has shown this is not necessarily the case. In this study over three quarters of all of the MEQs tested lower level cognitive skills, and were essentially knowledge testing items. When compared to the MCQs, which tested knowledge in only 47% of the items, the issue of using MEQs in an exit examination leading to internship must be considered.

The correlation between the MEQ and MCQ examination was particularly high, suggesting that the two components of the examination were testing similar traits. This can be contrasted with the correlation between these examinations and the OSCE, which suggests that the OSCE is testing different abilities and competencies, as would be expected from the structure of these examinations. If MEQs are so difficult to write to effectively test higher order skills and so time-consuming and variable in marking, it could be argued that their use in development is not an efficient use of valuable faculty time. The high correlation between the MCQ and MEQ casts further doubt on the utility of the latter assessment in the exit examination of the MBBS programme.

One of the problems we have identified with the MEQ paper is the need for expert markers or, if non-expert markers are employed, to have much more comprehensive marking templates. To produce a tight template with no ambiguity, no errors, no vagueness and to include all the possible answers, authors are more likely to create a Bloom level 1 question, rather than one that tests the higher cognitive skills of synthesis and analysis. In practice, it is difficult for experts to predict all the valid possible answers to questions that test higher order thinking. If this is accepted, then it does challenge the notion that MEQs can be written that are reliable and test higher order thinking. This raises the question of whether or not we are capable of testing higher order thinking in a reliable and reproducible way. MCQs provide a potential option (our data offer limited support to that observation), but perhaps more resources should be targeted at more clinically realistic assessments such as OSCEs and ward-based assessments such as the mini-CEX (Norcini et al. 2003) or newer types of assessment such as the Script Concordance test (Charlin et al. 2000).

Much of the criticism of MCQs revolves around the assertion that they can only test knowledge and much of the appreciation for MEQs relies around the assumption that they test higher order thinking. There is little doubt that IWFs in MCQs can have an effect on student performance (Harasym et al. 1998). A flawed question is more likely to help the weak student whereas the tough and well-structured question will be difficult for all – the poor students being given a leg-up in the overall examination through better performance in the poorly structured questions. This is not to ignore the effects of IWFs on the credibility of an examination, such as those illustrated by recent research (Tarrant & Ware 2008), which indicated that IWF's favoured poorer performing students. It is of concern that 36% of MCQs in this study contained IWFs, but this is less than that reported in other studies (Jozefowicz et al. 2002; Stagnaro-Green & Downing 2006).

Most universities and examining bodies still believe in the principle of an exit examination and for credentialing bodies such as the AMC and the Medical Council of Canada this is the only feasible means of ensuring an appropriate standard of clinical competence. To do this, the examination must meet the curriculum blueprint, have sound validity and fidelity, be able to delivered and marked in an uncomplicated manner as possible and must be able to withstand the rigours of potential legal challenge. Our data would suggest an examination based on the MEQ might not be able stand up to such critical review.

We have shown that for this exit examination the whole performed better than the parts, as was expected. What was unexpected was that the MEQ performed poorly in relation to its primary purpose, and in relation to our MCQ, despite a substantial effort on the part of our faculty to produce a quality MEQ examination. This suggests that formats such as the MEQ should be reserved for low-stakes processes or formative assessment. Given the practical difficulties of engaging a sufficient number of discipline-based experts in the development and marking of the MEQ we conclude that the MEQ is not worth the effort.

The University of Adelaide's MBBS programme has since dropped the MEQ paper from its exit examination and is evaluating in its place the Script Concordance test.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the article.

All authors contributed to the manuscript by analysing and interpreting data and contributing and approving the manuscript.

Notes on contributors

EDWARD J. PALMER Lecturer in assessment and evaluation at the University of Adelaide, Australia.

PAUL DUGGAN Chair of the Assessment Committee of the MBBS Program and Senior Lecturer in Obstetrics and Gynaecology, The University of Adelaide and the Royal Adelaide Hospital.

PETER G. DEVITT Head, Professorial Surgical Unit, Royal Adelaide Hospital.

ROHAN RUSSELL Intern at the Royal Adelaide Hospital.

References

Australian Medical Council 2009. Australian Medical Council. Retrieved July 2009.

Bloom BS. 1956. Taxonomy of educational objectives, Handbook I: Cognitive domain. New York: David McKay.

Charlin B, Roy L, Brailovsky C, Goulet F, Van der Vleuten C. 2000. The script concordance test: A tool to assess the reflective clinician. Teach Learn Med 12(4):189–195.

Epstein RM. 2007. Assessment in medical education. N Engl J Med 356(4):387–396.

Feletti GI, Smith EKM. 1986. Modified essay questions: Are they worth the effort? Med Educ 20(2):126–132.

Ferguson KJ. 2006. Beyond multiple-choice questions: Using case-based learning patient questions to assess clinical reasoning. Med Educ 40(11):1143–1143.

Gijbels D, Dochy F, Van den Bossche P, Segers M. 2005. Effects of problem-based learning: A meta-analysis from the angle of assessment. Rev Educ Res 75(1):27–61.

Harasym PH, Leong EJ, Violato C, Brant R, Lorscheider FL. 1998. Cuing effect of "All of the above" on the reliability and validity of multiple-choice test items. Eval Health Prof 21(1):120–133.

Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. 2002. The quality of in-house medical school examinations. Acad Med 77:156–161.

Lockie C, McAleer S, Mulholland H, Neighbour R, Tombleson P. 1990. Modified essay question (MEQ) paper: Perestroika. Occasional paper (Royal College of General Practitioners) 46:18–22.

Norcini JJ, Blank LL, Duffy FD, Fortna GS. 2003. The mini-CEX: A method for assessing clinical skills. Am Coll Physicians 138:476–481.

Palmer E, Devitt P. 2006. Constructing multiple choice questions as a method for learning. Ann Acad Med Singapore 35(9):604–608.

Palmer EJ, Devitt PG. 2007. Assessment of higher order cognitive skills in undergraduate education: Modified essay or multiple choice questions? Research paper. BMC Med Educ 7(1):49.

Schwartz RW, Witzke DB, Donnelly MB, Stratton T, Blue AV, Sloan DA. 1998. Assessing residents' clinical performance: Cumulative results of a four-year study with the objective structured clinical examination. Surgery 124(2):307–312.

Stagnaro-Green AS, Downing SM. 2006. Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. Med Teach 28(6):566–568.

Stratford P, Pierce-Fenn H. 1985. Modified essay question. Phys Ther 65(7):1075–1079.

Tarrant M, Ware J. 2008. Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. Med Educ 42(2):198–206.

Wass V, Van der Vleuten C, Shatzer J, Jones R. 2001. Assessment of clinical competence. Lancet 357(9260):945–949.