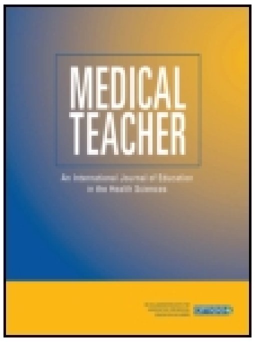

Medical Teacher

ISSN: 0142-159X (Print) 1466-187X (Online) Journal homepage: http://www.tandfonline.com/loi/imte20

Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations

James Ware & Torstein Vik

To cite this article: James Ware & Torstein Vik (2009) Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations, Medical Teacher, 31:3, 238-243, DOI: 10.1080/01421590802155597

To link to this article: https://doi.org/10.1080/01421590802155597

Published online: 03 Jul 2009.

2009; 31: 238–243 MEDICAL TEACHER

Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations

JAMES WARE[1] & TORSTEIN VIK[2]

[1]Health Sciences Centre, Kuwait University, Kuwait, [2]Norwegian University of Science and Technology (NTNU), Norway

Abstract

Background: One Norwegian medical school introduced A-type MCQs (best one of five) to replace more traditional assessment formats (e.g. essays) in an undergraduate medical curriculum. Quality assurance criteria were introduced to measure the success of the intervention.

Method: Data collection from the first four year-end examinations included item analysis, frequency of item writing flaws (IWF) and proportion of items testing at a higher cognitive level (K2). All examinations were reviewed before after delivery and no items were removed.

Results: Overall pass rates were similar to previous cohorts examined with traditional assessment formats. Across 389 items, the proportion of items with ≥5% of candidates marking two or more functioning distracters was ≥47.5%. Removal of items with high *p*-values (≥85%), this item distracter proportion became >75%. With each successive year in the curriculum the proportion of K2 items used rose steadily to almost 50%. 31/389 (7%) items had IWFs. 65% items had a discriminatory power, ≥0.15.

Conclusions: Five item quality criteria are recommended: (1) adherence to an in-house style, (2) item proportion testing at K2 level, (3) functioning distracter proportion, (4) overall discrimination ratio and (5) IWF frequency.

Introduction

Multiple choice questions (MCQ) in high stakes examinations are characterized by high validity and reliability if appropriately constructed. Moreover, MCQs may be useful tests of a candidate's knowledge base and what they can do with it. However, several reports show that in-house high stakes examinations do not achieve the required high quality without training by skilled item writers and those knowledgeable in the principles of assessment (Jozefowicz et al. 2002). When it was decided to introduce the MCQ format for undergraduate medical assessment at NTNU a series of workshops were organised to support the process.

It was also felt appropriate to look for markers of excellence and build them into the training process. Susan Case and David Swanson's NBME web monograph gives useful guidelines and tips how this goal might be achieved (Case & Swanson 2004). Other sources include a series of papers written by Haladyna and Downing (1989), who have also produced evidence suggesting that item writing flaws (IWFs) can prejudice the outcome of high stakes examinations (Haladyna & Downing 1989; Downing 2005). It is an important area of educational research that has stimulated further examination in the health sciences (Tarrant & Ware 2008).

The investment that any school makes when introducing most forms of educational change is great, both in terms of Faculty time spent and the costs of setting up workshops. Little information is available about markers for quality assurance of the process. This report briefly outlines the whole process from workshop, through production to outcome with the results. It is proposed that a series of criteria would be of value to not only increase accountability but also serve as incremental goals by which improvement can be measured and objective feedback based upon.

Practice points

- MCQs benefit from a consistent style.
- Assessment quality assurance is measurable.
- Criteria of quality should be used.
- Review against set Q&A criteria.

NTNU chose to introduce A-type, best one of five, multiple choice questions, to replace traditional, labour intensive and less reliable constructed response item formats like essays. The additional use of vignettes was considered important to raise the level of cognition tested (Case & Swanson 2004).

Methods

All Norwegian medical schools have a 6-year curriculum, however, there is no national licensing examination, and each school has their specific educational model. At NTNU, an integrated, problem based learning curriculum was introduced

Correspondence: J. Ware, Director of Medical Education, Centre of Medical Education, Faculty of Medicine, Health Sciences Centre, PO Box 24923, Safat, 13110 Kuwait. Tel: +965 7854153; fax: +965 531 8454; e-mail: jamesw@hsc.edu.kw

ISSN 0142–159X print/ISSN 1466–187X online/09/030238–6 © 2009 Informa Healthcare Ltd.
DOI: 10.1080/01421590802155597

in 1993, with one summative, integrated (short stations and written) examination at the end of each year. Students, who fail the yearly examination on a second attempt, have to redo the year. The examinations are aligned according to pre-specified learning objectives. Learning outcomes have been specified and published (Hegstad et al. 2004), and were based upon amendments of the Scottish doctor (http://www.scottishdoctor.org/) and the Danish 'The Future specialist' documents (Fremtidens speciallege 2000). As result of a minor revision of the curriculum in 2004, it was decided to extend written examinations to comprise both modified essays (MEQs), which were the only type of written examination, and A-type MCQs.

In order to support the introduction of the MCQ examinations five workshops were run in 2004 and 2005 to support the strategy. Over 70 Faculty staff and students attended the workshops. The format was similar for each workshop with a summary of selected response item format theory and research, a review of the known consequences of item writing flaws, a presentation of NTNU's in-house style, Appendix 1, followed by several sessions given to item writing, review and critiquing. Particular emphasis was put on recognising items testing lower levels of cognition, K1 (recall and comprehension) and higher levels, K2 (application and reasoning). This is a modification of a proposal made by Irwin and Bamber (1982) and also accords with the classification used by the IDEAL Consortium (Prideaux & Gordon 2002). An arbitrary goal was set for NTNU examinations being delivered with at least 50% K2 items.

As new questions were submitted for inclusion in the year-end summative examinations, they were reviewed by the examination committees for each of the first four years of the 6-year programme: curricular alignment, clarity, style, accuracy and elimination of the more common IWFs were the main focus. Appendix 2 gives a brief motivation for the IWFs sought for exclusion (Tarrant et al. 2006).

Following the delivery of each MCQ paper the results were computed without negative marking (Downing 2003). Item analysis was carried out using the IDEAL software (vers. 4.1). The *post hoc* reviews confirming the results were done with the aid of these data. The item statistics are based on classical test theory and the output is presented in the format shown in Table 1, (Osterlind 1998). Also available on the output sheet is the mean group result, variance, SD, Kuder–Richardson Reliability and SE of Measurement. Data have been used both from the performance data of each item and also the whole test, particularly *p*-values and upper-lower item discrimination based on the top and bottom 27% of candidates. The frequency of candidates marking each option was also noted and where ≥5% of candidates marked a distracter it was determined to be functional. A further analysis was carried out after the removal of items with *p*-values ≥85%.

The KR-20 was used as the internal consistency reliability coefficient, while the test mean result was used as a general comparator of class performance against historical data (not shown).

Arbitrary levels of discrimination were used to create ranges which reflect three levels of the discrimination power: >0.40, excellent; 0.30–0.39, good and 0.15–0.29, moderate. Below 0.15 was considered as having no discrimination power of significance.

The End of Year Examinations consisted of an MCQ and an MEQ paper with a University standard of 65%. The overall pass rates were the combined results of the MEQ and MCQ paper with equal weighting.

The two authors undertook a post-test review of the level of cognition tested by each item and classified the IWFs remaining in the examinations. Finally, TV gave feedback to the exam committee. Feedback was also obtained from members of each examination committee and student representatives at meetings held after each examination.

Table 1. An example of the item analysis output for one item from a first year summative examination. There were one hundred items and 120 candidates. See below for annotation of abbreviations.

Item 4	DIF = 0.583, RBIS = 0.384 RPB = 0.304, CRBIS = 0.322 CRPB = 0.255 (95% CON = −0.160–0.894) IRI = 0.150								
Group	N	INV	NF	OMIT	A**	B	C	D	E
Total	120	0	0	0	0.58	0.13	0.16	0.13	
High	32	0			0.75	0.03	0.16	0.06	
Mid	55	0			0.62	0.09	0.16	0.13	
Low	33	0			0.36	0.27	0.15	0.21	
Test score mean %:					74	62	73	66	
Discrimination power					0.39	−0.24	0.00	−0.15	
SE of discrimination power					0.12	0.09	0.09	0.09	

Notes: DIF:Difficulty/easiness/facility or *p*-value, proportion of candidates with correctly keyed option.
RPB: Point-biserial correlation between item and total test score.
CRPB: Corrected point-biserial, correlation as above with present item removed from total test score.
RPBIS: Biserial correlation between item and total test score.
CRPBIS: Corrected biserial as above.
IRI: Item reliability index, point-biserial correlation times the root of DIF times one minus DIF, reflecting item difficulty variance.
N: Number of candidates.
INV: Number of candidates not making a valid marking for this item.
NF: Number of candidates not finishing the test from this point forwards.
OMIT: Number of candidates omitting this item.
HIGH: Top 27% of candidates in the whole test.
LOW: Bottom 27% of candidates in the whole test.
SE of DP: Standard error of measurement.
**: Correctly keyed option.

Results have been presented as group means, some with 1SD and proportions are represented as percentages.

Results

Table 2 shows the overall results of the 4 Year-end Examinations. There was little difference in the passing rates across the 4 years compared with historical data. For these in-house examinations the range of KR-20s was 0.75–0.85, with the average error of measurement being 3.5%. Table 3 shows that almost 65% of all items had a significant Discrimination Power and almost one third were in the good or excellent ranges.

Table 4 shows the proportions of functioning distracters. About one quarter of items had one or more non-functioning distracters, 77% had one or more functional distractors and 47.5% had two or more. In the example shown in Table 1, option E was non-functional and options B, C and D were functional.

When the items with *p*-values ≥85% were removed and the proportion of the remaining items used to determine the overall proportion of functional distracters, this became 76.2 ± 5.8%. This result is to be expected due to the removal of items with high *p*-values.

The authors found that the first two End-Year Examinations had a significantly low proportion of K2 items, only the latter 2 years nearly meeting the criterion set, Table 2. The item review for IWFs found totally 31/379 that were significant and should have been expected to be picked up by the review committees; however, this was only 8% of the total items. The longest option was the commonest IWF (55%) and the four others were: word repeats in vignette and correct option (2/31), logical clues (4/31), sentence completions (6/31) and a negatively worded question (2/31).

Feedback from students suggested that they found their attendance at the workshops, an important part of the acceptance that was reported to be general and the change in assessment was not found to be threatening or unduly stressful. The students' own views were considered more influential in achieving this acceptance than the information given by seminars held by faculty staff. Staff had not experienced what is often complained of by many Faculties that item construction of MCQs was a tedious and onerous task. At the time of the workshops there were still those opposed to the change with

Table 2. Results of the first 4-year-end examinations in the 6-year undergraduate medical programme. Overall pass and fail decisions for the year-end examination based on equal weighting of MCQ and MEQ papers.

Exam year	1	2	3	4	Totals/mean ± 1SD
No of items	100	97	92	100	389
No of candidates	120	102	82	93	397
Mean diff. (*p*-value)	0.71	0.71	0.70	0.76	0.72 ± 2.71
KR-20*	0.85	0.79	0.82	0.75	0.80 ± 4.27
Proportion of K2 items	28%	22%	49%	46%	35%
Proportion MCQ passes	78%	85%	91%	91%	86%
Overall failure per year	16%	8%	9%	0%	9%

Note: *Kuder-Richardson 20, internal consistency reliability coefficient.

Table 3. Proportions of items in the three discrimination categories plus those below. 65% of all items have a discrimination power ≥0.15 across all 4-year-end examinations.

Exam year					
Discrimination power	1	2	3	4	Mean ± 1SD
≥0.40	21.0%	15.2%	11.3%	8.90%	14.1% ± 5.29%
0.30–0.39	19.0%	17.4%	15.5%	10.0%	15.5% ± 3.92%
0.15–0.29	38.0%	32.6%	37.1%	32.2%	35.0% ± 2.99%
0.00–0.14	15.0%	21.7%	27.8%	35.6%	25.0% ± 8.67%
≤0.00	7.0%	13.0%	8.2%	13.3%	10.4% ± 3.26%
Total	100%	100%	100%	100%	

Table 4. Proportions of items with functioning distractors and ≥5% of students marking each option. 47.5% had two or more functioning distracters ≥5% level.

Exam year					
No. functioning distractors	1	2	3	4	Mean ± 1SD
0	15.0%	23.7%	26.1%	27%	23.0% ± 5.5%
1	35.0%	35.1%	27.1%	22%	29.8% ± 6.4%
2	30.0%	26.8%	28.3%	34%	29.8% ± 3.1%
3	18.0%	10.3%	15.2%	15%	14.6% ± 3.2%
4	2.0%	4.1%	3.3%	2%	2.9% ± 1.0%
Total	100%	100%	100%	100%	

concerns that selected response item formats merely created superficial learning strategies. After the examinations this view had diminuished substantially.

Discussion

The introduction of the new assessment format was a smooth transition and the goal of producing items of acceptable quality seems to have been met. Although achieving very similar pass-fail results as previous editions of these examinations is not absolute confirmation for the validity of the assessments based on a new test format, it does represent some of the evidence needed.

The only criterion set, and not achieved, was to use at least 50% of K2 items in all four Year-end Examinations; although, the latter two Year-end Examinations were getting close to this goal. Despite this failure the declaration of the goals was probably useful. In this respect, five criteria can be highlighted: adherence to an in-house style, the frequency of items with a range of discrimination considered to serve their purpose, a proportion of items with sufficient functional distractors to make a valid test and a low proportion of items with IWFs. The fifth criterion was meeting the 50% proportion of K2 items.

These criteria alone are not sufficient to create high quality examinations. Institutional high stakes' tests shall be a valid sample of the curriculum and the items themselves should reflect the outcomes or learning objectives set by the curriculum planners. Determination of success of this is a matter of judgment and it matters little what the quantitative data show if the test has little relationship to the course of study.

In contrast to national certifying examinations, institutional assessments reflect the curriculum as stated by the outcomes and objectives. The assessments should be aligned with these thereby making the assessment process criterion referenced. This is an important qualifier when the proportion of items with *p*-values $\geq$85% is considered, 156/389, 40%. These results would be considered typical of criterion referenced examinations where the number of items with significant discrimination falls as student performance produces more items with higher *p*-values.

There still remains much discussion about what constitutes an item writing violation, with available empiric data being rather few (Haladyna & Downing 1989). Notwithstanding the controversies, it still remains reasonable to set rules for an institution, and we believe the list given in Appendix 2 are worth avoiding until such time as we know that any inclusion does not affect the test outcome. This becomes more important when the guessing factor, inherent in any selected response item format test, is accounted for. At NTNU negative marking was not used and there is good evidence for avoiding such a strategy (Downing 2003).

The five criteria for a high stakes end of course (viz., graduation) or year-end examinations we would recommend are the following:

1. Strong adherence to an in-house style: for NTNU see Appendix 1.
2. The proportion of K2 items is at or above 50%.
3. Greater than or equal to 50% of all distracters shall be functioning at the 5% level.
4. Greater than or equal to 60% of items shall have moderate or better discrimination using set ranges.
5. The frequency of IWFs agreed for the institution shall be <10%.

Jozefowicz and coworkers stresses the importance of item writing training, having found that this substantially enhances item quality as judged by peer review (Jozefowicz et al. 2002). They set no quantitative criteria to be met, but had some very experienced assessment panellists to assist with their item reviews. These will not always be available to most institutions and some other method for quality control may have to be sought. But the use of quantifiable data should not be a substitute for peer review by experienced Faculty teachers. The data from item analysis are invaluable tools and should always be followed by a structured discussion. NTNU chose to use five options, although there is evidence that four or even three option MCQs function as well (Haladyna & Downing 1993). Whatever number chosen, and this may be a quite arbitrary decision, an important part of quality assurance is to determine that the number of options that function justifies the number set as a policy. We believe that after removing items with *p*-values $\geq$0.85 the desirable functional distracter proportion should be >50%.

The influential Maastricht school (Schuwirth & Van der Vleuten 2003) stresses that item format is not the arbiter of cognitive level tested, but rather the content. Therefore, MCQs are useful tests of a candidate's knowledge base and what they can do with it. But, using vignettes does not guarantee an item testing at K2. Although it is unsurprising that those items used at NTNU with images and diagrams were all classified as K2 items. We support the school of thought that a K2 item must always tap into a candidates knowledge base, and provided the test samples widely teachers will not be better informed about their students by using K1 items.

Although there will be a direct relationship between item discrimination and the number of functioning distracters (Haladyna & Downing 1993), we believe that both shall be included as criteria of quality. The former is often reported as a marker of quality for a test but seldom the latter. While separately the use of five options must be subjected to a test of justification for the decision. Although maybe more important is to ensure that all the options fall on the same continuum, and are plausible. These are matters of judgment, and success may not be quantifiable.

In conclusion, we believe that the costly decisions of educational change must be an accountable process and we have presented an attempt to achieve that goal.

Acknowledgements

The item analysis programme, originally developed by various personnel at the University of Alberta and British Columbia, was redesigned into a Windows-based application specifically for use by the IDEAL system. Daniel Precht was one of the original authors at University of Alberta and assisted this work under contract with funding from the Hong Kong WebMED Project. This project received funding from the University Grants Committee, Hong Kong Government.

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

Notes on contributors

JAMES WARE was formerly an academic surgeon, now whole time working in the field of medical education.

TORSTEIN VIK, as former Vice Dean for Medical Education was responsible for the introduction of new examination formats at Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

References

Case S, Swanson DW. 2004. Item writing manual: Constructing written test questions for the basic and clinical sciences, National Board of Medical Examiners publications, retrieved in 2004, from: http://www.nbme.org/aboutitem/writing.asp

Downing SM. 2003. Guessing on selected-response examinations. Med Educ 37:670–671.

Downing SM. 2005. The effects of violating standard item writing principles on tests and students: The consequences of using flawed items on achievement examinations in medical education. Adv Health Sci Educ 10:133–143.

Fremtidens speciallæge. 2000. Betenkning fram speciallægekommissionen. København: Sundhedsministeriet, pp 178–179.

Haladyna TM, Downing SM. 1989. Validity of taxonomy of multiple-choice item writing rules. Appl Meas Educ 2:51–78.

Haladyna TM, Downing SM. 1993. How many options is enough for a multiple-choice test item? Educ. Psychol Measure 53:999–1010.

Hegstad AC, Materstvedt LJ, Kaasa S. 2004. Undervisning i medisinsk etikk: Trondheims-modellen (Education in medical ethics: The Trondheim Model). Tidsskr Nor Lægeforen (J Norw Assoc) 124:2104–2106.

Irwin WG, Bamber JH. 1982. The cognitive structure of the modified essay question. Med Educ 16:326–331.

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

Osterlind SJ. 1998. Constructing test items. 2nd ed. Boston: Kluwer Academic Publishers.

Prideaux D, Gordon J. 2002. Can global co-operation enhance quality in medical education? Some lessons from an international assessment consortium in medical education. Med Educ 37:404–405.

Schuwirth LW, Van der Vleuten CP. 2003. ABC of learning and traching in medicine: Written Assessment. BMJ 326:643–645.

Tarrant M, Ware J. 2008. The impact of item writing flaws in multiple-choice questions on student achievment in highy-stakes nursing assessments. Med Educ 43:198–206.

Tarrant M, Kneirim A, Hayes SK, Ware J. 2006. The frequency of item writing flaws in multiple-choice questions in high stakes nursing examinations. Nurse Educ Today 26:662–671.

Appendix 1

A 45 year-old smoker develops an attack of acute bronchitis following surgery for his inguinal hernia. The intern treats him with a course of oral amoxicillin. After 3 days the patient developed abdominal cramps, diarrhoea and a fever of 38.2°C. The registrar carried out a sigmoidoscopy and found a markedly erythematous and ulcerated mucosa with a fibrinous covering membrane which was easily rubbed off.

What would be the treatment of choice?

A. Oral metronidazole
B. Intravenous steroids
C. Oral trimethoprim
D. Surgical resection
E. Complete bowel rest

Note the following: The vignette has all the information required to answer the question, which cannot be answered without using this information; the question is clearly stated in a separate paragraph and the options are short and of equal length.

Appendix 2

IWFs to be avoided

1. Grammatical clues, found when using sentence completions. The option with an incorrect grammatical flow is automatically eliminated by most candidates
2. Logical clues, based on information in the stem also being used in the correct keyed option. Test wise candidates are quick to spot this flaw.
3. Words repeat, where the stem has a complete or part of a word that is clearly identified in the correct keyed option.
4. Convergence cues, usually based on multiple facts used in the options. The good candidate quickly adds up these facts and finds the correct option having most repeaters in it. Or, where more than two options deal with similar areas to the exclusion of others, which are the distracters and then serve little purpose.
5. The longest option is the correct keyed option because of the number of qualifying statements added to justify it as the best choice.
6. Lost sequence in presentation of data, failure to use ranges and mixed units, as well as overlapping data, or no normal values given. All these flaws add to the uncertainty and, therefore, become confusing.
7. Use of absolute terms such as never, always, only etc which are seldom appropriate qualifiers for clinical statements and the option is eliminated by a good candidate.
8. Use of vague terms such a frequently, occasionally or rarely (among others) which then cause uncertainty and are usually eliminated as being fillers.
9. Use of negative(s) in the question. These items are frequently misunderstood as one is not expecting the formulation to be in the negative. Alternatively, the

correct option is so implausible so that it shall not apply under any circumstance.

10. Use of EXCEPT in the stem as part of the question formulation. Although seldom confuses, these items identify the correct keyed option as often being out of sequence with the others without the use of any knowledge.
11. The use of none or all of the above (NOTA or AOTA) as the last option. Writing options that fulfill these absolutes: NOTA, often provide clues; while AOTA rewards partial information.
12. Failure to pass the Hand Cover Test (HCT) increases uncertainty about the question being asked, or leaves the examinee guessing.
13. Unclear language, ambiguities, gratuitous information, vignette not required etc.
14. Use of interpreted data. Not infrequently a complex vignette is followed by a reference to the condition, disease or diagnosis followed by a question which requires no reference to the information given in the vignette, only knowledge of the condition.
15. Inaccurate information, including implausible options.