A confidence and safety approach to multiple-choice question scoring

M.J. Tweed[1], M. Thompson-Fawcett[2], P. Schwartz[2] & T.J. Wilkinson[3]

Abstract

Aims: This study explored the effect of a scoring system that incorporates a rating of the candidates' confidence in their responses and the safety of the chosen responses in practice.

Background: Multiple-choice question (MCQ) tests with number-correct scoring are widely used in high-stakes examinations. Concerns about number-correct scoring include the implication that guessing is acceptable practice and uncertainty whether an incorrect response reflecting an unsafe decision in practice is a true belief or a random guess.

Methods: A scoring system was developed in which responses included descriptors for levels of confidence, as defined by the likelihood of contacting a colleague/reference. Incorrect responses were also reviewed in advance for the level of safety in practice. An optional MCQ paper was then offered to medical students at the University of Otago. A randomised cross-over design with four versions of the paper was used. Each paper had two sets of 12 questions in different order, one set to be answered with number-correct scoring instructions and the other with the new system instructions.

Results: The survey was completed by 372 students. Score reliability was sufficient, and the scoring system instructions and year of the student both led to differences in responses. Students gave fewer unsafe responses when scored using the safety and confidence instructions.

Discussion: Making students aware of potential unsafe responses, and allowing them to admit a lack of confidence, had the effect of reducing answers that reflected unsafe decisions, without decreasing the number of correct responses.

Keywords: MCQ, confidence, safety.

1 University of Otago Wellington
2 University of Otago Dunedin
3 University of Otago Christchurch

Correspondence
Dr Mike Tweed
Associate Dean Medical Education
Medical Education Unit
University of Otago Wellington
PO Box 7343
Wellington 6242
New Zealand
Tel: +64 4 3855541
Email: mike.tweed@otago.ac.nz

Introduction

Multiple-choice-questions (MCQ) are widely used in healthcare professional high-stakes examinations. Number-correct scoring is a frequently used scoring system which credits the candidate with a positive mark for every question answered correctly. Whilst the content of such examinations may reflect clinical practice, a possible consequence of number-correct scoring is that it may encourage guessing. Concerns over guessing include the invalid raising of a score but, more importantly, that guessing when uncertain is not good clinical practice (Tweed 2006).

In addition to being incorrect, distractor responses from a selection list may reflect unsafe or dangerous practice within the context of the question. If a candidate does give an unsafe response, it may be: a guess because the scoring system has no adverse consequences for incorrect, unsafe answers; a guess because of the risk-taking personality of the candidate; or incorrect knowledge held with excessive confidence. Each of these, if applied in clinical practice, would not be desirable. Incorrect knowledge held with over-confidence may lead to incorrect patient care. Self-improvement in practice is desirable for clinicians, but over-confidence leads to a failure to search for contradictory information and even to disregard that which becomes apparent (Garb 1986).

The development of a scoring system that can incorporate safety and confidence would be useful (Dory et al. 2010; Tweed 2006). This study investigated the responses of undergraduate medical students to questions with instructions that incorporate confidence and safety.

Method

We used a randomised cross-over design.

Examination questions combined into papers and randomisation

Twenty-four questions were selected from the University of Otago Faculty of Medicine extended matching MCQ item bank. Each question consisted of a clinical scenario stem, a question task and ten possible responses, with the tenth response always being "I don't know and would seek advice". In their responses, the candidates were instructed to use "don't know" rather than not answering at all. Although taken from the extended matching MCQ item bank, each item stem and response list was used only once, and there was only one correct response per question.

There were two sets of 12 questions (A and B), and four papers were prepared. Each paper had all of the same questions, but the question group order and the nature of the instructions and their order were varied. Order, instructions and content were balanced between papers (Table 1).

A paper of 24 questions was given to each candidate. These were taken from a pile containing all four papers in a random order.

Examination candidates

The potential candidates were undergraduate medical students at the University of Otago. The medical degree course at the University of Otago is a six year course with a first-year common to all health sciences, followed by an Early Learning in Medicine section (Years 2-3), an Advanced Learning in Medicine section (Years 4-5) and

a Trainee Intern Year. This formative MCQ exam was offered to all students in Years 2-5. Invitations to participate and information sheets were sent to all students, and all who agreed to participate gave written consent.

Safety and confidence scoring

Safety scoring used the judgment of an expert panel to grade the safeness of the distractor responses of the MCQ items. For each incorrect response, the individual expert specified the response as being not unsafe, low unsafe, moderately unsafe or highly unsafe. The median of all the expert opinions was used to define the level of unsafeness of each incorrect question response. This was undertaken before the candidates sat the examination. In this process, the expert panel also reviewed the questions and set the standard for exit in Year 5 and for number-correct scoring, using a modification of the Angoff method (Livingston & Zieky 1982).

Confidence-based (certainty-based) scoring incorporates the confidence of a candidate into the response and scoring rubric (Gardner-Medwin 2006). The positive and negative weightings of scores are constructed so that a negative score for an incorrect response is given greater weighting than the equivalent positive score for the same degree of confidence. The descriptors used for the candidate's level of confidence were defined by perceived level of knowledge and the readiness to ask for assistance. In addition, there was a "don't know" option representing no confidence.

The confidence and safety scoring system was presented to the participants in the form of a table that was included with the question instructions (Table 2).

Students received feedback on the number of their responses that were correct, incorrect, "don't know" and unsafe. They also received a number-correct score equivalent to exit the Year 5 standard.

Analysis

Score reliability was calculated using Cronbach's α (Cronbach 1990) for all question responses as if marked using number-correct scoring.

It was envisaged that any effect of the order of presentation of the instructions on the outcomes (number of correct, incorrect, unsafe and "don't know" responses) would be reduced by the randomised approach with balanced papers. The effect of the order of the papers was analysed using independent sample t-test in SPSS (SPSS Inc., Illinois).

The effects of the year of the student and scoring instructions on the outcomes (number of correct, incorrect, unsafe and "don't know" responses) were analysed by repeated measures ANOVA in SPSS. This took account of both within subject (scoring instructions) and between subject (year) effects. Interaction terms were tested to see if the effect of instruction depended on the student's year. Pairwise differences in means between individual years (i.e. whether scores were different by year, on average across both instructions) were not explicitly tested using post-hoc tests as these were not a focus of the investigation.

Results

Three hundred seventy-two students participated in the study (39% of the total classes). Of the 8976 question responses, 3896 (43.6%) were correct, 4183 (46.9%) were incorrect and 849 (9.5%) were answered "don't know" (Table 3). Reliability was α=0.80.

Question order did have an effect on responses. Comparing the 12 questions the candidates got first with the 12 questions answered second showed no difference in the mean number answered correctly (5.3 v 5.2, p=0.81) or incorrectly (5.2 v 6.1, p=0.73) or number that were unsafe (2.9 v 3.0, p=0.34), but there was a significant difference in "don't know" responses with fewer in the group answered second (1.5 v 0.7, p<0.001). This pattern was seen for both sets of instructions, suggesting that any differences resulting from the order of presentation of the groups of questions were the result of the order and not the instructions.

The year of the candidate and type of instructions both led to differences in responses (Table 3). Compared with number-correct scoring instructions, safety and confidence scoring instructions led to fewer unsafe responses (F (1,368)=5.3, p=0.022), with the same number of correct (F (1,368)=2.5, p=0.11), incorrect (F (1,368)=0.96, p=0.33) and "don't know" (F (1,368)=0.38, p=0.54) responses. The higher the year, the more correct, fewer incorrect, fewer "don't know" and fewer unsafe responses (all Fs (3,368)>35, all p<0.001).

For each outcome variable, an interaction term was included in the model to test whether the effect of instruction type differed across student years; none of these interactions were significant (all Fs (3, 368)<1.3, all p>0. 1). This suggests that the effects of instruction did not differ across years.

Discussion

Both scoring system instructions and student year affected responses. Safety and confidence scoring instructions led to fewer unsafe responses. As expected, increasing year group levels led to more correct, fewer incorrect, fewer "don't know" and fewer unsafe responses.

Unlike the results from an earlier study (Tweed & Wilkinson 2009), there was no evidence of an interaction between year and instructions. In that study, the effect was not seen in the more advanced students, and it was postulated that the proximity of high-stakes examination for these students may have had an effect. This study builds on that previous work, by sampling an increased number of participants and more year groups, and incorporates students' confidence, or willingness to seek advice, in the responses.

Dichotomous scoring systems encourage guessing. A candidate with partial knowledge must either guess based on this partial knowledge or respond "don't know" without benefit to their score. Other scoring schemes have been tried to recognise partial knowledge, including partial credit scoring (Frary 1989; Masters 1988) and subset selection scoring (Frary 1989; Jaradat & Sawaged 1986). Scoring systems that recognise dangerous and unsafe responses have also been developed (Mankin, Lloyd & Rovinelli 1987; Slogoff & Hughes 1987). The advantage of the current scoring system over these

is that it combines recognition of partial knowledge, clinical uncertainty and safeness.

A strength of this study is the randomised methodology. This should reduce the bias caused by order of questions or order of scoring instructions. Many year groups were represented and as expected the responses changed with increasing year, confirming construct validity of the questions. Another strength is that the descriptors of confidence were linked to practice in terms of contacting a colleague or referring to relevant literature, making this authentic to reality. The distractors were reviewed by a panel of practicing clinicians making the decisions about safety authentic to practice. Objective safety measures based on actual data of patient outcomes may be possible (Tweed 2006), but it is unrealistic for all distracters, and limiting the number of distractors to those where safety data are available might reduce the number of distractors to a level that is impractical.

There are limitations to this study. It may not be authentic in terms of the results produced. All the participants were aware that this was not a summative assessment and was part of a research project. Their responses may have been different if this was a high-stakes examination or they were in medical practice. Question selection may not be representative as there were many more unsafe response options in this study than the average in the question bank. Although the instructions did cause the participants to alter their responses, the effect size was smaller than that between consecutive year cohorts.

The greatest value of this system may be in formative-only assessments. Just the presence of the instructions and awareness of safety implications can have an effect on responses (Tweed & Wilkinson 2009). Knowledge of the number of responses that a candidate has for each level of confidence and safety may be helpful in developing their future learning, examination responses and perhaps even practice. In addition, it discourages guessing without regard for consequences. This benefit should not be confined to undergraduate medicine but be realised in other healthcare professionals and post-graduates.

Even though the scoring rubrics, as originally described (Gardner-Medwin 2006), were used as examples to the participants, they generate scores for each candidate for each level of confidence and safety. As such, these scores cannot be aggregated to make summative decisions. There are two reasons for this: firstly, the size of the penalty for completion error, a candidate filling in an unsafe response by mistake with high confidence when meaning to mark a different one; and more importantly, although it is accepted that compensation between different attributes may not be valid, compensation within attributes such as knowledge-ignorance, safety-danger and confidence-doubt may particularly threaten validity (Tweed 2010). Instead, a count of responses in various categories with requirements needed in each category to pass may be more appropriate (such as the suggested scheme shown in Table 4). This recognises the importance of correct knowledge that is held with appropriate confidence. It also means that correct responses cannot be used to compensate

for unsafe responses especially when held with high confidence. Patient safety and appropriate confidence in decisions are important aspects of medical practice.

In this study, a substantial number of incorrect responses appeared to have some degree of unsafeness, and this occurred across all years. As mentioned, this may not reflect responses in high-stakes assessment or actual practice. Further exploration of this may include investigation of the interactions of knowledge, confidence and safety.

There are several avenues for future work. Issues of safety and confidence are important to all healthcare professionals, and this scoring system should be developed and evaluated for assessment of those candidates. Further exploration of how to use this information summatively to make decisions on progression for individual students would add to its utility. A particular area of future interest would be to determine if the subsequent practice of candidates who hold unsafe information with high levels of confidence differs from that of other candidates.

Ethical approval

This study was approved in accordance with policies and procedures of the Ethics Committee of the University of Otago regarding research involving human subjects.

Acknowledgement

We wish to acknowledge advice on statistical analysis from James Stanley, Biostatistician, University of Otago Wellington.

Table 1: Development of papers

	First 12 questions		Second 12 questions	
	Question group	Instruction	Question group	Instruction
Paper 1	A	Confidence and safety	B	Number correct
Paper 2	B	Number correct	A	Confidence and safety
Paper 3	A	Number correct	B	Confidence and safety
Paper 4	B	Confidence and safety	A	Number correct

Four papers were prepared such that order, content and instructions were evenly distributed.

Table 2: Scoring system related to confidence and safety.

		Confidence			
		None	**Low**	**Moderate**	**High**
Confidence descriptors	Related to knowledge	The candidate has no idea of correct response and any response would be a guess.	The candidate has no clear idea of correct response but has some knowledge on the subject. Any response would be based on limited information.	The candidate has a reasonable idea of correct response on a basis of moderate knowledge on the subject. Any response would be based on sufficient information	The candidate is certain of correct response on a basis of detailed knowledge on the subject. Any response would not be a guess.
	Related to practice	The candidate would need to consult a colleague or references prior to considering any response.	The candidate would need to consult a colleague or references but would be able to give a response first.	The candidate would need to consult a colleague or references to confirm the correctness of the response.	The candidate would have no need to consult a colleague or reference.
Correct		-	+1	+2	+3
Incorrect	Not unsafe	-	0	-2	-6
	Low unsafe	-	-1	-4	-10
	Moderately unsafe	-	-2	-6	-14
	Highly unsafe	-	-3	-8	-18
"Don't know"		0	-	-	-

Table 3: Percentage of responses by year group and instructions

Year	Scoring instructions	Correct	Incorrect	"Don't know"	Any unsafe
2	Safety and confidence	23.32	59.93	16.76	31.65
	Number correct	28.10	55.50	16.40	31.47
	Total	25.71	57.71	16.58	31.56
3	Safety and confidence	32.16	54.08	13.77	26.99
	Number correct	35.24	53.89	10.87	30.62
	Total	33.70	53.99	12.32	28.80
4	Safety and confidence	47.44	45.35	7.21	21.96
	Number correct	45.99	44.71	9.29	24.36
	Total	46.71	45.03	8.25	23.16
5	Safety and confidence	61.38	34.95	3.67	16.29
	Number correct	62.31	35.14	2.55	18.16
	Total	61.85	35.04	3.11	17.23
All	Safety and confidence	42.59	47.45	9.97	23.61
	Number correct	44.69	46.26	9.05	25.47
	Total	43.64	46.85	9.51	24.54

Compared with number-correct scoring instructions, safety and confidence scoring instructions led to fewer unsafe responses, with the same number of correct, incorrect and "don't know" responses. Increasing year led to more correct, fewer incorrect, fewer "don't know" and fewer unsafe.

Table 4: Combined confidence and safety scoring used to make summative decisions

		Confidence that response is correct			
		Nil	**Low**	**Moderate**	**High**
Correct		-	a	a	b
Incorrect	Not unsafe	-	c	c	c
	Low unsafe	-	c	c	d
	Moderately unsafe	-	c	d	e
	Highly unsafe	-	d	d	e
Don't know		c	-	-	-

By considering the combination of safety and confidence, a risk table can be produced.

To pass the assessment, these four criteria need to be met:

- The total number of correct responses (total of cells with a's and b) has to be above a threshold.
- The total number of correct responses held with high confidence (cell b) has to be above a threshold.
- The total number of unsafe responses held with confidence, significant in combination, (cells with d's and e's) has to be below a threshold.
- The total number of moderately or highly unsafe responses held with high confidence (cells with e's) has to be below a threshold.

All responses in cells with c's would be useful for feedback without contributing to a summative decision.

References

Cronbach LJ (1990) *Essentials of Psychological Testing.* New York, Harper Collins.

Dory V, Degryse J, Roex A, Vanpee D (2010) Usable Knowledge, Hazardous Ignorance: Beyond the Percentage Correct Score. *Medical Teacher* 32: 375-380.

Frary RB (1989) Partial-credit Scoring Methods for Multiple-choice Tests. *Ethics & Behavior* 2: 79-96.

Garb HN (1986) The Appropriateness of Confidence Ratings in Clinical Judgment. *Journal of Clinical Psychology* 42: 190-197.

Gardner-Medwin AR (2006) Confidence-based Marking: Towards Deeper Learning and Better Exams, in: C Bryan & K Clegg (Eds) *Innovative Assessment in Higher Education.* London, Taylor & Francis.

Jaradat D, Sawaged S (1986) The Subset Selection Technique for Multiple-choice Tests: An Empirical Inquiry. *Journal of Educational Measurement* 23: 369-376.

Livingston SA, Zieky MJ (1982) *Passing Scores: A Manual for Setting Standards of Performance on Educational and Occupational Tests.* Princeton, NJ, Educational Testing Service.

Mankin HJ, Lloyd JS, Rovinelli RJ (1987) Pilot Study Using Dangerous Answers as Scoring Technique on Certifying Examinations. *Academic Medicine* 62: 621.

Masters GN (1988) The Analysis of Partial Credit Scoring. *Ethics & Behavior* 1: 279-297.

Slogoff S, Hughes FP (1987) Validity of Scoring 'Dangerous Answers' on a Written Certification Examination. *Academic Medicine* 62: 625.

Tweed M (2006) Negative Marking Can Be Justified in Marking Schemes for Healthcare Professional Examinations. *Medical Teacher* 28: 579-580.

Tweed M (2010) Passing Assessments Should Not Just Be Jumping Hurdles. *Focus on Health Professional Education* 11: 86-90.

Tweed M, Wilkinson T (2009) A Randomized Controlled Trial Comparing Instructions regarding Unsafe Response Options in a MCQ Examination. *Medical Teacher* 31: 51-54.