Reviewer's report

Title: Rarely selected distractors in high stakes medical multiple-choice examinations and their recognition by item authors: A survey and simulation

Version: 1 Date: 20 July 2010

Reviewer: James Ware

Reviewer's report:

The paper is another trying to justify the reduction to four or three option multiple-choice questions as opposed to the usual five options favoured in health sciences. The authors first identify the least chosen multiple-choice item distractors based on the annual Swiss Federal Graduation Examinations and then study the effects of reassignment by random allocation and purposeful allocation to the correct keyed option. A survey was also reported determining if experts could also identify the least selected item distractor and then correlate their results with the statistical data.

It is assumed that this information would be considered of value to institutional readers and not medical licensing bodies. However, there are major differences between a final graduation examination delivered by a federal authority and what local institutions are faced with. Not least it has to be assumed that weak students have long since been removed from any medical programme by the time they reach their final graduation year. It also seems by using the word graduation in the title of the exam the local institutions may use the results as a surrogate for their own graduation examinations.

Methods

Three years of examinations provide the frequencies of functioning and non-functioning distractors (NF-D). However, many would argue that insufficient evidence was presented for defining a NF-D as one chosen by less than 1% of candidates rather than using only distractors chosen by <5% and or with positive discrimination (less robust). It was unclear whether there were five separate discipline examinations or one single paper with 250 items for 2005.

Having analysed functioning distractor frequencies from three years of annual examinations, inexplicably only the first year’s data were used for the rest of the paper. This should be explained.

There are five components representing the major disciplines/subjects in the graduation examination. However, and once again without explanation, only one discipline with 55 items were allocated to 37 experts divided into two panels reviewing a maximum of eighteen items only.

The attractiveness of distractors to the less able candidates was highlighted using delta medians and the NF-Ds defined by both <5% and <1% combined to illustrate selections made by specific proportions of the candidates. Once again, except to increase the numbers, the reason for this was not explained, Fig 1. The
data as such seem sound.
The outcomes measures are the same as those reported many times before. To then justify the paper as needed in the health sciences based on their paucity, without stating who needs the information is an omission. However, it does not matter what subject is used when we all use theoretical and mathematical manipulations to prove the same results. If this had been an experimental study there might have been more justification.

Discussion and conclusions
The discussion and conclusions were reasonable but for a federal graduating examination this paper seemed to omit two important issues (a) what happened to the pass-fail rates (if all five components were separate, 15 could have been calculated) and also (b) no mention was made of the very high mean test scores which led to the very low functioning distractor frequencies, as one would have expected. Limitations were stated.

Formalities
Language, title, abstract and use of references were all as expected and correct.

Recommendations
The justification to include the definition of >1% as evidence for distractor functionality based on failure rates of about 5% is not acceptable and as this is a cornerstone of the paper better and more compelling evidence must be submitted. It is obvious that functioning distractor frequencies follows a sigmoid curve fit and as the mean test scores were so high in the data presented, the frequency would always be low.

If the authors cannot do this, then only the <5% rule (empirical, but does at least have historical references) should apply.

The survey of the experts does not justify to being included, unless it can be expanded to include much larger numbers.

Having calculated a range for score change, surely what the reader needs to know is (a) would this affect the pass-fail rates and (b) specifically what would this do for borderline or weak testing examinees?

Judgment
An article of limited interest.

**Level of interest:** An article of limited interest

**Quality of written English:** Acceptable