Reviewer's report

Title: An assessment of functioning and non-functioning distractors in multiple-choice questions

Version: 1 Date: 14 April 2009

Reviewer: David Swanson

Reviewer's report:

Overall, this is a useful, clearly written paper reporting on the measurement characteristics of teacher-developed tests in the health professions. However, many of the conclusions drawn do not really follow from the results presented.

Major Compulsory Revisions

1. The authors appear to take the point of view that there is an ideal number of options for MCQs and that all items should have that same number of options. It would be good if they could explain why this is the case and why the number of options should not vary depending on the content area, skill to be tested, level of the learner, and the logical number of options given the stem (eg, 12 may be a good number of options for items related to cranial nerve abnormalities).

2. The authors should describe the process used to develop the seven tests that were studied. In particular, if the items had been used on previous exams and item statistics from those exams influenced the decision to reuse items, there may be some selection bias that affects results to an unknown degree. The authors may wish to calculate statistics similar to those reported separately for items with and without prior use.

3. The decision to drop items with item difficulties greater than 0.9 from the dataset has a major impact on results, and it is unclear why this is appropriate. It may be better to leave these items in the dataset in calculating the reported statistics, though this will clearly increase the proportion of items with multiple non-functioning distractors. Some topic areas are straightforward and well-learned; if they are important, that does not mean they should not be covered on tests. This seems like it should be an educational decision, not a psychometric one – there is no reason not to use easy items testing important topics. Further, it may not even be appropriate to rewrite non-functioning distractors: in some instances, this will increase the sophistication required to answer a question, making it too specialized and educationally inappropriate for the learners for whom the test is developed. There can also be good substantive reasons for including non-functioning distractors: for example, in an item describing a patient with a benign breast lump, including various benign and malignant causes as distractors in the option list may be appropriate, even if the malignant causes are non-functioning, because the benign/malignant differentiation is an important one educationally and clinically.
4. If all items included only three options as the authors recommend, the proportion of items with item difficulties over 0.9 would increase and this is ignored in the paper. This should be particularly true if item authors selected distractors without pretesting. The fact that items rarely have more than one or two functioning distractors does not mean that item authors will be good at identifying which distractors those are. It is an empirical issue what happens to item characteristics if authors only include three options when they write items, and, at this point, the authors have no information about this; there may not even be much time saved in writing items with fewer distractors if most author time is spent thinking of functioning distractors.

5. Dropping a distractor because of a positive discrimination index is particularly problematic if the option is, in fact, incorrect. Estimates of the value of this index will not be very stable with the sample sizes used in the study: if the item were given again, the sign on this index would reverse for a reasonable proportion of items for which it is near zero. (The same is true, but to a lesser extent, for the proportions picking the correct answer and each distractor: these values are also sample dependent.) This should at least be acknowledged as a limitation somewhere in the manuscript.

Minor Essential Revisions

6. It would also be good for the authors to report additional information about how the calculations underlying the correlations in Table 3. Were the values in the table used directly in calculations or were the number of functioning distractors correlated with item difficulties with items as the unit of analysis? Given the definition of non-functioning distractors, it is somewhat counter-intuitive that the correlations are so low.

Discretionary Revisions

7. The authors may wish to discuss the school's use of 50% as a passing score for all tests. Intuitively, this seems like a questionable policy: passing scores should be set in light of test difficulty, not despite test difficulty. Decisions about reuse of items with difficulties over 0.9 become very important with such a policy in place.

Level of interest: An article whose findings are important to those with closely related research interests

Quality of written English: Acceptable

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests:

I declare that I have no competing interests.