Author's response to reviews

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

Authors:

Marie Tarrant (tarrantm@hku.hk)
James Ware (jamesw@hsc.edu.kw)
Ahmed M Mohammed (ahmedm@hsc.edu.kw)

Version: 2 Date: 15 May 2009

Author's response to reviews: see over
Dr. Marie Tarrant  
Assistant Professor, Dept. of Nursing Studies  
University of Hong Kong  
4/F, William M. W. Mong Block  
Li Ka Shing Faculty of Medicine  
21 Sassoon Road, Hong Kong  

May 15, 2009

Professor Melissa Norton  
Editor in Chief, BMC Medical Education

Re: An assessment of functioning and non-functioning distractors in multiple-choice questions  
(MS 6733148332571344)

Dear Professor Norton:

Further to the editorial decision on the above cited manuscript, we have made the recommended revisions and are resubmitting the revised manuscript for your consideration. We would like to thank the reviewers for their thoughtful and helpful comments and hope that we have adequately addressed their concerns. Below you will find the detailed responses to the reviewers’ comments and we have highlighted the edited and added passages in the manuscript. The revisions have been read and approved for publication by all authors. Should you have any further queries, please feel free to contact me at tarrantm@hku.hk.

Sincerely,

Marie Tarrant RN MPH PhD
Response to Reviewers

Reviewer #1 (DS):

Major Compulsory Revisions

R1-1. 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).

Response: We thank the reviewer for pointing out our overzealousness in advocating 3-option items and agree that the number of options can vary with the content. We have edited our discussion of this point in the introduction (p.2, lines 14-16), discussion (p.8, lines 16-22 & p. 9, lines 1-10) and the conclusion (p.14, lines 11-19) to include the point made by the reviewer and to further clarify the recommendation for the use of 3-option items under most circumstances.

R1-2. 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.

Response: Further information about the test development process has been added to the Methods section (p.3, lines 18-22). While we agree that it would be useful to identify items which have been used on previous exams and compare the item statistics, we do not have the data available to do this. We have added further discussion on this point to the limitations section (p.13, lines 14-16).

R1-3. 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.

Response: We thank the reviewer for providing his input on this issue. This was discussed and debated among the authors during the data analysis. The reason for removing items with difficulties >.90 was to be consistent with previously published literature on this topic [1]. However, we agree that including all items presents a more accurate analysis of non-functioning distractors and we therefore have reanalyzed the data leaving these items in. The results and discussion have been amended to reflect the updated data analysis.
R1-4. 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.

Response: This possibility has been further discussed along with the relevant research on the discussion section (p. 10, lines 11-22).

R1-5. 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.

Response: We have added this point to the limitations section of the discussion (p. 13, lines 16-20).

Minor Essential Revisions

R1-6. 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.

Response: Further information has been added to the data analysis section to clarify the statistical computations (p.5, lines 12-20).

Discretionary Revisions

R1-7. 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.

Response: Further discussion of this point has been added to the discussion section (p.9, lines 11-19).
Reviewer #2 (SMS):

Major compulsory revisions:
R2-1. The number of examinees per test for 6 out of the 7 tests sampled was <100 (range 73-75). I think this is too small a sample to give an accurate estimation of discrimination power (D) of an item option, when using only the upper 27% and lower 27% (i.e. only 54% of the total number) of examinees for calculating the D value. Have you tried using the discrimination coefficient instead of discrimination index? The advantage of using the discrimination coefficient instead of discrimination index is that with the first method each and every person evaluated is taken into account, while with the second, only 54% of the total number of examinees are considered.

Response: We have now reanalyzed the data using the point-biserial correlation coefficient instead of the discrimination index as recommended. The methods, results and discussion have been amended to reflect the updated data analysis.

R2-2. What was the content of the test papers examined in this study? Were they discipline-based, multidisciplinary or fully integrated tests? This may have a bearing on the predictability of a test item for the whole test performance.

Response: The curriculum in the Bachelor of Nursing programme is a credit-based curriculum. Therefore, the content of all of the test papers was discipline specific. In this analysis there were no comprehensive or multi-disciplinary papers. A brief note has been added to the Methods section to highlight this (p.4, line 1).

R2-3. Although the data demonstrated that in these tests there was a low frequency of test items with >2 functioning distracters per item, it’s perhaps over simplistic to draw the conclusion that “ideally multiple choice item should consist of three options: one correct answer and two plausible distractors.” I read in here the assumption that by reducing the number of options per MCQ item from 4 to 3, there will naturally be an equivalent reduction in the proportion of non-functioning distractors. This is not necessarily true as there is no guarantee that an otherwise untrained/unskilled examiner will naturally construct a test item with 2 effective/functioning distractors just because of a reduction in the number of options needed per test item.

Response: Further discussion has been added on these points. Please see the response to points R1-1 and R1-4 above.

R2-4. The decrease from 4-options to 3-options per test time appeared to selectively have favoured the low ability examinees without helping the high ability examinees, albeit the change was small and insignificant statistically. Nevertheless the message is still clear that reducing the number of options helps the lower ability students to pass the examination. It is believed that larger number of options gives the lower ability students more opportunity to show what they don’t know. The question is whether in a test, we should only look for what examinees know, or we should also be interested in what they don’t know!

Response: We thank the reviewer for making this point. Actually, we don’t know what the impact of reducing the number of options would be for high-ability students as that was not examined. In Table 5, we simulate what would happen to pass rates when the number of options was reduced. As was noted in the results and discussion sections, there is a small and insignificant increase in the number of students who would pass the three-option test. As well, Table 6 shows that the overall mean test score increases
slightly with three-option items. This increase is likely shared among all the examinees and not just lower ability students. As most researchers have pointed out, the increase in pass rates that occurs when items are reduced to three options is not substantial or significant enough to outweigh many of the other psychometric advantages of three-option items. Additionally, the three-option test may actually be a more accurate reflection of student achievement than the four-option test [2].

Minor essential revisions:
Table 2:
R2-1. Please check the value in the parentheses of “Total” column (23.6) for distractors with both frequency <5% and discrimination >0. It looks larger than expected.

Response: The value should have been 13.2. However, most of the values in this table have been amended in light of the revised data analysis.

References: