Reviewer’s report

Title: Assessment of higher order cognitive skills in undergraduate education: Modified Essay or Multiple Choice Questions?: Research Paper

Version: 2 Date: 10 August 2007

Reviewer: Jon Veloski

Reviewer’s report:

Comments on Assessment of higher order …Modified Essay
8/8/2007

The authors have done an excellent job of revising the earlier manuscript. The conclusions are now very relevant and clearly-stated. This will make an important contribution to the literature by reminding readers that these two item formats can be used to measure higher order skills. Alternatively, either format can be misused. However, the MCQ has certain practical advantages, as the writers point out.

The Abstract is improved. I, personally, would still like to see more detail in the abstract – number of items reviewed, number of examinees, number of raters, etc., but this is really up to the authors and the Editor. The present abstract is certainly informative. (Discretionary)

The research question outlined at the end of the introduction is much clearer.

The substitution of Kappa for Rho is a substantial improvement in the paper. The results are now much more convincing!

Bottom of page 4. The statement that “The significance of these values...” at the bottom of the page does not make sense. Kappas are interpreted by the values – the values of .70 to .80 are quite reasonable. The statistical test reported by some computer programs is a test of the hypothesis that the Kappa is equal to zero. I would recommend deleting this sentence. Or, if there is some reason for including it make it clearer what it tells the reader. (Major – delete the sentence or please expand to explain what is being reported.)

Top of page 6. “At a minimum...” This is a small point, but I believe you mean content validity, not face validity. You refer nicely to content validity in the Conclusion. This is the foundation of test development – do experts judge the content of the test to be valid. On the other hand, face validity is a more informal criterion – does the average person on the street or layman judge the test to be valid. Most testing professionals feel that face validity is of little or no value. I would suggest that you refer to content validity here in order to be consistent with your conclusion, which is excellent. (Minor)

(Discretionary) Bottom of page 7. You discuss extended matching and uncued
items, and suggest that they focus exclusively on knowledge recall. I would suggest that you reword this slightly to say that these formats can be misused by test faculty members because these formats can readily be used to test knowledge recall. Like all formats, these two can be adapted to measure higher-order skills. In particular, at some point you might want to take a closer look at Veloski et al (1993), which addressed the cueing effects of MCQs. The abstract is included here:

Although the cueing effects inherent in conventional multiple choice questions (MCQs) present serious limitations, this format continues to dominate testing programmes. The present study was undertaken to estimate the effects of cueing when MCQs are used to test medical students, and to evaluate the reliability, validity and feasibility of an alternative testing format. Equivalent items in both MCQ and open-ended, or uncued (Un-Q), formats were administered to 34 third- and fourth-year medical students. The students’ mean % correct score on the MCQs was 11 percentage points higher than their mean level of performance on equivalent Un-Qs. When a second set of more difficult items was administered to 16 of these students, their mean performance on the MCQ items was 22 percentage points higher than their performance on equivalent Un-Qs. The results support the feasibility of large group administration of tests constructed in an open-ended format that can be scored by computer. Not only is this format equally reliable and economical when compared with the MCQ, but it also provides important advantages that strengthen its face validity. The Un-Q format can be used to test either simple recall or certain higher level problem-solving skills that cannot be tested by MCQs. Even more important, the results also suggest that the Un-Q format may be a more effective discriminator of academically marginal examinees.

This study was replicated in a larger study of over 7,000 resident physicians.


(Minor) The title on Figure 1 might be expanded at the Editor’s discretion. Figure 2 is not clear. Title might be something like “Number of MEQs Within Average of Ratings Assigned by Two Raters on Bloom’s Taxonomy”

(Discretionary) I’m not an expert on content classification systems. However, as I recall from my reading it is generally referred to as Bloom’s Taxonomy. Obviously, this is a matter of editorial preference.

In summary, this is a careful revision that is much clearer and will make an important contribution to the literature.

What next?: Accept after minor essential revisions
**Level of interest:** An article of importance in its field

**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.