Reviewer’s report

Title: The Modified Essay Question: does it still have an effective role in assessment? Research Paper

Version: 1 Date: 30 April 2007

Reviewer: Jon Veloski

Comments on The Modified Essay Question: Does it still ....

This study addresses the question of which format, MEQ vs. MCQ, is psychometrically most effective for the assessment of clinical problem-solving abilities.

The Abstract is very vague and not very informative. Please consider putting more substance into the Abstract to enable the reader to get a better idea of what was done and what the findings are. Also, please paginate the manuscript because it is then possible for a reviewer to refer to specific page numbers in the review comments.

The research question outlined at the end of the introduction is not clear. How exactly is 'effectiveness' defined? It appears from the Methods section that effectiveness is being defined as faculty members’ judgments of the cognitive level of the items (Bloom’s Taxonomy) and the absence of item-writing flaws. The authors state that the content was balanced in the two forms.

One concern is that an important criterion for the effectiveness of tests is their validity – how well do they measure what they purport to measure? The authors might want to consider the following study that was published in 1989:


The Department of Family Medicine at Jefferson Medical College used the modified essay question as the final examination format for its required third-year clerkship in the 1970s and 1980s. To compare the family medicine modified essay question format with the MCQ format used in the other five required clerkships, examination scores from 2,174 Jefferson graduates (1976-1985) were correlated with scores on the examinations of the National Board of Medical Examiners (NBME), faculty ratings of students' clinical performance in the third-year clerkships, and program directors' ratings on four global areas of postgraduate competence. Grades on the multiple choice examination in internal medicine consistently yielded the highest correlations with NBME scores and with postgraduate ratings of medical knowledge. Performance on the modified essay examination in family medicine had the lowest correlations with these other measures of knowledge. The family medicine scores, however, consistently yielded the highest correlations with students' overall third-year clinical performance and with their postgraduate performance as residents in the areas of data-gathering skills, clinical judgment, and professional attitudes. These results indicate that the modified essay question provides a different and important parameter in the evaluation of the problem-solving abilities and professionalism of medical students.

which provided support for the concurrent validity of the MEQ in relation to performance ratings in clerkships and its predictive validity in relation to performance two years later in postgraduate education programs (residencies). Also, it was a substantial study covering 10 years and a sample of over 2,000 physicians.
In the Results section, inter-rater agreement is analyzed by indices such as Kappa, which corrects for chance agreement. It is not clear why Spearman’s Rho was used, which does not include a correction for chance agreement when there are only 3 options. This is especially critical because the distribution in the figure is biased toward the value of 1. If the two raters were assigning ‘1’ randomly at the rates of about 70% and 50% reported, they would agree about 35% of the time by chance alone! The rate reported needs to be corrected for this high probability of chance agreement. Consider reporting Kappa rather than Rho, or explain why Rho is reported when the convention is to use an index such as Kappa.

The concluding statement ‘...as a viable alternative to other types of written..’ is not supported by the data presented. As discussed in the introduction and Discussion section, there is a broad spectrum of item formats that has been used over many decades of objective testing. This ranges from true/false through MCQ to extended matching, long menu/uncued, short answer, MEQ and essay. The present study evaluated MCQs against MEQs on a few limited criteria, but did not evaluate the broad array of other formats. Please consider moderating this conclusion to reflect the analysis that was actually performed. A complete review of the advantages and disadvantages of all item forms would be complex. For example, at one extreme it has been argued that true/false items are the most cost-effective because they make it possible to achieve reliable scores and cover the broadest content domains (S. M. Downing. True-false, alternate-choice, and multiple-choice items. Educational Measurement: Issues and Practice Fall:27-30, 1992.) At the other extreme there are the standardized testing programs in North America that use essay tests in addition to MCQs because of their essential content validity (SAT and MCAT, to name several). They achieve reliability by using multiple graders, and justify the cost because of the corresponding increase in the content validity of the tests that are used for high-stakes decision-making.

Consider rewriting the Discussion section to summarize more clearly the pros and cons of the different formats.

The tables are not well-presented. Table 1 is probably not necessary, summarize it in the text in a few sentences. Table 2 is not clear. Provide a complete title and more explanation of what it is. Also, might simply summarize in text.

The figure has no legend. It is not clear what is being reported.

In summary, this study reminds readers that it is possible to write MCQs that measure higher order cognitive abilities. You have focused on the criteria of classification by Bloom’s Taxonomy and review for item-writing flaws. However, there are other criteria such as reliability, validity (content, concurrent, predictive) and cost that need to be considered and discussed more carefully in the introduction and Discussion.