Reviewer’s report

Title: Changes in standard of candidates taking the MRCP(UK) Part 1 examination, 1985 to 2002: Analysis of marker questions

Version: 2 Date: 5 May 2005

Reviewer: Deborah Saltman

Reviewer’s report:

General

This is a well-written article in an area of great importance to the development of vocational assessment in medicine. The authors are to be congratulated not only reviewing their data, but also exposing their results to the wider research and clinical communities for scrutiny. Very few Medical Colleges take this step. The capacity to analyse the data in a longitudinal way is fortuitous and yields greater depth to the article.

The paper contains a great deal of complex and elegant statistical analysis which requires an in-depth understanding of measurement tests in educational psychology. Unfortunately, it is the over-emphasis on the analytic modelling that is one of the limitations of the paper and leads to a study that has a mismatch between the stated research questions and the interpretation of the results.

The authors describe the aims of the study as “whether, on average, aggregate performance on items has increased or decreased between the first and second usage.” They have made a good attempt at answering this question which is mainly about outcomes of the examination at two points in time.

The claim that this analysis can provide an answer to their other question “By assessing how such a change relates to the dates when the items were used, one can estimate the changing overall performance of candidates in the Examination” requires a major leap of faith (from analysis of an examination to supposition about a candidate), but preferably a re-analysis of all the data about the examination and a re-write to include some significant literature in the area.

For example:

1. The possibility that the marker questions were unreliable markers due to:

a. Imprecise terms making the questions unintelligible. Holsgrove and Elzubeir examined three published papers for the Part 1 membership examination of the Royal College of Physicians (MRCP 1991). They found imprecise terms on 234 occasions.

Has this issue been addressed? Where some of the marker questions deemed “imprecise”? What impact would “imprecision” have on the two cohorts?

What has been the impact on the way the Board reviews marker questions. For example, with an annual review of marker questions by the Board, could one would expect reduced sensitisation to any “imprecision”?

Reference: Holsgrove G. Elzubeir M. Imprecise terms in UK medical multiple-choice questions: what
b. Norm-referencing of marker questions. Wilson asserts that criterion referenced scores are in no way absolute scores. They are norm-referenced and this is done prior to the test construction process at the item level, and not at the total test level during a specific application of the test. This may impact on their validity as a criterion referencing tool.


The authors should provide some analysis norm-referencing process that they undertook to review the marker questions, as they describe in their conclusion:

“If the marker questions used were out-of-date then that may explain the apparent decline. However, not only is the content of marker questions always approved by MRCP(UK) Part 1 Examining Board before each inclusion in the Examination, but the questions shown in table 1 clearly relate to core conditions and their underlying disease mechanisms, and hence changes cannot be shrugged off as resulting from irrelevant or outmoded questions.”

What is the impact of a prior process of norm-referencing on the criterion referencing process? For example: How are “core conditions” defined? Have the “underlying disease mechanisms” for the core conditions in the marker questions changed? How is the relevance of a “core condition” established – by prevalence, by prior use in an examination?

2. The Use of Criterion Referencing in Medical Education

Recent evidence suggests that the criterion referencing may not always be the most appropriate method of assessing candidates, particularly over time. McHarg and others concluded that norm referencing was preferable to criterion referencing.


What evidence is there that criterion referencing tell us anything about candidates ability and competence rather than performance in an exam? How can the implications derived in the conclusion be grounded in the research output?

In summary, this paper needs a major re-write to include more about the context of the examination, including an appropriate literature review and perhaps a little less about the analyses. As it stands the conclusions are not directly related to the research undertaking. If the statistical analyses remain the focal part of this paper, it would be more suited to an educational psychology journal, rather than a medical/health sciences one.

Thank you very much for the opportunity to review this paper.

-----------------------------------------------------------------------------------

Major Compulsory Revisions (that the author must respond to before a decision on publication can be reached)

**Which journal?**: Appropriate or potentially appropriate for BMC Medicine: an article of outstanding merit and interest in its field

**What next?**: Unable to decide on acceptance or rejection until the authors have responded to the major compulsory revisions
Quality of written English: Acceptable

Statistical review: Yes

Declaration of competing interests:

I declare that I have no competing interests