Reviewer's report

Title: Implementing statistical equating for MRCP(UK) Parts 1 and 2

Version: 2 Date: 1 July 2014

Reviewer: Chris Beauchamp

Reviewer's report:

General Questions (details follow)

1. Is the question posed by the authors well defined?
Yes. The issue is very clearly stated.

2. Are the methods appropriate and well described?
The methods are very appropriate for this type of research.

3. Are the data sound?
The data is sound and contributed significantly to the field because it is based on actual rather than simulated data.

4. Does the manuscript adhere to the relevant standards for reporting and data deposition?
Yes it does.

5. Are the discussion and conclusions well balanced and adequately supported by the data?
Please see detailed notes below. The discussion and conclusions are appropriate from a statistical perspective, but I would have liked to see a bit more attention paid to eligibility and examination policy issues (or at least have these considerations ruled out).

6. Are limitations of the work clearly stated?
Limitations are not provided in this manuscript. The inclusion of such a section is my only essential revision.

7. Do the authors clearly acknowledge any work upon which they are building, both published and unpublished?
The introduction does not contain a significant amount of cited research. However, it is possible that not much research exists in this area.

8. Do the title and abstract accurately convey what has been found?
Yes.

9. Is the writing acceptable?
There were very few typographical, spelling or grammar errors in this manuscript.
This made the manuscript much easier to review.

Overall comments
1. Given the public scrutiny faced by regulators and examination boards, the authors have done an excellent job of exploring a real-world phenomenon using a multi-faceted and sophisticated approach. Any concerns raised by individuals who question the increase in pass rates for non-UK candidates are addressed and nullified in this article.

2. The authors employed a number of statistical methods in the exploration of the issue at hand. Some of these methods are complex. They did an excellent job of describing these methods in a succinct and easy-to-understand manner. As a result, the statistical analyses employed added to the narration rather than providing an additional layer of complexity that could be seen as a distraction by some readers.

3. Undoubtedly, a large amount of data was produced during this analysis. The data reduction techniques used by the author (e.g., tables, figures) succinctly and clearly illustrated the findings in a comprehensible manner.

Minor Essential Revision
4. The limitations to this research are few. However, the authors should include a brief section on limitations. Suggested follow-up research would also be helpful (although the latter is not an essential revision).

Discretionary Revisions

Introduction
5. The introduction lacks citations from other organizations who have introduced statistical equating as a replacement or supplement to the Angoff and/or Hofstee method. Is the observation that the pass rate increases for IMGs unique to the MRCP (UK)? What has occurred for other high-stakes examination programs in the healthcare sector in the UK? What has happened in other countries when this change was made to the medical licensing process (e.g., USMLE, MCC)? If this observation is truly unique to the MRCP (UK), it bears mention. If there is a dearth of previous relevant research, this should be mentioned as well.

6. I assume that the successful completion of Part 1 is a pre-requisite for writing Part 2. However, I did not see this explicitly stated. This could explain the differences in findings between the two examinations.

7. P. 4 lines 153-154: If I read this correctly, the standard setting panels for Part 1 ranged from 5-8 raters, and for Part 2, ranged from 6 to 14 raters. In addition, in some cases, the ratings from the “hawks” and “doves” were removed (p.5 line 199-200). I realize that the point of this article is not to justify the Angoff/Hofstee method that was used, but some consideration should be given to the fact that the small panel sizes (especially for Part 1) were sub-optimal. Many of the improvements seen in the program may not be due to the introduction of statistical equating alone, but may be due to the discontinuation of such a
sub-optimal process. If I have misread this section and the rating panels were larger, then it should be stated more clearly.

8. P. 5 lines 189-190. Unless I’m missing something, there appears to be a small arithmetic error when describing the pass-rate limits. For example, 35% ± 6% (i.e. 29% to 41%).

9. P. 6 lines 232-234. I have no doubt that the Rasch model is appropriate for this analysis; especially in the case of an achievement-based pass/fail assessment. However, it would improve the article if some justification could be provided for the selection of this model (e.g., consistency with prior research, model fit, parsimony).

10. P. 7 line 262. Along the same lines as the previous comment, some rationale for the selection of UK graduates as the reference group would be helpful. I assume it is because it is a “stable” and homogenous group, but this should be stated.

11. P. 7 line 282-283. At this point in the article, the only criterion stated for the selection of anchor items is statistical item performance. The reader is left to wonder why the representativeness of anchor items to test specifications is left out. This is addressed and clarified by the authors later in the article (p. 11 line 411), but should be mentioned here as well.

Method

12. P. 12 lines 473 and 477. My understanding is that lowess and loess curves are interchangeable terms. “Loess” is used on lines 473 and 477, while “Lowess” is used in the figures and also on line 492 of the same page.

13. P. 13 line 512-p. 14 line 554. I realize that it is generally not feasible to correlate examination performance with an external metric of physician performance (e.g., supervisor ratings, patient ratings, patient outcomes). As a result, using the intercorrelations of candidate scores on the three components of the MRCP (UK) is the best available method. Much like previous comments, the article could be improved if the authors provided a rationale for this decision.

There is a small logical inconsistency in this section as well. By adopting the Rasch model, the authors are theorizing that examination performance is based on a single latent trait. On page 14, lines 537-542 the authors claim that the lower correlations between Parts 1 and 2 to the PACES are due to the first two being assessment of knowledge, while the PACES is an assessment of clinical skills. This argument implies a multidimensional model of candidate performance. The solution to this inconsistency can be clarified with a statement that the individual components of the MRCP (UK) are unidimensional, but the larger examination program is multidimensional.

14. P. 14 lines 555-569. The issue of measurement error and repeated examination items is one that has been studied extensively. The weight of evidence suggests that repeating some examination items does not introduce
error. I believe that in this paragraph, the authors were theorizing that increased performance for non-UK may be due to a leak in examination content. This was not the case. However, the mini-hypothesis (“leakage of items is unlikely to be explaining the changes across time shown….”) is not introduced until the last sentence of the paragraph. I would suggest opening the paragraph with the mini-hypothesis and then providing the data to disprove it. (As a tangent, this specific data related to repeated item exposure may warrant a publication of its own)

Discussion

15. P. 17 lines 678-685. In this paragraph, the authors have provided one of the best explanations I’ve seen on the distinction between construct-relevant and construct-irrelevant DIF. The example provided further solidified this explanation. Well done.

16. P. 17 lines 685-699. As noted earlier in the article, it is unusual to see so many items showing DIF. Even more surprising is the significant number of items where non-UK candidates outperform UK candidates. The authors have provided additional details on this (e.g., treatment of conditions which were more common in the west, but are not much rarer). However, I would have liked to see the authors move beyond additional observations and try to posit theories. This is a noteworthy observation that merits more exploration. I am not aware of much research showing DIF favouring international candidates.

17. P. 18 line 730 to p. 19 line 761. This section of the article almost takes an apologetic tone where the authors are attempting to justify why it took five years to resolve and explain this phenomenon. While I’m albeit ignorant of the political pressures associated with this research topic, I’m not sure this level of rationalization is necessary. The reality comes down to 1) you have to determine if this is a random occurrence or a trend and 2) if it’s a trend, you need data before you can study it.

18. P. 19 lines 762-775. This section is harsh and unfair to the Angoff/Hofstee process. First, based on information contained in the introduction section of this article, the Angoff/Hofstee method was not being used according to best practice (e.g., small number of panel members). Second, statistical equating introduces its own error and is highly dependent on a large number of theoretical assumptions. I disagree fundamentally with the assertion that the Angoff/Hofstee process is “merely norm-referencing through the back door.” This is not the case if the methods are used as they should.

For many examination programs, an initial passmark is set using a judgment process (e.g., the Angoff method). This typically occurs at the conclusion of a job analysis process where the operational definition of a “barely passing” candidate is established (or re-established). Then, statistical equating is used to adjust this passmark to compensate for differences in examination forms. If the judgment process is flawed, statistical equating will only propagate this error into the future. As a result, I feel that judgement methods and statistical equating should work synergistically. It is not the case of one over the other. I would recommend that
the authors revisit this paragraph. At most, they can conclude that, in the context of the MRCP (UK), statistical equating is an improvement over the way the Angoff/Hofstee method was being conducted. It is not, however, an indictment of the Angoff/Hofstee method.

Conclusion

19. As mentioned, the authors have done an excellent job of exploring a phenomenon using a multi-faceted and sophisticated approach. However, in this section, I expected to see some consideration given to examination policies, especially those related to examination eligibility. In addition, I expected to see if the characteristics of IMG applicants has changed over time.

20. The reader is left to assume that there were no major policy decisions made to coincide with changes in pass rates for non-UK candidates. For example, were English language requirements for non-English speaking international applicants changed at that time? Did the application processing and prior learning assessment done for IMGs become more stringent? There are a number of factors that could explain increased pass rates for IMGs that are not mentioned in the conclusion of this article. The arguments are limited to those of a statistical nature while alternative policy-based hypotheses are ignored. If it is truly the case that increased pass rates for IMGs occurred without any major policy changes, then this should be mentioned.

21. As mentioned in a few places in the article, non-UK candidates are a heterogeneous group. I assume that examination performance varies tremendously based on the demographics for this group, and that the demographics for this group change over time. In order to study changes in pass rate, we are left to assume that this heterogenous group did not change over time. However, I know that in other countries, there have been changes in pass rates because of changes in international applicants. For example, the source country of the international applicant has a tremendous impact. Applicants from English-speaking countries with a similar medical system would, for example, fare much better on the MRCP (UK). In addition, some countries have introduced bridging programs or additional training to help applicants before they challenge an examination. Was this the case in the UK? Once again, if this was not the case, it should be mentioned rather than ignored. I realize that the focus of this article is not on eligibility or examination policies. However, they cannot be ignored completely.

Level of interest: An article of outstanding merit and interest 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