Author's response to reviews

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

Authors:

I C McManus (i.mcmanus@ucl.ac.uk)
Liliana Chis (Liliana.Chis@mrcpuk.org)
Ray Fox (Ray.Fox2@ggc.scot.nhs.uk)
Derek Waller (derek.waller@suht.swest.nhs.uk)
Peter Tang (PTang@gmc-uk.org)

Version: 3
Date: 20 August 2014

Author's response to reviews: see over
Reviewer's report

We are grateful to the reviewers for their kind and helpful comments, which we believe have improved the manuscript. Detailed responses are below. *Our responses to the comments are in italics.*

**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.

*Response: No response necessary*

2. Are the methods appropriate and well described?

The methods are very appropriate for this type of research.

*Response: No response necessary*

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.

*Response: No response necessary*

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

Yes it does.

*Response: No response necessary*

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).

*Response: See below.*

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.

Response: We have included a section on Limitations in the discussion section. The major limitation is that only two examinations at a single institution have been studied.

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 Do the authors clearly acknowledge any work upon which they are building, both published and unpublished?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.

Response: This is an interesting area as there is a lot of tacit knowledge, mainly from other institutions who have implemented IRT (and one of them helped us to implement it), but there is almost nothing in the conventional scientific literature, at least as far as we are aware, which discusses how to implement IRT equating, and what the issues might be. Nor is there anything that we know of which compares IRT equating with judgmental equating. We have added a paragraph to that effect, in the introduction. Needless to say, if the reviewers are aware of further studies we would be delighted to include them.

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

Yes.

Response: No response necessary

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.

Response: No response necessary

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.

Response: No response necessary

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.

Response: Thank you for that comment, and the one below, which are very much appreciated.
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.

Response: See above.

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).

Response: As mentioned above, we have included a section (albeit fairly brief) on limitations.

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.

Response: As mentioned in a previous response, there does indeed seem to be a dearth of published work on implementing statistical equating, despite informal networks suggesting it is quite common, and we have added a paragraph to that effect. There are many technical accounts

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.

Response: The reviewer is correct and we have now explicitly stated that in the opening paragraph of the Introduction.

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.

Response: The Angoff panels that we used were relatively small, but we note that in the recent paper on Angoffing USMLE, run by NBME, that N ranged from 7 to 11. The review by Brandon (2004) looks at various recommendations, and eventually comes down as “at least 10 and
ideally 15 to 20 judges”. NBME in the Clauser et al JEM paper met the minimal criterion on only 7/18 cases. We wonder how many institutions actually meet the criterion of 20 judges....

Clauser et al’s generalizability analysis suggests that the majority of variance is not between judges but between panels, and almost no institutions use multiple panels for the same items. We dispute that our Angoff panels are ‘sub-optimal’ in that they are seriously different from other major players. That we do not follow Clauser et al and have three panels of 30 judges is indeed true, but we wonder how many institutions meet that criterion. We had added some comments into the Limitations section.

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%).

Response: Ooops, yes, that was wrong, but in a more interesting way. The next line was also in error with its +/- 5%, and in fact Part 1 had 5% and part 2 had 6% and they, but not the ranges, had got swapped around. Now corrected. Thank you for picking it up.

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).

Response: We have added a line and a reference here to justify using the Rasch model.

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.

Response: We have ourselves wondered about this, and while it is probably not appropriate for the main text, there are some interesting questions here. Traditionally exam boards use the nationals of their own country (and NBME did exactly that in the 1990s). However a while ago we realised that were education to change in the UK (and it might by fiat of a single government) then that could change the standard. However international graduates come from a host of countries, and it would be much less likely that all of those governments could simultaneously change their medical education in the same way, potentially making it more stable.

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.

Response: We have added an extra sentence or two at (the old) lines 282-3, so that the two sections are consistent.

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.
Response: Thank you for picking this up. It seems that ‘lowess’ and ‘loess’ curves are not identical but are extremely similar. Wikipedia comments that, “A smooth curve through a set of data points obtained with this statistical technique is called a Loess Curve, particularly when each smoothed value is given by a weighted quadratic least squares regression over the span of values of the y-axis scattergram criterion variable. When each smoothed value is given by a weighted linear least squares regression over the span, this is known as a Lowess curve; however, some authorities treat Lowess and Loess as synonyms”. SPSS actually says that it produces a ‘loess’ curve, and since we used that to produce these figures we have made sure that ‘loess’ is used throughout the manuscript.

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.

Response: We are currently working on looking at predictive validity of MRCP results onto later professional behaviours, but that is a long-term exercise. We have included two references to other exams in which, very unusually, there is a clear relationship between exam performance and professional behaviour (and in one study mortality/morbidity) in patients.

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.

Response: This is an interesting and important point. We do not believe that Part 1, Part 2 or PACES is unidimensional (and it is unlikely even that any knowledge exam is unidimensional – candidates who have done lots of cardiology and less gastroenterology will do better on cardiology questions and vice-versa). However, we are not carrying out a psychological study of the structure of knowledge and its dimensionality. The exam is set as a single whole, physicians are required to have a reasonable working knowledge across all of the ‘ologies’, the exam is blue printed so all ‘-ologies’ are included in particular proportions, and the exam has a single final mark. In practical terms therefore the mark for the exam is unidimensional, and we are interested in the extent to which the items load onto that dimension. We could run an exam in which there are separate pass marks for each ‘ology’, and then we would be doing something rather different (and would presume unidimensionality within ‘ologies’, even if that were unlikely to be true, some candidates knowing more about interventional cardiology, some more about hypertension, etc.). As for as PACES is concerned, we are now breaking it down into seven skills, with a separate pass mark for each, and hence the multidimensionality is being directly recognised. We have not included this in the paper itself, but a joy of the BMC journals is that comments and discussions such as these are available for readers who wish to dig in more deeply.

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)

Response: That is a good suggestion and we have included ‘item leakage’ in the sub-heading, and have also been more explicit in the text about testing that hypothesis.

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.

Response: Thank you – that is very kind. As ever, it is only by struggling to come up with a good example and write it down that one fully understands the issues properly oneself.

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.

Response: You are right that the number of items showing DIF is unusual. That probably reflects DIF more usually being looked for in different sorts of test, such as aptitude tests used in school or college selection, where there is good reason to believe that there is genuine unidimensionality in construct terms, and also very good reasons, educationally, politically and legally, for trying to remove items showing DIF, and hence they get eliminated at an early stage. Medical knowledge is a different sort of construct though, and is inherently multidimensional in terms of learning and acquiring it. How unusual our numbers of DIF items are is difficult to be certain of. Although USMLE IMGs show differential performance on sub-scores (ref 33 in our original MS), information is not provided on individual items, and hence comparison is difficult to make. The Geriatric ABIM exam did find 8-13% which suggests DIF is not that unusual.

We do not wish to make further hypotheses here, interesting though the phenomenon is, as ultimately it is tangential to the major thrust of this paper. We will however be returning to it. An interesting issue, for instance, is whether items with DIF predict later outcomes differently from those without DIF. It might be for instance that high scores on items with DIF on which IMGs do better indicate less exposure to westernised medicine, and later problems with high-tech medicine. Clearly these are interesting questions and complex.

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.
Response: Some of the apologetic tone which the reviewer has correctly detected is correctly perceived. Some of it is also reflecting the current regulatory environment in the UK, which is becoming increasingly concerned with top-down management, with sometimes an excessive attempt from regulatory bodies to wish to intervene too soon because of what they perceive as risks. As you say, the correct scientific approach here was undoubtedly to sit and wait until there was sufficient data to make a reasoned analysis. Those in charge of examinations do not, though, always see it that way, and are worried about external challenges from candidates, politicians, regulators, and so on. It should also be said that the UK has recently had a major judicial review of another postgraduate examination, the MRCGP, with accusations of direct or indirect discrimination against IMG doctors or doctors from ethnic minorities, and that has made many examinations slightly twitchy. That is not discussed here, but in fact is the subject of a different paper which is currently being reviewed by BMC Med Ed.

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.

Response: As mentioned below, in a response to another reviewer, we are not convinced that the number of panel members is atypically small compared with, e.g. NBME exams, and neither are we clear what is the appropriate number. It may well be much larger than the numbers we have used, but were it to consist, as Clauser et al imply, of three panels of 30, then we doubt if any examination would use it.

We agree that judgment methods and statistical equating should be used together (and we still do carry out an Angoff of our Part 1 and Part 2 exams, at intervals of a couple of years, to check that results look acceptable).

As far as “norm-referencing by the back door” is concerned, we have removed the phrase, but have replaced it with a direct quotation with a quote from a 2013 paper by Brian Clauser, whom we have worked with, and whose studies to our eye seem to provide a major challenge to the Angoff method. If the Angoff method is less valid, at least in our hands, then the Clauser work potentially provides an explanation of that, and it must be referred to for the reader to make up their own mind. We have also added, as suggested, “for MRCP(UK)” at appropriate places to indicate that the conclusions may only apply to our exams.

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.

Response: It is extremely hard to know whether IMG applicants are changing over time. IMGs can apply to take MRCP(UK) at any time from anywhere in the world, and they no doubt do so for many reasons. However we know relatively little about them before they apply, only having basic information about age, sex, place of qualification and the like. Gross changes in candidate mix can be detected, but mostly it is demographic. Like many examining boards, MRCP(UK) is not an educational institution and therefore has little ability to collect data on background and progress beyond what we have. For those IMGs who are working in the UK that is different, but the majority of the IMGs are outwith the UK and will never work there. We have little therefore that we can say of use, although we continually to monitor candidate characteristics.

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.

Response: There were no major policy decisions concerning MRCP(UK). In fact there are no English language requirements for the exam, although candidates are advised that it is in English and that an IELTS level of 7 should be a minimum requirement. IMGs who wish to work in the UK must pass the GMC’s PLAB exam, for which there is an IELTS requirement which has been increasing over recent years, but that is only a tiny proportion of the IMGs taking this exam. There is no prior learning requirement beyond holding a recognised registrable qualification in medicine. If the characteristics of IMGs do change then that probably reflects self-selection, rather than anything done by the MRCP itself.

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.

Response: We take the point entirely that this is potentially of interest, but we worry that a) the paper is already long, and b) the primary interest is in the standard-setting process, and not on sociological correlates of examination performance. There may well be long-term secular trends in the characteristics of non-UK graduates in the pool taking MRCP(UK), but that
is unlikely to cause any ‘step-change’ in pass rates (and it is was that apparent step-change which was the primary impetus to the present analyses). That the step change may be a more gradual and more prolonged shift, beginning before the introduction of equating, is strongly suggested by our data, but digging deeper into that is not easy. Not the least of the problems is that IMGs choosing to take MRCP are not a random sample of the graduates from their countries; if that self-selection changes then ability levels may change, but we have no way of being able to measure it, or, to compare and contrast it with changes in the overall quality of the graduates from their medical schools (and there are very many medical schools…). So we hope the reviewer will forgive us if we do not go down this route. We have included a summary of what was said above in the limitations section.

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

Reviewer’s report

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

Version:2 Date:13 July 2014

Reviewer: Mona Nasir

Reviewer's report:

Minor Essential Revisions

1. Limitations of the study should be mentioned

Response: This has been done – see responses to the previous reviewer, and in particular the section on Limitations in the Discussion.

Discretionary Revisions

2. Lines 257 -297 may be moved into the Methods section since these are giving the details of analyses carried out.
Response: Lines 257-297 are not the method for the present study, but are instead a description of the method by which statistical equating has been carried out for the MRCP(UK). It is the method used in the back office, therefore, and not the method used for collecting and analysing the present data, and so we believe it is in an appropriate, introductory, position.

3. The tables and figures may be moved in with the corresponding text for better alignment.

Response: That is not standard BMC format, which wants figures submitted separately, and tables and figure captions at the end of the manuscript.

4. The headings of the tables and figures may be shortened since the details are already presented elsewhere in the text.

Response: That is true but readers tend to find it easier to read a caption with the figure, particularly when it is complex. We have therefore left it but can change it if editors or sub-editors request it.

Level of interest: An article whose findings are important to those with closely related research interests

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.

----

Reviewer's report

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

Version: 2 Date: 1 July 2014

Reviewer: Syeda Ali

Reviewer's report:

Minor Essential Revisions:

1. line 87 "this study will primarily consider...." should be changed to "this study primarily considers...."

Response: Done.
2. line 109-110 a reference is needed for the statement made.

Response: Done

3. line 113 'a just passing candidate' is not the term that is used in defining a 'borderline candidate' as one who is 'minimally competent'. This term may be changed accordingly.

Response: The literature seems to refer variously to the minimally competent, borderline or just passing candidate, and so we have retained the latter, since it is the term we use within the MRCP(UK), but on the first usage also give the other terms.

4. line 118 "Angoffing an exam" is not scientific language and needs to be rephrased

Response: It may be ugly and new, but it is what people actually say. English is an evolving, living language, and the conversion of nouns and names into verbs is common-place (are people not hoovering their floors, googling for books, or whatever?). As for scientific language, science frequently takes everyday words and phrases and uses them as it wants (and see molecular biology for that going on on an everyday basis). When 'goggling' "Angoffing" [and one has to put it in quotes or Google strips off the –ing, we found 92 results, including the Canadian Association of Occupational Therapists, who in March 2013 went Angoffing in Ottawa. http://www.caot.ca/exam/CEC_AR_2013_FNL.pdf; and the Canadian midwives seem to do it as well. We have left it therefore.

5. references are not mentioned in the para starting line 232 on wards while explaining IRT

Response: This has now been done – thank you for pointing out the omission.

6. line 258 I wonder why it is important to say "American Psychometricians" why not "psychometricians" only.

Response: Well they were from America, from a well known institution which does a lot of IRT... It does emphasise as well that there was at that time almost no UK expertise on this topic. We have changed it though.

7. line 335 The sentence is a repetition of line 333

Response: It does indeed – thanks for picking it up. It has been removed.

8. line 351 the percentage needs correction it is 11.7% in the table while 11.1% in text

Response: Again, thanks – the table was correct and the value in the text has been corrected.

9. line 473 mentions 'loess curve' while line 492 mentions 'lowess curve'. are these being used as synonymous or is there a difference as in weighted quadratic least squares regression (loess) and weighted linear least squares regression (lowess). both the figures mention 'Lowess curve'

Response: See response to reviewer 1, who also picked up this point. We in fact had used SPSS's loess function, and we have corrected that throughout the MS.

10. there is some confusion while reading lines 661 to 663. suggest re reading.
Response: This has now been reworded and hopefully makes more sense. Thanks for the close reading and picking it up.

11. in the figures a new term OS has been introduced for Non-UK examinees. I suggest that this term may also be mentioned in the text or removed from the tables & figures.

Response: OS ('Overseas') has now been removed throughout and non-UK used consistently.

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'