Author's response to reviews

Title: Cross-comparison of MRCGP & MRCP(UK) in a database linkage study of 2,284 candidates taking both examinations: Assessment of validity and differential performance by ethnicity

Authors:

Richard Wakeford (RichardWakeford@hotmail.com)
Mei Ling Denny (withden@btinternet.com)
Katarzyna Ludka-Stempien (katarzyna.ludka.10@ucl.ac.uk)
Jane Dacre (j.dacre@ucl.ac.uk)
IC McManus (i.mcmanus@ucl.ac.uk)

Version: 3
Date: 19 August 2014

Author's response to reviews: see over
Responses to reviewers

We thank the editor and reviewers for their helpful and interesting comments. Our responses to each are below in italics.

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

Editorial Comment:

A comparison of performance by the same candidate on two different but related examinations is intrinsically interesting. I have doubts, shared with the referees, that this comparison can legitimately be called a validation since there is no hard evidence that either examination has predictive validity. At best it can be said that both are measuring similar parameters. I am not sure what the main learning points are from this study. The data on ethnicity are certainly important but at present seem to be a side issue within the paper. If the authors could revise the paper to address the referees’ comments it could be reconsidered for publication.

Response: In response to this and other comments below, we have first re-thought the paper and then thoroughly revised it. It is clear that we have two issues to illuminate and discuss, that of validity broadly, and that of differential performance by ethnicity (hereafter DPE), though of course the two are closely linked, the latter effectively concerning validity within different ethnic groups. An important occurrence that has occurred between the original submission of the paper and the present re-write is that the High Court action against the MRCGP CSA, pending when we wrote the paper (and lurking behind it), has been thrown out with costs awarded against the complaining party. This means that we collectively feel able to be more up-front with the DPE issue and indeed agree with the reviewers about its apparently secondary profile in the paper. It should not be a “side issue” (above) and now it is not. The outcome is that the two related issues, of validity and of differential performance, get equal prominence, and there is a major re-write of much of the paper – other than the Methods and Results sections, of course, although they have been tweaked.

We are not sure that we agree with the editor’s strong line on validity. It is a position one can take, but one fears that if taken then there would not be a single item left in any undergraduate curriculum which could be regarded as valid. We have put in an extra paragraph on the topic, which in particular describes two prospective studies which relate examination performance (one using licensing exam, and the other USMLE 2 CK) to actual clinical behaviours and outcomes in practicing doctors. It may be that it is only these assessments which are valid, and all others are invalid, including MRCP and MRCGP, but that seems somewhat unlikely.

Also, could you kindly include all the email address of all the authors in the title page?

Response: We had put these on the BMC website, so had presumed they were available. We have now added them to the MS as well.

----------------------------------------------------------------------------------------------------
Reviewer's report

Title: Cross-validation of MRCGP and MRCP(UK): a database linkage study of 2,284 candidates taking both assessments

Version: 2 Date: 28 February 2014

Reviewer: John C Mclachlan

Reviewer's report:

This paper addresses a particular aspect of post graduate performance – effectively, the correlation of performance of those doctors who undertook both MRCP(UK) and MRCGP assessments in the assessments themselves. The analysis is relatively simple in nature, but appropriate, and generates a surprisingly large amount of interesting information! The paper is well written and makes its central points clearly. It represents a valuable, if specific, contribution to the literature. In my view it is worthy of publication. All my comments are Discretionary.

Response: Thank you! Hopefully these positive views will not diminish after the re-write.

The title uses the term ‘cross validation’. But ‘validation’ is a complex term, with both competing definitions and sub definitions. It could be argued that this is a kind of concurrent validation, although this generally implies that the events take place closely linked in time. From another perspective, it could be argued that this study is not really a ‘validation’ of either set of tests. Perhaps it would be less complicated to title the paper simply: “A database linkage study of 2,284 candidates taking both MRCGP and MRCP(UK) assessments”. But I leave this up to the authors.

Response: The title has changed a) to respond to the above, using the possibly less contentious term ‘cross-comparison’ and b) to place the DPE issue ‘up there’. We agree with the reviewer that this is not really ‘concurrent validation’ (there is a gap of several years), but feel that it is surely some form of validation, albeit if it is not clear what. [And if it is not validation what is it and what would it have to be to be validation?] We have therefore used the term ‘validation’ on its own, but follow the reviewer in saying “a form of concurrent validity” – the validity mavens can make their own minds up exactly what it is!

We have also turned the title right around so that it is now “Cross-comparison of MRCGP & MRCP(UK) in a database linkage study of 2,284 candidates taking both examinations: Assessment of validity and differential performance by ethnicity”. This has the advantage a) that it puts up front what we have done and emphasizes the two exams, their linkage and the number of candidates; and b) then suggests that those data are being used to look at both validity and DPE.

The Abstract mentions the ethnicity findings in the Conclusions, but not previously, although these findings are likely to be of considerable interest to a number of parties. I would include mention of the relevant methodology in the Abstract (Methods), indicate briefly the outcome in Results, and include ethnicity as a key word, for the particular benefit of readers who initially have access only to the abstract and/or computer key word searching.
Response: Ethnicity findings have been included in the abstract and further discussed elsewhere in the paper (Background & Discussion). Title changes are relevant here. Ethnicity and Black and Minority Ethnic have been included as key words. The Abstract was lengthened to make it clearer how ethnicity comes into the analysis, but that took it up to nearly 700 words, and we have now, after quite a lengthy struggle, got it back down to the BMC limit of 350 words. If it is less than perfectly clear that follows the constraints imposed by word length.

In the Abstract (Conclusions), the authors use the phrase “cognate knowledge”. This phrase on its own is sometimes used to describe word matching in language acquisition. The terms ‘cognitive knowledge’ or ‘declarative knowledge’ are alternatives, as would be the phrase ‘knowledge cognate to both assessments’. I can see the original phrase has been used, since in the Conclusions (2) the authors use the phrase “assessing cognate areas of relevance to medicine”, where ‘cognate’ is indeed the most appropriate term.

Response: The points are good ones, and we have followed the reviewer and changed the term in the abstract to, “suggesting they assess knowledge cognate to both assessments.”

The authors state “That they are in fact substantially linked supports the idea that both are assessing cognate areas of relevance to medicine. Of course that alone cannot demonstrate validity, but, as has been emphasised earlier, the demonstration of validity requires information from multiple strands. The correlations shown here certainly support the validity of both MRCGP and MRCP(UK)”. As I read this, the last sentence contradicts the first clause of the previous sentence. This also lead to the major substantive comment I would wish to make about the paper, below.

Response: The correlations support the argument for validity but cannot alone make that argument. It was however confusing, we agree, and we have reworded it. “That they are in fact substantially linked supports the idea that both are assessing cognate areas of relevance to medicine. Of course that alone cannot demonstrate validity, for, as has been emphasised earlier, the argument for validity requires information from multiple strands of evidence. The correlation is however compatible with validity, and the argument for validity would be compromised if such a correlation not present.”

As the authors indicate, the concept of ‘validity’ is complex: as I have said above, it has various definitions, each of which may have sub definitions. To my mind, the most relevant version in medical training environments is that of predictive validity, with the framing “Does performance on this test predict how effectively the candidate will perform in real clinical settings?”. Most medically related assessments are indeed intended to grant or decline candidate access to positions of greater clinical responsibility. From this view point, the article, while interesting and significant, does not advance our knowledge of predictive validity of either test (nor was it intended to, of course). If predictive validity data is subsequently identified for either set of tests, then the findings become even more valuable.

Response: We agree that predictive validity is the most important form of validity (and there is an argument that our data, since MRCGP is taken long after MRCP(UK), is actually predictive validity). However to argue that ONLY predictive validity matters would be to paint oneself into a corner. Do children who go on to study medicine need to do arithmetic at school if there is no direct evidence that those who do better at arithmetic in primary school subsequently make fewer prescribing errors when calculating dosage in infants and children? Accurate prescription requires arithmetic facility, and that is a theoretical issue, not one ultimately justified by predictive validity? The validity of teaching children arithmetic is of a different type from predictive validity.
As a consequence, for my taste the authors are a little too positive about having confirmed the ‘validity’ of the tests. It remains possible that both sets of tests are measuring the same construct but that that construct is not the one we really intend to measure. I would have stated my conclusions in rather more cautious terms with regard to validity alone. However, I appreciate that my view is personal, and possibly idiosyncratic, and I propose this as a consideration for the authors (and the editor) rather than as a requirement.

Response: Thank you for this suggestion. We have downplayed the use of the V-word generally, and made our final conclusion (see abstract) more cautious. Having said that, for postgraduate (and indeed undergraduate and all) examiners we believe there is nothing more important than validity, and that it should be driving all that we do.

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.

Reviewer's report

Title: Cross-validation of MRCGP and MRCP(UK): a database linkage study of 2,284 candidates taking both assessments

Version: 2 Date: 3 February 2014

Reviewer: Richard Hays

Reviewer's report:

I read this paper with interest but was soon disappointed. By the end it reads like a psychometrician’s dream - I think that the statistics are fine (but am not a statistician) and the results are interesting, but I am not sure just what they mean and how they fit into medical education.

Response: We hope that our comprehensive re-write with the clear stating of our dual objectives – and the associated Discussion – will persuade this reviewer of the focus and value of the paper. We hope too that a re-ordering of the material in the Results section will help this. As a rhetorical question, do psychometricians dream? The use of “like a psychometrician’s dream” suggests that perhaps an adjective was missing in the sentence, as the phrase appears to be disparaging….  It must be said that the paper is psychometric, and it addresses some serious and interesting questions psychometrically. We can hardly apologise for that, and neither can we make it go beyond what it is not.
The interesting parts for me are that 1. knowledge performance predicts knowledge performance. This is hardly either new or surprising, but I do understand that the validity of particularly the MRCGP examination may be deemed more valid by such an analysis (as a non UK person I do not understand why the initial MRCP examination is done by so many people as it does not confer any particular advantage). 2. the new CSA may be more valid than the old CSA, although the reliability is still a little disappointing.

**Response:** Our claim is not that the CSA is made ‘more valid’ by our analyses, but that the two assessments as a whole can as a result be seen to be ‘assessing cognate areas of relevance to medicine’ (see other review and responses to it). And, in the UK at least, we know of no other such cross-comparisons of performance in different high stakes assessments. (FYI: As a GP Training Programme Director, RW’s experience was that trainees who possessed – or who during training took – the MRCP were either the badge-collectors, thinking probably correctly that it would make them stand out from others in the jobs market or those who had originally embarked on a hospital medicine career and later changed their minds, for one reason or another.) From an MRCP perspective, MRCP does confer many advantages (and not having it is clearly disadvantageous); it is the entry to specialist training, which presumably is advantageous. Maybe we are also a little old fashioned, but why might it be of no use to know a lot about internal medicine, as determined by a large number of experienced physicians, if one wants to be a physician (or even a good doctor)?

I found the extensive results and discussion about ethnicity to be distracting. This dominates the paper although is not mentioned in the title or the abstract. Indeed I wonder if there is a separate paper somehow pasted into the middle of this one?

**Response:** We have as described in the response to the other reviewer re-titled and re-organised the paper and thus hopefully made clear its dual aims.

The risk of the approach taken is that the paper reads like a statistician’s fishing expedition - great numbers, but poorly related to the context.

**Response:** See next. We also should say that this is far from being a ‘fishing expedition’ in the sense of trawling and dredging through multiple variables with scant regard for significance levels, and the inevitable finding of occasional results which happen to reach some rather low alpha level. If instead it is meant that we have tied our fly knowing precisely what we needed to tease the piscine beast we are interested in to take our hook, then we cannot see the problem. As for the context, it is precisely the context of knowing how different examinations relate to one another, particularly in relation to ethnicity. It doesn’t pretend to do anything else, but it surely does well what it does do.

Mandatory modifications: The authors should make clear just what the paper is about (its purpose) and present the results and discussion relevant to the purpose.

**Response:** In response to this and other comments, we have first re-thought the paper and then thoroughly revised it. It is clear that we have two issues to illuminate and discuss, that of concurrent validity broadly, and that of differential performance by ethnicity (hereafter DPE), though of course the two are closely linked. An important occurrence that has occurred between the original submission of the paper and the present re-write is that the High Court action against the MRCGP CSA, pending when we wrote the paper, has been thrown out
with costs awarded against the complaining party. This means that we collectively feel able to be more up-front with the DPE issue and indeed agree with the reviewers about its apparently secondary profile in the paper. It should not be a “side issue” (as another reviewer commented) and now it is not. The outcome is that the two related issues get equal prominence, hence a complete re-write of much of the paper – other than the Methods and Results sections, though in the latter we have borne in mind the very direct comment above.

Apart from that it is well written and clear, although perhaps a bit dense at times when focusing on the numbers.

**Response:** We have revised the results section, not to make substantive changes but to enhance its clarity.

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

**Quality of written English:** Acceptable. *(Thank you, Professor!)*

**Statistical review:** Yes, but I do not feel adequately qualified to assess the statistics.

**Declaration of competing interests:**

I declare that I have no competing interests