Author's response to reviews

Title: The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: An analysis of MRCP(UK) examinations

Authors:

Jane Tighe (jane.tighe@mrcpuk.org)
Chris McManus (i.mcmanus@ucl.ac.uk)
Neil G Dewhurst (nei.dewhurst@mrcpuk.org)
Liliana Chis (liliana.chis@mrcpuk.org)
John Mucklow (john.mucklow@dsl.pipex.com)

Version: 2 Date: 30 April 2010

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

Original comments are in Roman type, and our responses are in Italic.

Reviewer's report

Title: The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: An analysis of MRCP(UK) written examinations, 2002-2008, and Specialty Certificate Examinations

Version: 1 Date: 22 February 2010

Reviewer: Kent Hecker

Reviewer's report:

This manuscript outlines an often misunderstood concept; the relationship between SEM, SD and reliability (in this case either coefficient alpha or a correlation coefficient).

Minor Essential Revisions

For completeness of the article I would strongly encourage the authors to:

1. Provide further information regarding the range of restriction problem, a restriction of range algorithm as well as a worked example showing the change in the reliability coefficient for the values from the Monte Carlo experiment. Given the values provided, I calculated an adjusted correlation coefficient as being 0.87. This should be included on page 10 and 11. It would nicely reinforce your arguments regarding the the relationship between SD and reliability coefficients.

Response: This is a good suggestion, and we have included that calculation using the standard formula.

2. Provide a description of how to interpret the SEM. While most people understand that reliability is reported from 0-1, with a value closer to 1 as being more reliable, what do the percentages mean that are calculated from SEM?

Response: The interpretation of SEM is not as straightforward as that for reliability (which is perhaps why people like reliability), and this paper is not really the place to go into a full explanation. We are however preparing a separate paper on precisely that issue, and therefore here have merely included a reference to the excellent paper by Dudek.

3. Carefully proof read and edit the document, for instance:
a. Being a non-European I do not understand the term “diet” first found on page 7. I think that this paper will have a wide readership throughout the world and therefore would like to see this term explained.

Response: Diet is the conventional (if perhaps somewhat old-fashioned) word used for the MRCP exams (think of the "Diet of Worms", from mediaeval history). The Oxford English Dictionary has, "4. Sc. A session or sitting of a court or other body on an appointed day;" with several nineteenth-century examples (and the MRCP was first sat in 1859). We have put "(sittings)" after the first use of the word, but continue to use diets since that is the official term used in all of the web material etc that describes the examination (see e.g. http://www.mrcpuk.org/Pages/Home.aspx). Hopefully it is no more confusing than the problems for European readers of finding references in North American scientific papers to sophomores, 13th-graders, etc, none of which occur in the UK or other parts of Europe.

b. SCE’s? What are they? Specialty certificate examinations? Pg.8

Response: This abbreviation had actually been spelled out on page 7 of the original manuscript.

c. There are two page 2’s at the end of the manuscript. One before the reference list and one being the 2nd page of the appendix.

Response: There are tedious technical reasons for that. We will try to ensure it doesn’t happen in the final version of this revision.

4. Finally, and this is a subjective comment, I would recommend that the point can be made for calculating and reporting SEM using the PMETB document as an example, not as a criticism of the PMETB document. Through the discussion, I found that I was distracted by the comments regarding the PMETB documents, which became discouraging. I would encourage the authors to stay true to the message of the use and rationalization of SEM and remove/rephrase paragraphs that are distracting, such as the paragraph at the bottom of page 15 and top of page 16 as well as the comments made in the final paragraph of Appendix 2.

Response: It should be remembered that the PMETB document is not merely opinion but also has a prescriptive value, being written by the authority responsible for the regulation of UK postgraduate examinations, and therefore there is an obligation to be correct and accurate. Having said that, we have toned down some of the comments so that the point of the paper is not obscured by such matters. We have also added in that from April 2010, PMETB has been merged into the General Medical Council.

Level of interest: An article of importance in its field
Quality of written English: Needs some language corrections before being Published

Statistical review: No, the manuscript does not need to be seen by a statistician.

Reviewer’s report

Title: The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: An analysis of MRCP(UK) written examinations, 2002 2008, and Specialty Certificate Examinations

Version: 1 Date: 26 February 2010
Reviewer: Dwight Harley

Reviewer’s report:

General Comments:

The study’s objectives were clearly stated and the data collection methods were well described. The data appears to be credible and was developed in an appropriately described manner.

Response: We thank the reviewer for those kind comments,

The manuscript was adequately laid out. The tables and figures were clear. I am not sure that appendix 1 or 2 is necessary and the to the spreadsheet discussed is to be broken. I was unable to access it.

Response: We believe that they are necessary, as otherwise our analysis appears to be merely empirical (i.e. based on calculation of values from particular data sets), rather than being also based in a clear theoretical analysis which shows that the conclusions derive necessarily from the nature of reliability, SD and SEM, and their inter-relationships. That could have been included in the main text, but would probably have put off many readers, but in the appendices they can be referred to by those readers who wish to dig more deeply (or to check what we have done). We have therefore retained them.

Apologies for the broken link to the spreadsheet. We will include it in the paper in its correct form, but have also uploaded the spreadsheet as such to the BMC website as supplementary material, so that the problem should not arise in the future.
The discussion and conclusions were balanced and reasonable given the results. The title reflected the content and the writing was adequate.

Minor Essential Revisions:

Explain meaning of the term “diet”

Response: We have added in the term "sittings" after the first usage of the word diet, which is the proper description of the repeated sittings of the exam, and therefore, as explained above, we have retained 'diet'.

Discretionary Revisions:

Consider deleting the two appendices – I’m not sure that they are necessary. Perhaps make reference to a text reference.

Response: We have considered it, but as explained earlier to one of the other reviewers, we believe they are necessary for some readers (but would be overwhelming and offputting within the text proper). We wonder if something has got garbled in the phrase "Perhaps make reference to a text reference" as we could not understand its intent.

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

Quality of written English: Acceptable

Statistical review: No, the manuscript does not need to be seen by a statistician.

Declaration of competing interests:
I declare that I have no competing interests.

Reviewer's report

Title: The standard error of measurement is a more appropriate measure of quality for postgraduate medical assessments than is reliability: An analysis of MRCP(UK) written examinations, 2002 2008, and Specialty Certificate Examinations

Version: 1 Date: 13 February 2010

Reviewer: Tanya Beran

Reviewer's report:
This paper compares the ‘reliability’ of reliability to standard error of measurement. The conclusions in this paper have significant implications for high stakes testing in medicine, and, thus, are very appropriate for this journal. Some suggestions are made below to strengthen the quality of the paper.

Response: We thank the reviewer for these kind comments, and for the careful reading of the manuscript.

The title can be shortened.

Response: One of the other reviewers commented that the title reflected the content accurately. We have therefore compromised by reducing the subtitle, but have left it indicating that it refers to MRCP(UK) examinations, since that is probably important for readers to know.

The third line on page 5: candidates should be candidate’s

Response: Corrected

Conjunctive adverbs need commas around them when in mid-sentence.

Response: We are grateful to the reviewer for their attention to detail, and, in most cases, we have attempted to correct these errors, so, as far as it is possible, the text does not consequently become less rather than more readable.

In appendix a the sentence after the 3rd formula should read: As a result, the equation can be simplified to provide an approximation to SEM, which, to avoid confusion, is called SEM*.

Response: We assume the reviewer meant Appendix 1 and not a. The correction has been made.

There should be a space before and after all = signs.

Response: Done

In the first formula in appendix b there is a period after 10/9

Response: Again, it should be Appendix 2 and not b. Corrected.

In the fifth last line of appendix b there needs to be a space between remains non-zero

Response: Corrected.
In the first paragraph on p.6 the authors state that reliability is not the sine qua non of an assessment. Given that reliability and validity are often presented hand-in-hand when discussing psychometric properties of tests, it is important to briefly mention the importance of validity in determining the ‘quality of the assessment’, as the authors indicate.

*Response:* We have added an extra sentence here to that effect.

On p.7 ‘very variable’ should be reworded to ‘highly variable’.

*Response:* Done

There should be no punctuation in the header on p.7.

*Response:* Done

On p.7 the first line of the last paragraph: ‘radically’ needs to be re-worded.

*Response:* The change was radical from the perspective of those carrying it out, but we have nevertheless followed the reviewer’s advice.

On p.10 in the second line: ‘in figure’ can be deleted. In the fourth line there are too many periods. In this same line there should be no comma after seen.

*Response:* Corrected.

On p.10 fourth last line of first paragraph: (929 is an error. These types of errors occur throughout the paper, requiring a careful proof reading.

*Response:* These errors occur due to different versions of Windows having different fonts, and despite the PDF format being designed to prevent it, they still occur in many of the manuscripts that I receive as an editor or reviewer. I presume the matter needs taking up with Microsoft or Adobe.

The statement on p. 10 that ‘something has to be wrong’ is overstated and needs to be clarified.

*Response:* ‘Overstated’ is surely overstated. We have however changed the wording.

The first paragraph of the discussion is a run-on sentence.

*Response:* As Wikipedia says, a run-on sentence, "may be used as a rhetorical device". That was its use here. It has been left as it is.

Given the recommendation of the use of SEM over reliability in the conclusion
section, it may be worthwhile in the discussion section to compare the interpretation of reliability and SEM. That is, people easily understand the range of reliability coefficients (0 to 1.00 with higher reliability closer to 1.00), but how can people interpret a SEM of 3.06 for example? Some guidance to the reader here may encourage its use.

Response: We agree that the interpretation of the SEM is an important issue (and it is often used wrongly, that is also clear). There is not however space to discuss the issue here (and we have a long paper in preparation that discuss the matter). We have added in a sentence therefore which refers the reader in the first instance to Dudek’s excellent paper.

Level of interest: An article of importance in its field

Quality of written English: Needs some language corrections before being published

Response: Done

Statistical review: Yes, and I have assessed the statistics in my report.

Declaration of competing interests: ‘I declare that I have no competing interests’

Comment of Associate Editor

Associate Editor comments:

All three reviewers have responded very positively to this paper and recommended that the paper be accepted for publication with some minor mandatory revisions. Please address each of the reviewer’s comments and indicate your response to them. In particular, address Hecker’s comment:

I would recommend that the point can be made for calculating and reporting SEM using the PMETB document as an example, not as a criticism of the PMETB document. Through the discussion, I found that I was distracted by the comments regarding the PMETB documents, which became discouraging. I would encourage the authors to stay true to the message of the use and rationalization of SEM and remove/rephrase paragraphs that are distracting, such as the paragraph at the bottom of page 15 and top of page 16 as well as the comments made in the final paragraph of Appendix 2.
Response: We have removed most of the references to PMETB, but not all. As mentioned earlier they are/were the principle regulator of post-graduate examinations in the UK, and therefore what they say is of great import to those running high-stakes examinations.