Author's response to reviews

Title: Accuracy of portrayal by standardized patients: Results from four OSCE stations conducted for high stakes examinations. 1. Lubna A. Baig
Corresponding Author Email: lbaig@ucalgary.ca and baigl@ksau-hs.edu.sa 4755 Hamptons Way NW Calgary, AB # T3A 6K1, Canada 2. Tanya N Beran Email: tnaberan@ucalgary.ca 3. Andrea Vellavand Email: avalleva@ucalgary.ca 4. Zarrukh Ali Baig Email: zabaig@ucalgary.ca 5. Mauricio Monroy-Cuadros Email: mmonroy@ucalgary.ca

Authors:
- Lubna A Baig (baigl@ksau-hs.edu.sa)
- Tanya N Beran (tnaberan@ucalgary.ca)
- Andrea Vellavand (avalleva@ucalgary.ca)
- Zarrukh A Baig (zabaig@ucalgary.ca)
- Mauricio P Monroy-Cuadros (mmonroy@ucalgary.ca)

Version: 4 Date: 12 March 2014

Author's response to reviews: see over
Dear Dr McKinley,

Following changes have been made in accordance with the comments of the reviewers:

**Reviewer 1 Major revisions**

1. The information that is the basis for training was considered the 'gold' standard. I assume that this provided the basis for the global rating scale as well as the checklists for each case, but that wasn't clear.

   *Yes it is and we have removed the word Gold standard and made it clearer in the methods section*

2. I would recommend that the authors change the phrasing when referring to the SPs giving information from 'asking' to 'being asked'.

   *These changes have been made in the text and also on the tables*

3. In some administrations of SP examinations for licensure and certification, presenting information is presented before the encounter begins by providing a printed sheet. In this examination, does the SP present the complaint to begin the encounter? More information about these details of the administration are needed to assist with interpretation of the information presented on accuracy of presentation. Does this refer to the chief complaint or the affect used in beginning the encounter?

   *We have clarify this in the procedures section*

4. I assumed that the emergency management case was one that presented with distended abdomen in obstetrics/gynecology (Case B), but that wasn't clear in the text.

   *Yes that is correct and the case presented in ER, we have tried to clarify it a little more. This was problematic as we have to keep the secrecy of the exam as well.*

5. The internal consistency of the checklist rated by physicians for Case C was fairly low, and no potential explanation was offered for this in the discussion section of the paper. Why might the raters supply such inconsistent ratings for this particular case?

   *We have tried to discuss this in the discussion session. The reason could be that the SPs across all tracks were very inconsistent and in our opinion as this was a counseling case with a bulky script the SPS were inconsistent.*

6. Can you explain why generalizability (variance components analysis) were not used to examine the variance due to case, rater, SP, or some interaction?

   *Candidates scores were not available for us to do the analysis hence it was not done.*
7. Although there were findings that the SP portrayal accuracy could be improved, what effect did this have on candidate performance? Here, I think a bit of descriptive statistics would be helpful. For all candidates seeing these SPs (not only those encounters included in the study), were there differences for the cases that were common across SPs and locations?

Same response as above

8. The method in which scores are equated over time were not the focus of the study, but in considering whether to provide feedback, it will be essential to consider the administration cycle for the examination. Changes in portrayal are likely to affect candidate performance, and whether there is score equating or not, stability in the prompts received may be preferred. Improvement can occur at points in the administration period that have the least potential for adversely impacting the examinees.

We agree that the portrayal can affect the scores of the candidates; however as that was not the focus of the study we did not have ethical approval to do this analysis.

Minor essential revisions

While the authors have done a commendable job in reviewing research that is most directly relevant to SP portrayal accuracy, I recommend that they consider two other articles that affect the validity of the interpretation of the results: McKinley and Boulet (2004), The effect of task sequence on examinee performance, Teaching and Learning in Medicine; and McKinley & Boulet (2004), Detecting score drift in high-stakes performance based assessment, Advances in Health Sciences Education. Although these are not specifically about SP portrayal, they do offer some rival hypotheses about differences between SPs performing the same station and about time (AM vs PM).

We are very familiar with the work of McKinley and Boulet and have tried to add them in the intro and discussion.

Reviewer 2 Major revisions

1. The title specifically states results are based on 10 OSCES, but the author states in the Abstract and Methods section that “Four stations from a ten-station high-stakes OSCE were selected …” To establish consistency between the title, abstract and methods either remove the number of OSCEs from the title or use the same number throughout.

Title changed

2. Describe the Multiple Mini Interview in greater detail.

Done

3. The Procedure section needs to be rewritten for clarification regarding the
design for the study. For example, the tracks (sessions?) are labeled yellow, red, blue, and green in discussion and numbered in Table 1. Keep the discussion consistent with associated labels in Tables and other sections of the manuscript. Why were six SPs selected from each assessment track instead of using all SPs? Were the ones selected chosen randomly?

Description done in the methods section and reasons for selection clarified

4. Cohen’s Kappa indicated sufficient reliability between raters. Clarify if ratings were combined between raters or kept separate for the analysis.

Modal value was used for analysis and this is explained in the methods and the results section

5. Provide details about how scores were derived.

Tried to clarify this in procedures section

6. Expand the analysis section with the statistical method used in reported results.

Some minor changes have been although there was not a lot of stats done

7. Consider including a generalizability analysis to yield information about case and other effects of interest in this study.

Not possible as we did not have approval for using examinees scores

8. To improve clarity, report the same p value in the discussion that is associated in each Table. Due small sample size justification needs to be provided for the type of statistical method used to derive statistical significance (see comment 5 above). Perhaps a descriptive approach is sufficient. The last row in Tables 2-5 was not included in discussion within the Results section or needs to be discussed more clearly.

Errors rectified and changes made in the tables, last row also explained

9. Include additional limitations that are associated with this study. For example, only 4 out of 10 OSCE stations were included in the study.

More limitations added

Minor Essential Revisions

1. Correct typos and use of written language throughout the manuscript.

We have tried our best to catch all the errors

2. Check that references have been correctly cited in the manuscript.

Done
3. On page 4 in the Background section is the first sentence referring to items or sources of measurement error “… the interaction of one or more of these items [sources of measurement error] with candidates.”

   This is rectified

4. On page 6 revise the sentence at the top of the page. Assessment and instrument are redundant and use of the term Gold standard seems a bit awkward.

   Revised it is not called Gold standard anymore and explanation given in the methods as well

5. On page 8 when giving examples of items on the instrument use the same wording for each item shown on the checklist for each case.

   Done

I hope these changes will make it publishable for your journal.

Kind regards

Lubna Baig

Corresponding author