Reviewer's report

Title: Accuracy of portrayal by standardized patients: Results form a 10 station OSCE 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: tnbaberan@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

Version: 2 Date: 22 January 2014

Reviewer: William Roberts

Reviewer's report:

General comments

The purpose of this study was to rate the accuracy and realism of standardized patient (SP) portrayal by experienced clinicians. I have considered a number for points in this review to help improve clarity throughout the manuscript. I did find the research question clear and relevant to understanding the potential sources of measurement error associated with SP ratings of examinee performance. However, written discussion and technical weakness will need to be addressed before the manuscript is ready for publication. Specifically, the following comments are provided in the next sections of this review.

Major Compulsory 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.

2. Describe the Multiple Mini Interview in greater detail.

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?

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

5. Provide details about how scores were derived.

6. Expand the analysis section with the statistical method used in reported results.
7. Consider including a generalizability analysis to yield information about case and other effects of interest in this study.

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.

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

Minor Essential Revisions

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

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

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.”

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.

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.

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

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.

Declaration of competing interests:

I declare that I have no competing interests.