Reviewer's report

Title: Accuracy of portrayal by standardized patients: Results from 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
2. 4755 Hamptons Way NW Calgary, AB # T3A 6K1, Canada
3. Tanya N Beran Email: tnaberan@ucalgary.ca
4. Andrea Vellavand Email: avalleva@ucalgary.ca
5. Mauricio Monroy-Cuadros Email: mmonroy@ucalgary.ca

Version: 2 Date: 17 February 2014

Reviewer: Danette McKinley

Reviewer's report:

Major compulsory revisions

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.

I would recommend that the authors change the phrasing when referring to the SPs giving information from 'asking' to 'being asked'. For example, on page 9.: 'In all the tracks SPs did not give information without asking for questions on cardiac risk factors, past history, and on diet and weight'. I think the authors are referring to the SPs providing information without being asked - 'In all the tracks, the SPs did not give information without being asked questions on cardiac risk factors, past history, and on diet and weight'. Similar changes would be made throughout the results section.

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?

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.

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?

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

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?

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.

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).

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

Declaration of competing interests:

I declare that I have no competing interests