Reviewer's report

Title: Modification of an OSCE format to enhance patient continuity in a high-stakes assessment of clinical performance

Version: 1 Date: 22 November 2010

Reviewer: Danette McKinley

Reviewer's report:

Major Compulsory Revisions (which the author must respond to before a decision on publication can be reached)

1. I have a few comments on the manuscript. First, there was no mention of how passing standards were established for the pilot, although there was a note for the first table that indicated which ratings constituted ‘failing performance’. There was no mention of how participants were scored across the six sections to determine passing performance, although this was reported in the results. The manuscript is fairly brief, and it should not be difficult to add some information about the way that decisions were made across sections and scenarios. Since the format was a bit different, I was interested in how many participants were classified as 1-3 by section, particularly for the second station where the standard deviation was quite small (e.g., Station 2, section c. mean=4.17; sd=0.65).

2. Please say more about the independence of the sections. With correlations within sections accounting for 30% to 48% of the variation in the ratings of performance, it seems that knowing the correct answers in one section (physical examination, perhaps) would have some affect on knowing proper management. It was not surprising that the correlation with counseling could be lower (within stations), but were these ratings higher across stations?

Discretionary Revisions (which are recommendations for improvement but which the author can choose to ignore)

3. Also, I could see why the design was set up the way it was for the generalizability analyses – but wondered whether there should have been a rater facet – perhaps nested in scenario. So little variance was due to station (0) and section (0), and so, this facet may have been redundant in the design. To be fair, this was also addressed a bit in the inter-rater reliability correlations that were reported.

4. While I think that having participant and examiner reaction was essential and added an important element to the report, I would appreciate it if the authors could address some of the concerns of the examiners in their discussion. How would increasing the number of scenarios address the concerns of the examiners? Would this permit the combination of a ‘long case’ with the traditional OSCE? Would it be feasible to include two scenarios as described and have the
remaining scenarios in the ‘old’ format? What scoring challenges might this introduce? Can the separate scores (physical examination, management, and communications) still be calculated with this change in design? The discussion seems to be the ideal place to speculate about possible challenges and solutions, and may inform other researchers as well as the future research of these authors.

**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.