Reviewer's report

Title: Multiple tutorial-based assessments: a generalizability study

Version: 1 Date: 20 August 2013

Reviewer: William Roberts

Reviewer's report:

GENERAL COMMENTS

The objective of this study was to investigate the reliability of a shortened version of the Tutotest-Lite using generalizability theory. Results indicated sufficient reliability with fewer items on the test.

Major Compulsory Revisions

This review includes a list of comments to address areas of strength and weakness with recommended corrections. Technical and general areas of weakness in the manuscript will need to be addressed before acceptance for publication.

1. The first sentence in paragraph 1 of the Methods section is awkward. It is not clear to me a “unit” is a scored component or dimension of the Tutotest-Lite. The reference to Table 1 could be clarified by stating entries are summary statistics.

2. I would like more be said about the formation of new groups to each PBL content area? Is PBL standardized among newly formed groups?

3. The sample for this study appears to be from a single medical school. This should be mentioned as a limitation when generalizing findings to the population of all medical students.

4. In paragraph three in the Methods section it is stated “Committee members determine pass/fail standings from the individual scores”. It would improve clarification if the number of members on the committee were given and how the standard for a pass decision is operationalized.

5. In the section describing the Tutotests-Lite it is stated in the first paragraph “Clinical domains to be developed within PBL sessions were identified through focus groups composed of faculty members”. If applicable expand discussion to include the extent of overlap between medical practice and domain content identified by committee focus groups.

6. The design notation for the effects G-study design needs to be specified more clearly. For example, should this be conceptualized as a univariate or multivariate G-study design? The definition for “unit” and “clinical domains” is not clear to me. Perhaps the definitions would be clarified if expanded in detail when first mentioned on page 5. As part of the design specification state if this is a
norm referenced or criterion based examination. If this is a criterion based examination, then the index of dependability is the appropriate measure of reliability.

7. In the second paragraph in the Analysis section it is stated “Analysis for Year 1 was conducted using only the three common abilities …”. This seem to contradict what was stated earlier in the section describing the Tutotest-Lite, “Clinical skills are not taught and therefore not assessed in three of the eight units in Year 1 …”. Clarify what is meant by these two statements.

8. In the first paragraph within the Overall Reliability section it is not clear what “Additional file 1” is referring to when discussing variability components. Perhaps this is meant to refer to Table 2? The reference “Additional file 1” also shows up in the second paragraph.

9. Include dependability coefficients in Figure 2.

Minor Essential Revisions

1. For consistency, in the third paragraph of the Background section numbers and words should not be mixed when referring to quantity “(e.g. three to 31 items)”.

2. The first paragraph in the Data section could be shortened by simply reporting that the sample consists of 384 first-year and 374 second-year students representative of the school in this study.

3. In the second sentence of the first paragraph in the Results section keep the format consistent in reporting means and standard deviations for each of the two years. Also, clarify if statistics are SEM or SD. In Table 1 these are listed as SD.

4. For each year, include test grand means and standard deviations in Table 1.

5. Lines in Figure 2 can be better contrasted if one of the lines is segmented.

Discretionary Revisions

1. In the second paragraph of the Background section remove “for example” at the end of the first sentence.

2. I suggest referring to the scale as a four-point rating scale and dropping the word Likert.

3. I suggest “Unit percentage scores means” be rephrased to unit means in percentage metric?

Level of interest: An article of importance in its field

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.