Reviewer’s report

Title: Tutor Assessment of PBL Process: Does tutor variability affect objectivity and reliability?

Version: 0 Date: 20 Oct 2018

Reviewer: Grace Wong

Reviewer’s report:

I found your manuscript interesting and relevant to the journal's readership.

Abstract

The abstract is concise. However, there are discrepancies on the number of tutors between the abstract and the manuscript. There is no mentioning of student sample size and significant values of findings in the abstract.

Literature Review:

The authors started the manuscript by describing the advantages of Problem-Base Learning (PBL) and listed a range of assessment methods to examine student's learning. Are there any problems with these assessment methods? Which assessment methods will give less discrepancies among raters? Which methods are better in maintain objectivity of raters and ensure reliability? The authors described the Hawk-dove effects. You will need to link this with using of a rubric/criteria may eliminate this effect.

The authors need to provide the year of the study in literature review on page 4. Although there are few studies mentioned but the authors failed to identify the gap of knowledge and build up a good justification for this research.

The aim of the study was to examine the differences between lenient stringent tutors in assessing PBL tutorials. Is this an assumption? How do you define lenient and stringent in this study?

Methods:

The authors used an appropriate research design. However, there are confusion on number of tutors and number of groups that these tutors are assessing. The authors did not state what type of assessment was carried out? Were they group presentation? Individual presentation? Care plan? The authors listed the twelve performance criteria used in this assessment, however, no details were given if these are 'yes' or 'no' criteria or if there are different sub-scale in each of these criteria.
The assessment instrument has been used by the school for more than 25 years. Reference 26 is not a publication or reliable source of information. How do you justify it is comparable to the other instruments which were used by the other pioneer schools? Did you adapt their instrument? Were the tutors trained to use this instrument? Have all the tutors used this instrument before?

It was not clearly stated how the data was collected? Was the tutor leading the group the assessor? Did each tutor assess each group? Were the assessment carried out in one day or in a period of time? Was all the PBL tutorial content the same for this assessment?

The authors stated that ethics approval was not necessary as this was part of the school assessment monitoring strategies. However, the results stated certain tutors were lenient or stringent which infringed on individual's privacy and reputation.

Results:

The first two bar charts are clear. The first figure showed the mean rating of individual group with G8 scored 58 the highest and G9 scored 40 the lowest. Figure 2 showed the mean rating of individual tutor ranges from 31 to 63 which revealed a statistically significant difference among the tutors. However, the rest of the results are confusing.

How do you classify stringent and lenient tutors? Are the tutors with a higher mean rating scale lenient, with a lower mean rating scale stringent? I would define the tutor with similar rating scale stringent instead. I am not convinced about the necessity of comparing the mean ratings of each tutor. I would be interested in the the discrepancies in different criteria. Which criteria were rated consistently or inconsistently among tutors? Using a line chart to plot out the rating by different tutors and the 12 marking criteria would be useful.

Any results on the years of tutor experience and the mean rating score?

Discussion:

The authors should link the findings to literature which has been discussed in the literature review section. As this instrument was comparable to other instruments, could results be compared?

Although the results revealed significant discrepancies in rating the assessment, there is no reason to define the high rater as lenient and the low rater as stringent. For the assessment results to be valid, you need to maintain consistency. It would be interesting to find out which are the criteria causing the discrepancies in rating. The authors can use these results to discuss how these can be rectified to ensure reliability of the assessment.

The authors have addressed a number of limitations of this study, such as PBL contents. If this tutors were not assessing the groups with the same tutorial content, this study is not valid.
Conclusion:

The training of tutors should be put in discussion. It is a good suggestion to explore tutor assessment with self and peer assessment of PBL.

In general, the authors have established a sound research question. The research design is reasonable but not written in a style that is logical and clear to follow. It is obvious that the abstract, the limitations and results were written by a different author as the writing style is different. There are problems with referencing. Some part of the manuscript has a reference to each of the sentence.

Are the methods appropriate and well described?
If not, please specify what is required in your comments to the authors.

No

Does the work include the necessary controls?
If not, please specify which controls are required in your comments to the authors.

No

Are the conclusions drawn adequately supported by the data shown?
If not, please explain in your comments to the authors.

No

Are you able to assess any statistics in the manuscript or would you recommend an additional statistical review?
If an additional statistical review is recommended, please specify what aspects require further assessment in your comments to the editors.

I recommend additional statistical review

Quality of written English
Please indicate the quality of language in the manuscript:

Needs some language corrections before being published

Declaration of competing interests
Please complete a declaration of competing interests, considering the following questions:

1. Have you in the past five years received reimbursements, fees, funding, or salary from an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?
2. Do you hold any stocks or shares in an organisation that may in any way gain or lose financially from the publication of this manuscript, either now or in the future?

3. Do you hold or are you currently applying for any patents relating to the content of the manuscript?

4. Have you received reimbursements, fees, funding, or salary from an organization that holds or has applied for patents relating to the content of the manuscript?

5. Do you have any other financial competing interests?

6. Do you have any non-financial competing interests in relation to this paper?

If you can answer no to all of the above, write 'I declare that I have no competing interests' below. If your reply is yes to any, please give details below.

I declare that I have no competing interest.

I agree to the open peer review policy of the journal. I understand that my name will be included on my report to the authors and, if the manuscript is accepted for publication, my named report including any attachments I upload will be posted on the website along with the authors' responses. I agree for my report to be made available under an Open Access Creative Commons CC-BY license (http://creativecommons.org/licenses/by/4.0/). I understand that any comments which I do not wish to be included in my named report can be included as confidential comments to the editors, which will not be published.

I agree to the open peer review policy of the journal.