Author’s response to reviews

Title: Impact of video feedback system on medical students' perception of their clinical performance assessment

Authors:

Bee Sung Kam (beesung@pusan.ac.kr)
So Jung Yune (cc139@pnu.edu)
Sang Yeoup Lee (saylee@pnu.edu)
Sun Ju Im (sunjuim11@hanmail.net)
Sun Yong Baek (sybaek@pnu.edu)

Version: 2 Date: 30 May 2019

Author’s response to reviews:

May 30, 2019

Editor-in-Chief
BMC Medical Education.

MEED-D-17-00837R1 Title: Impact of video feedback system on medical students’ perception of their clinical skills assessment
Dear Sir:

We are very grateful to you and the reviewers for the comments on the 1st revised version of our manuscript. This letter represents our response to the reviewer’s comments about our paper. We carefully considered all suggestions by the reviewer and submit a 2nd revised manuscript. We highlighted with red color all changes made in the revised manuscript. Please find below a point by point response to editor’s and reviewer’s comments and clarify the important points of your main concerns. We did our best to revise the 1st revised manuscript according to the constructive advice of the peer reviewers. We believe these actions address the deficiencies and comments noted by the reviewer. We hope that you will be pleased with this revision and the revised manuscript is accepted and published in your esteemed journal, the “BMC Medical Education” for publication. We thank you again for the constructive review by the editor and reviewer.

Sincerely yours,

Sang Yeoup Lee M.D.

Family Medicine Clinic, Pusan National University Yangsan Hospital, Beomeo-ri, Mulgeum-eup, Yangsan, Gyeongsangnam-do, 626-770, South Korea

Phone: +82-55-360-1442, Fax: +82-51-510-8125, E-mail: saylee@pnu.edu

[Responses to the comments of Sylvia Heeneman (Associate Editor)]
For the comments of Reviewer 1:

Q1. Previous Q4 ("don't quite understand line 15 "students do not always agree with their score." At first, I thought, perhaps "students do not feel the score generated from the checklist is an accurate representation of their performance or skill." Having read the article further I am not sure if they disagree with how the checklist was filled out or if the checklist is a fair assessment of their skill."). This comes back twice, it has been adapted on p4 (manuscript p3, line 16), but the statement is also made on p4 line 47, and the same nuance should be made there

A1. Thank you for your sharp comments. We also amended the statement you pointed out as follows: “Providing personalized video feedback of student performance during CPA in addition to hand-written comments is expected to make students more likely to accept with their score results based on only an analytical checklist.” □ “Students may accept that the score generated from a checklist properly assessed for their performance if personalized video feedback of their performance during CPA is provided in addition to hand-written comments and an analytical checklist.”

Q2. Previous Q13 ("I'm not sure if student agreement with their score or comments is an appropriate outcome.") is along the same lines of Q4 of this reviewer and I agree that acceptance of score may not be a good measure of feedback acceptance, if it is, it is at most an indirect measure and this needs more attention, it should at least be mentioned as a limitation. A limitation section in the discussion is not present and I would advise to include one.

A2. We thank your very helpful comment. We added it in the limitation section as follows: “It should be noted that this study was limited in that acceptance of the test score was part of the overall feedback system acceptance, which may not be sufficient alone, because this is an indirect measure that requires more caution when interpreting the results. Feedback might be useful if it is accepted by those being evaluated. However, the results of this study revealed some students who, although accepting their scores at first, no longer accepting them viewing the video feedback. Accordingly, additional interviews should be conducted to ensure that students understood their scores well and considered the test results to be fair and appropriate; unfortunately, however, such interviews were outside the scope of this study.”
Other comments:

Q3. The writing is not always clear, the reviewers commented on this as well, e.g. had specific questions for clarification and I agree. The articulation of certain statements raises questions on what is meant. Also, there are some typo-s in the use of plurals

E.g. Background p4, line 38-40: ‘feedback can be viewed as not feedback to the students’: should this be read as feedback is not accepted by the student?

E.g. Background p4, line 43, here I would use ‘teacher’ or ‘evaluator’ (as was done previously) and not ‘professor’

A3. Yes. As your suggestion, we have amended those parts to make it more clearly including the following: “if the student does not agree to the feedback, then the feedback can be viewed as not feedback to the student.”
‘professor’→‘evaluator’

Q4. Although that was not your question: while reading I agreed with Reviewer 2 (comment on the design of the questionnaire, that is still not addressed, p5, line 51, at least add the literature that was used. And the statement p7, line 10-11 ‘conventional Delphi’ etc is still not very informative (cannot be replicated)

A4. Thank you for your comments. We added literature[23, 25-27], informs of the design of the questionnaire and conventional Delphi’ including the following: “For the development of the questionnaire, two rounds of Delphi expert consultation were conducted with five faculty members in the department of medical education and a faculty focus group (n=10) selected among members of the Clinical Skills Committee. During the first round, experts were asked to provide their opinions in a questionnaire consisting of open-ended questions about the evaluation area and evaluation items. Items selected in the first-round analysis were presented to each expert by email in a second round, when experts were asked to use the Likert 5-point scale to evaluate whether they agreed with inclusion or exclusion of items according to the importance of each factor and item. Experts were also asked to describe the suitability of the evaluation system and the comments pertaining to the items to be revised and supplemented by the evaluation factors. Experts did not meet face-to-face, and they completed their assessments independently. The content validity was based on the content validity ratio (CVR) proposed by Lawshe [28]. The CVR ranged from a maximum of +1.0 to a minimum of -1.0. If the CVR was positive, more than half of the respondents answered 'appropriate', which meant they were rated 4 or 5 on the Likert 5-point scale. The CVR gives the minimum value according to the number of panels.
When the value was above the minimum value, it is judged that there is content validity for the item. The number of panels in this study was 15, and the content validity was found to be more than 0.49. In the second round, all of the developed items were available because the average of the validity responses was 4.5 or more. Finally, the questionnaire consisted of 4 items regarding CPA total score reports including CPX and OSCE, 12 items regarding CPX score reports, 2 items regarding OSCE score reports, 2 items regarding online written comments, and 2 items regarding video feedback system. The contents of the questions are shown in Table 1. The same questionnaire was administered before and after providing video feedback. Only questions pertaining to the usefulness of video feedback were added. Answers were given on a 5-point Likert-type scale from strongly disagree to strongly agree, which is used to allow the individual to express how much they agree or disagree with each question. Two open-ended questions concerning the CPX station that students disagreed with for the CPX station score and OSCE station score were presented at the end of the questionnaire. Completion of the questionnaire took approximately 30 min.

Q5. Likewise: Methods, p7-8, line 58, line 1, I disagree with the use of the statement ‘marginally significant’ and would advise against it. What the result show is a slight change, the students already scored high in the Likert scale at timepoint 1 (3.8-3.9), the overall change could be interpreted as meaningful, the intervention of the video feedback seemed to have an effect on how students perceived their performance, and my advice would be to present this as such. Valid suggestion for follow-up were already made by the reviewers and have been addressed in the discussion.

A5. Yes, as your suggestion, we removed the statements ‘marginally significant’.

Q6. Small comment: Discussion p9, line 31, a satisfaction with the video feedback does not necessarily indicate that the students were efficient in understanding and accepting CPA results, this is again a nuance to be made in line with Reviewer 1 comment (Q13).

A6. Yes. As your suggestion, we have amended the statement as follows: “The satisfaction rate of the video feedback system was more than 4 out of 5, indicating that students were very effective at understanding and accepting their CPX results.” “This change could be regarded as meaningful and indicates that the intervention of the video feedback seemed to have an effect on how students perceived their performance; however, care should be taken when interpreting these results.”
Q7. Page 5 - please be specific as to what literature was reviewed for the scale. Presumably literature on giving feedback on OSCE performance but should be specified other than saying "the literature".

A7. Thank you for your comments. We added references of the literature [23, 25-27].

Q8. Page 5 - I am not sure what "consensus of faculties" means. I assume it should be reworded to be something like "based on the consensus of faculty members in the department of medical education. Also, who were these faculty members? Were any of them actually qualified to be judging feedback in an OSCE? A bit more detail around the qualification of the judges is needed.

A8. The faculties mean faculties of Clinical Skills Committee and department of medical education. They all are expert in this fields. We have described it in more detail.

Q9. Page 5 mentions a consensus method is used and Page 7 mentions a conventional Delphi. The problem is that the Delphi method means many different things, so it needs to be described in a bit more detail. (See work recent work by Susan Humphrey-Murto on consensus methods). When I look at page 7, additional information needed includes, how were the items sent out, how many items were sent out, what was the response rate, and how was consensus defined.

A9. Yes. As your suggestion, we have described it in more detail as follows: “For the development of the questionnaire, two rounds of Delphi expert consultation were conducted with five faculty members in the department of medical education and a faculty focus group (n=10) selected among members of the Clinical Skills Committee. During the first round, experts were asked to provide their opinions in a questionnaire consisting of open-ended questions about the evaluation area and evaluation items. Items selected in the first-round analysis were presented to each expert by email in a second round, when experts were asked to use the Likert 5-point scale to evaluate whether they agreed with inclusion or exclusion of items according to the importance of each factor and item. Experts were also asked to describe the suitability of the evaluation system and the comments pertaining to the items to be revised and supplemented by the evaluation factors. Experts did not meet face-to-face, and they completed their assessments independently. The content validity was based on the content validity ratio (CVR) proposed by Lawshe [28]. The CVR ranged from a maximum of +1.0 to a minimum of -1.0. If the CVR was positive, more than half of the respondents answered 'appropriate', which meant they were rated 4 or 5 on the Likert 5-point scale. The CVR gives the minimum value according to the number of panels. When the value was above the minimum value, it is judged that there is content validity.
for the item. The number of panels in this study was 15, and the content validity was found to be more than 0.49. In the second round, all of the developed items were available because the average of the validity responses was 4.5 or more. Finally, the questionnaire consisted of 4 items regarding CPA total score reports including CPX and OSCE, 12 items regarding CPX score reports, 2 items regarding OSCE score reports, 2 items regarding online written comments, and 2 items regarding video feedback system. The contents of the questions are shown in Table 1. The same questionnaire was administered before and after providing video feedback. Only questions pertaining to the usefulness of video feedback were added. Answers were given on a 5-point Likert-type scale from strongly disagree to strongly agree, which is used to allow the individual to express how much they agree or disagree with each question. Two open-ended questions concerning the CPX station that students disagreed with for the CPX station score and OSCE station score were presented at the end of the questionnaire. Completion of the questionnaire took approximately 30 min.”

Q10. Statistical analysis - I am holding firm on not using the term marginally significant for two reasons. First, if I wanted to be a purist, there are actually 20 comparisons so the p-value should be corrected for the family-wise error rate and be .05/20 = .0025. It is an exploratory study to some degree so one could argue for keeping the p-value at .05 but given that the p-value should actually be considerably lower, there is no justification for discussing questions in which the p-value is between .05 and .10. Second, while I thank the authors for adding in effect sizes, I note that they are all small. A p-value between .05 and .10 combined with a small effect size indicates to me that there is nothing there of interest. I think just pointing out that all the means are higher and then focusing on the subset that is statistically significant and not marginal is advisable.

A10. Thank you for your reasonable and thoughtful comments. As you pointed out, we removed “the term marginally significant”.

Q11. Upon re-reading this version, table 2 and 3 triggered a thought. There is a paper by Kruger and Dunning on self-assessment. They showed that people who were good performers tended to rate their performance lower than they should have and people who were poor performers tended to rate their performance higher than they should have. I am curious if the people who disagreed with the result were all poor performers. Presumably they are and I think having a look and reporting on this provides an interesting link to a major finding in self-assessment. I wish I had clued in on this pattern on the first read.

With regards to table 2 and 3, on page 8 there is a comment that the tables show a change in perception. This terminology could be more specific, the tables actually show the number or frequency of people who did not agree with their feedback.
A11. Thank you very much for your interesting point. We appreciate how you considered our findings. There were 15 students disagreed with the results. They ranked from 21th to 94th. As you suggested, they definitely were not among good performers, but rather they seemed to be middle or poor grade students as following gradings.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>26/126</td>
<td>20.63</td>
</tr>
<tr>
<td>34/126</td>
<td>26.98</td>
</tr>
<tr>
<td>37/126</td>
<td>29.36</td>
</tr>
<tr>
<td>45/126</td>
<td>35.71</td>
</tr>
<tr>
<td>47/126</td>
<td>37.30</td>
</tr>
<tr>
<td>49/126</td>
<td>38.88</td>
</tr>
<tr>
<td>54/126</td>
<td>42.85</td>
</tr>
<tr>
<td>55/126</td>
<td>43.65</td>
</tr>
<tr>
<td>88/126</td>
<td>69.84</td>
</tr>
<tr>
<td>91/126</td>
<td>72.22</td>
</tr>
<tr>
<td>92/126</td>
<td>73.01</td>
</tr>
<tr>
<td>102/126</td>
<td>80.95</td>
</tr>
<tr>
<td>103/126</td>
<td>81.74</td>
</tr>
<tr>
<td>119/126</td>
<td>94.44</td>
</tr>
</tbody>
</table>

Q12. Finally, I found some of the text, especially the new revisions to have some awkward wording. It might be of benefit to run the manuscript by an English grammar editor before final publication.

A12. Yes, this second revised manuscript was edited again by an English grammar editor before final publication.