Author’s response to reviews

Title: Voluntary vs. Compulsory Student Evaluation of Clerkships: Effect on Validity and Potential Bias

Authors:

Sola Aoun Bahous (sola.bahous@lau.edu.lb)
Pascale Salameh (pascalesalameh1@hotmail.com)
Angelique Salloum (angelique.salloum@lau.edu.lb)
Wael Salameh (wael.salame@lau.edu.lb)
Yoon Soo Park (yspark2@uic.edu)
Ara Tekian (tekian@uic.edu)

Version: 1 Date: 26 Sep 2017

Author’s response to reviews:

Ms. Ref. No.: MEED-D-17-00327

Title: Voluntary vs. Compulsory Student Evaluation of Clerkships: Effect on Validity and Potential Bias

BMC Medical Education

Dear Dr. Raupach and Distinguished Reviewers:

On behalf of the authors of the manuscript, I would like to thank you for your time and for your thoughtful review and consideration of our manuscript. Please find our responses to your comments and questions below. In brief, the authors agreed to make changes based on all suggested comments without reservation.

Sincerely,

Sola Aoun Bahous, MD, PhD, MHPE
Reviewer 1:

Comment 1:

Consideration of authority bias: “Wish the sham question had been included throughout and not just in the final cohort!”

Response 1: We agree that administering the sham question to a larger cohort would have added more meaning to the study. However, we thought that students who will notice the sham question would communicate with the rest, which makes future answers to the question biased. For this reason we elected to administer the question only once, at the middle of the year, to allow some practice of the authority-based approach. This information was added to the Methods section, page 9, lines 13-16.

Comment 2:

“Why only half the year in the new (compulsory) system? Does the data or conclusions change if we look through the entire year?”

Response 2: We agree with this pertinent observation. As stated above, we considered that some students would notice the sham question and communicate this to the rest of the class. After the sham question was introduced, further evaluations could become biased. Furthermore, to confirm our assumption of the bias, we had to administer a second survey to students. After this intervention, we found no utility from continuing the evaluations because students would now be attentive to the content after they went through the experience. We agree with the author that should the sample be higher, results could be different. This is indicated in the paragraph about study limitations under Discussion on page 16, line 22 and page 17, line 1.

Comment 3:

“Similar presence of comments (this would be a goldmine to explore in a future qualitative study).”

Response 3: We thank the reviewer for pointing this out. Indeed, comments should be explored in a next step to examine themes and trends in students’ comments about their learning experiences in clinical clerkships.
Comment 4:
“Table 1 comes across as very busy and less useful (baseline characteristics from voluntary group). It could potentially be included as just a supplementary table if there are space limitations.”

Response 4: We agree that the table is loaded with information. We eliminated the “Useful Teaching Activities” rows because they don’t add much to the study and we removed related sentences in the manuscript.

Comment 5:
“OB/gyn regresses to mean.”

Response 5: We acknowledge that the OB/GYN descriptive statistics might appear to regress toward the mean. Yet, the fact that there are variations across clerkships indicate that student evaluations differ and are not always regressing toward the mean. Since this paper reports these findings as preliminary results, we hope to confirm these findings again in a future project.

Comment 6:
“I am struggling to figure out Table 3 looking at nonresponse bias. This is also referenced in the abstract and requires clarification. R v. NR table clearly indicates a higher grade for the respondents but the authors minimize this saying that the difference disappears when whole class grade was considered. This will require rephrasing for clarification - are they saying only that throughout the clerkship R v. NR were the same even though individual clerkships looked different? I think there is a potentially significant finding here that is being underplayed that within each clerkship those who perform better are more likely to evaluate the experience even if in aggregate over the year there is no difference. ***This is a significant revision (p11 L8-12 and abstract p2 L17-19).”

Response 6: We thank the reviewer for raising this question. It made us realize that the message was not clear. In fact, in the voluntary group (which is the only group included in the table, as stated in the title), there are no significant differences between grades of respondents vs. non-respondents. Therefore, those who rated the clerkships and those who chose not to rate have similar academic achievement (as shown in the table; for example respondents in the Internal Medicine clerkship had non significantly lower grades by comparison to non-respondents: 77.51 vs. 79.34). Therefore, Table 3 presents non-response bias in the voluntary group and does not involve the compulsory group. On the other hand, Table 2 includes grades of students who participated in the voluntary group, that were significantly higher than grades of students in the
compulsory group (79.53 vs. 78.76, p=0.001, Table 2). Therefore, grade difference we talked about is not between respondents and non-respondents of the voluntary group; instead, it is between respondents in the voluntary group and respondents in the compulsory group. This raises a bias of baseline academic achievement between the 2 cohorts and we should correct for it. Yet, since the compulsory group (year of 2015) included the whole class while the voluntary group (2014) included only respondents, comparison of grades would be meaningless and biased if we do not compare whole classes. When we did such comparison including the whole class, there was no difference between grades. We agree that we should make this clearer. For this reason we re-organized the whole Discussion section and more specifically we made a clarification to this comment by including the following paragraph to the section on page 14:

“Interestingly, the ratings were comparable between the voluntary and compulsory groups except for one clerkship, and therefore, affecting the willingness of students to participate in clerkship evaluation using authority did not yield different ratings. One possible confounder in this comparison was students’ academic performance. Students in the voluntary group had significantly higher aggregate grades than in the compulsory group. To adjust for this confounder, we compared grade averages of the whole class between the two cohorts and found no statistically significant difference.”

Comment 7:

“Consequential validity evidence p12 starting on L4. Need to explicitly state that this is the final cohort with the sham question for clarity. This finding is probably underpowered to detect a significant difference and would have been very interesting to have a large sample size.”

Response 7: We agree with the reviewer. A clarifying sentence was added to this section: “Authority-induced bias was tested only in the final cohort where the sham question was added to the clerkship evaluation form.” We also agree with the low power to detect a difference and a sentence about this was added to the limitations section, as stated above.

Comment 8:

This comes from a Lebanese medical school following the US medical school training model. These specifications may raise a question about cultural bias and whether the findings can be abstracted to other training environments. However, this is a minor point that should be included in the limitations but overall does not degrade the quality of the work.

Response 8: We thank the reviewer for this point and we included this remark in the limitations section, which now reads: “A second limitation is related to the small sample size and
monocentric model (single institution experience) of the study, which may introduce a cultural bias and hence limit the generalizability of our results.”

Comment 9:

P2L3 Recommend changing "reliable" to "useful" or similar word in the background section of the abstract. A major aim of the paper is to assess evaluation reliability so would not use it here (see p12 1st paragraph of discussion where authors make this same argument).

Response 9: We agree and changed the word to “useful”.

Comment 10:

P2L13 Change "bogus item" to Sham question

Response 10: this was also changed.

Comment 11:

P2 L22-23 Too strong a statement given the limited data. "tendency toward" may make it more accurate.

Response 11: We totally agree and we made the suggested change so the sentence now reads the following: “Testing for authority-induced bias revealed that students might complete the evaluation randomly without attention to content.”

Comment 12:

P12 L 22-23 & P13 L1 Sentence reads awkwardly and could be rephrased for clarity.

Response 12: We re-phrased the sentence and removed expressions that are non-essential for the understanding of the manuscript. The sentence now reads: “Given the high-stake use of these ratings, interpretation should be made carefully, considering quality psychometric measures.”
Comment 13:

P13 L12-13 awkward phrasing "non-response bias is frequently non-significant in a homogeneous target population."

Response 13: We agree with the reviewer about the double negatives and we replaced “non-significant” by “can be ignored”.

Reviewer 2:

Comment 1:

“In fact, on p 9 the authors offer a hypothesis about compulsory response and bias with offering any information about directionality (hypotheses should always be stated directionally). It would be important to hear why low response rates would suggest unreasonably high or low ratings but this is largely unaddressed.”

Response 1: We agree with the reviewer that expressing expected direction of hypothesis is important. This has been added throughout the manuscript. In fact, we believe that using authority does not yield accurate information about the clerkship, which pushed us to test for attentiveness to evaluation content. We also agree that discussing the relationship between response rates and provided rating is interesting; however, literature is limited in addressing this aspect of evaluations, and existing data showed that increasing responses using incentives was associated with more favorable ratings toward the survey sponsor. Furthermore, a recent systematic review of factors influencing student ratings in UME did not find clear association between response rates and ratings. This information and the recent review were added to the discussion and to references.

Comment 2:

“The authors over-utilize the terms validity and reliability and sometimes misuse them. It would be so much more direct in presenting and discussing the findings to tell us what they say and mean rather than to constantly say that this makes the students' responses more or less reliable.”

Response 2: We thank the reviewer for this observation. We changed in many places these words by their direct meaning, whenever applicable. Examples include: utility and effectiveness of evaluations.
Comment 3:

“Beginning on p 8 the authors talk about AAPOR guidelines for response rates, etc. From what I can tell, they simply use number returned as the numerator divided by number distributed as the denominator, a simple mathematical calculation that appears pretty self-apparent. Why is complexity in terms of AAPOR guidelines relevant??”

Response 3: We agree with the reviewer that using AAPOR guidelines may add complexity to the reader. However, since the publication of the guidelines, and in light of the recent literature about using surveys, AAPOR has been recommended to be included in all studies involving response rates. We elected to include this to present a template/model making use of this standard of response rate presentation and increasing awareness about this reporting. Sometimes reporting response rate calculation may involve a different denominator (all students instead of eligible ones for example) and a different numerator (removing incomplete surveys for example). However, if the reviewer considers this to be complex for the readership, we can reconsider this rationale.

Comment 4:

“I am fascinated that the authors included the bogus item in the middle of their second round of questionnaires. Why did they not choose to include it in the first, for purposes of comparability? I didn't see any discussion, but did their students not comment and question this—this would be interesting for readers to know as a sort of sidebar. And, by the way, there is little or no reporting of this aspect of the paper highlighted in the findings.”

Response 4: We agree with the reviewer that administering the bogus question directly at the beginning of the year would have been more interesting. However, including the question at the end of the first clerkship would reduce our compulsory cohort to this group only. After the question is added and used in the evaluation, students who would have noticed it, may communicate this to their peers, which makes the inclusion of future evaluations useless (students may become more attentive just because they were informed about the question). Therefore, we allowed a certain time before administering it to have a larger number of evaluations in the compulsory group. This information was added to the Methods section, page 9, lines 13-16. Moreover, after we administered the question, we made an assumption about the bias introduced by students’ responses to this question. Confirming this assumption necessitated a second survey, as stated in the study. During that survey students commented about the presence of this question. Those who were biased did not notice it but those who were attentive to content felt very unreasonable to include the question and considered that it could be a mistake in the system. We added this note to the Discussion section on page 15, lines 11-13.
Comment 5:

“The fact that females are more likely to respond to surveys is well documented. This factor, like any other factor, is only important to the extent that females are likely to be more positive or negative in their ratings. Here again, the authors need to be more attentive to the directionality of bias and those factors that might introduce bias in one direction or another.”

Response 5: We totally agree with the reviewer and we added this important clarification to the Discussion section, page 16, lines 1-5.

Comment 6:

“The authors spend almost as much time talking about the factor structure of their questionnaire as other issues, and seem to want to focus our attention on the fact that the conditions of responding didn't affect this. Personally, I find this part of the story that this paper has to tell far less interesting and relevant.”

Response 6: We agree with the reviewer that this aspect is less interesting in our paper. We included it as a first step to demonstrate that our instrument is valid before it is used for high-stake. We reduced this part in the Results and Discussion sections to only pertinent information.

Comment 7:

“When the authors discuss authority-induced bias, they refer us back to the Methods section to remind us what they are talking about. They should describe what they mean briefly right there and then rather than send us back earlier in the paper for a reminder of what this term refers to.”

Response 7: We thank the reviewer for this observation and we revisited the hypothesis about authority-induced bias in the results section.

Comment 8:

“In the first paragraph of the Discussion, the authors summarize their main findings, and their third finding is that "...the compulsory approach might introduce..." The data presented either indicate that they do or do not—"might" is not appropriate here as a summary of what was found.”
Response 8: We agree with the reviewer about the indecisive nature of this sentence. In fact, our findings show that using authority was associated with a bias in about a third of the cohort, which may affect evaluation information. This has been now differently stated.

Comment 9:

“On p 14 in the Discussion session, a good deal of new literature is introduced. For the most part, this should have been raised in the Intro to anticipate the relevant issues, to be re-raised in the Discussion. Also, if new references are to be introduced at all, they ought to be presented in order to clarify the findings, but I did not find this literature functioned to throw light on the interpretation of data presented earlier in the paper.”

Response 9: We thank the reviewer for this interesting observation. We introduced most of the literature and important information in the Introduction section, we added new references, and we kept in the Discussion section only references that can help justifying our findings. The Discussion part was globally revisited and restructured to allow a smooth flow of ideas. We hope it is now improved.