Author’s response to reviews

Title: Mentored peer review of standardized manuscripts as a teaching tool for residents: a pilot randomized controlled multi-center study

Authors:

Victoria Wong (vwongmd@gmail.com)
Roy Strowd, III (rstrowd@wakehealth.edu)
Rebeca Aragón-García (ra2356@cumc.columbia.edu)
Yeseon Moon (ypm2102@cumc.columbia.edu)
Blair Ford (bfdord@neuro.columbia.edu)
Sheryl Haut (sheryl.haut@einstein.yu.edu)
Joseph Kass (kass@bcm.edu)
Zachary London (zlondon@umich.edu)
MaryAnn Mays (maysm@ccf.org)
Tracey Milligan (tmilligan@partners.org)
Raymond Price (Raymond.Price@uphs.upenn.edu)
Patrick Reynolds (preynold@wakehealth.edu)
Linda Selwa (lmselwa@med.umich.edu)
David Spencer (spencerd@ohsu.edu)
Mitchell Elkind (mse13@columbia.edu)

Version: 1 Date: 31 Mar 2017

Author’s response to reviews:

See attached Microsoft Word file for identical information but in an easier to read format.

We thank the reviewers for their helpful comments.
Reviewer 1:

It was difficult to assess the manuscript in some places, as the CONSORT checklist was missing. I would suggest that the authors submit the checklist with the revised manuscript, so that there is full information on the trial. For example, the sample size calculation was not addressed, so it is not clear whether the study had enough power to detect the effect of mentoring.

- CONSORT checklist completed and attached.

- The authors' understanding is that the CONSORT checklist is recommended for "the efficacy of healthcare interventions" as per the "CONSORT 2010 Explanation and Elaboration" document at [http://www.consort-statement.org/Media/Default/Downloads/CONSORT%202010%20Explanation%20and%20Elaboration%20(BMJ).pdf](http://www.consort-statement.org/Media/Default/Downloads/CONSORT%202010%20Explanation%20and%20Elaboration%20(BMJ).pdf). As our study focused on an educational intervention rather than a drug/device trial, and was also a pilot study, not all sections were applicable.

- This study was meant to be a pilot study from the very start. We worked with residents at only nine residency programs despite a high level of interest from many programs, since this was the number of program directors and IRBs that the authors felt it would be reasonable to work with in the context of limited funding, manpower, and other limitations. A sample size determination was not performed at the start since it would not have been feasible to work with a larger number of programs. See sample size calculation under reviewer comment #2 below for details.

- Reasons for "losses and exclusions after randomization" were not tracked stringently, again since this was not a medical intervention. We were working with busy neurology residents, and their primary reason for withdrawal from the study was that they were too busy with residency to continue their participation. "Additional File 3" lists some of these statistics. Likewise, there were no "important harms or unintended effects" to report.

- According to the ClinicalTrials.gov website, the registry is for "clinical studies on a wide range of diseases and conditions." ([https://clinicaltrials.gov/ct2/about-site/background](https://clinicaltrials.gov/ct2/about-site/background)) Ours was an educational intervention so we did not register our trial.

- All analyses performed were intentional primary or secondary analyses without further "exploratory" analyses since even secondary analysis data was based only on the contents of prepared survey questions.

- Lastly, please note that this study is, by the U.S. Department of Health & Human Services, exempt as human subjects research under their flowsheet: [https://www.hhs.gov/ohrp/regulations-and-policy/decision-charts/#c3](https://www.hhs.gov/ohrp/regulations-and-policy/decision-charts/#c3) (It was found to be exempt by some of our participating institutional IRBs whereas some of our participating institutional IRBs did not allow for an exempt status so in those cases, it was approved.)
I have several comments regarding the methodology and presentation of results - they are listed below in order of their appearance and not importance:

1. Abstract, page 3, lines 51-58: Significant results are only for secondary outcomes, ie. satisfaction with the intervention - this should be clearly stated here.

   - Added to abstract.

2. Methods: Description of important aspects of the trial have not been described in detail, so it is difficult to assess them, such as sample size, methods for sequence generation in randomization, mechanism used to implement the random allocation sequence, including who who generated the random allocation sequence, as well as who enrolled participants, and who assigned participants to interventions.

   - See explanation above regarding sample size and the fact that this is a pilot study. That said, here is the power calculation: Comparing the change in pre- and post-test between mentored and non-mentored groups using a two-sided t-test with alpha=0.05 would require a sample size of 332 with 80% of power to detect the mean difference of 5.4% (pooled STD of 17.5). And comparing RQI scores between equal size of two groups will have 80% of power with the sample size of 200 to detect the mean difference in RQI of 0.4 (pooled STD of 1.0) (based on manuscript5; Q8). This has not been added to the manuscript.

   - Again, as this was not a drug/device trial, we used a simple random number generator to allocate residents to two groups. We have added that the research assistant was the one who generated the random number sequence and allocated the participants to the two groups.

3. Methods, page 7, lines 1-15: The description of the manuscript used for peer review and the introduction of errors were not described in sufficient detail. Were these manipulated manuscripts pretested with experts to make sure that the errors are discoverable and logically introduced. This could be explained in more detail on the Additional file 2 (currently this document only states which items in a reporting guidelines were altered, but not what kind of alteration this was).

   - Manipulated manuscripts cannot be shared publicly for copyright reasons, though if the reviewer(s) or editor would like to see these manuscripts, we can share them. Manipulated manuscripts were reviewed by three co-authors, all of whom were neurologist editors (affiliated with the Neurology journal at the time of the study). We have edited “Additional file 2” to provide more detail.
4. Methods, page 6: I don't see the value of Figure 1 - it does not add important information, and the course is well described in the text.

- The authors feel that a visual timeline of the study is appropriate, as is commonly seen also in randomized controlled clinical trials. However, if the editors wish for Figure 1 to be removed, we are happy to do so.

5. Methods, page 7: Were the knowledge tests identical to those published previously in cited literature? If now, they should be provided as an additional material to this manuscript.

- Yes, the knowledge tests were identical to those published previously in cited literature, with all questions available in the referenced articles.

6. Methods, page 8, lines 21-45: The authors used means and standard deviations as measures of central tendency and variability, but then they use a non-parametric statistics for comparisons. Where the data normally distributed? If so, why then a parametric statistics was not used? The use of Cronbach's alpha for inter-rater reliability is not a usual measure - kappa statistics would be more appropriate, as Cronbach's alpha is usually used to test the reliability of an instrument. The values for inter-rated reliability should be reported in Results.

- Although t-test is robust against normality assumption, the Type I error rate could be substantially affected in the case of platykurtosis, which can be found often when the sample size is small. We used the non-parametric Wilcoxon rank sum test since we found that the sample size of each mentored and non-mentored groups were relatively small, and the distribution of some of variables demonstrated platykurtosis. Also as some scores were ordinal scale rather than continuous scale, we thought that non-parametric testing was more appropriate to compare the difference between the two groups. Nevertheless, as the reviewer recommended, we tested the difference using t-test to confirm the consistency and found similar results.

- We now use ICC as RQI is an ordinal scale in order to test reliability between two raters, and found that ICC was 0.7. We have made this changes in the manuscript accordingly, as requested by Reviewer #2.

7. Table 2: Although the numbers in groups are small, it would be useful to provide percentages for the groups and not only for the total column.
8. Results, page 12, line 46-48. The number of mentor-mentee meetings is presented as mean and standard deviation, but it would be more useful to present it as a median or as some categories, such as how many residents had more than 2 or 3 such meetings.

- Added to text.

9. Table 3: The headings for the first part of the table are missing.

The authors have addressed different biases of the trial; sample size and other methodological aspects that have not been described in detail should also be addressed in the revised manuscript. It would be useful to discuss what should be the focus of future studies, including methodological approach.

- Table 3 headers fixed. Methodologic limitations of the study have been added to the Discussion section including future directions.

Reviewer #2: RIPR-D-16-00030

1) This study is described as a pilot study but it is not clear why it was not conducted as a full RCT? Was the intention to conduct a full RCT and then the attrition was too high so you decided to report it as a pilot study? If so, this should be made clear. You say on p4 line 85 that participation interest was high, so it is not clear why this was a pilot study.

- See also the extensive response to Reviewer #1 above. This study started as a pilot study and had always been a pilot study because we were introducing a novel educational tool. There was no intention to complete a large-scale study. Participation interest was high in that 38 sites were interested, but it would not be feasible to coordinate this sort of preliminary educational study at that more than a small handful of residency programs, each with a small number of neurology residents. We chose to work with some of the largest residency programs to maximize our number of participants, and even then, working with nine IRBs and nine residency program directors was extremely challenging in the context of our limited resources. Attrition rate was disappointing, but our starting number was n=78.

2) The reporting of the study could be improved throughout in line with CONSORT eg eligibility criteria, description of intervention, how blinding was done, description of primary
and secondary outcome, justification for sample size, etc. More use of subheadings would make this easier to follow.

- We have completed the CONSORT checklist. Please see response to Reviewer #1 on why CONSORT is not entirely appropriate in this educational study that is technically exempt as human studies research. It is not a drug/device trial.

3) There are several mentions to the "feasibility" of mentored peer review, yet it is not clear how feasibility is being tested? Considering the attrition rates and results, one might question its feasibility. Even with "strong encouragement" to complete the test review, only 56% completed it (page 11 lines 194-6). There was no difference in change in knowledge or review quality.

- Mentions of feasibility have been removed.

4) Throughout the manuscript jumps around when describing the outcome measures and the focus tends to be on the positive results first rather than the planned outcome measures eg Abstract lines 61-62 on p3.

- Planned primary outcome measure now prioritized in the abstract and in Table 3.

5) The key assessment points are baseline and manuscript 5, however this is not immediately clear. In some ways the description of the additional reviews of other study designs is distracting. An improved description of the intervention under its own subheading and another on the outcome measures would help make this clearer.

- Additional subheadings added. The assessment points for RQI between Manuscript 1 (baseline) and Manuscript 5 re-emphasized under the Outcome Measures subheading.

6) You have inconsistently used different terms to describe some of the outcome measures and it can be difficult to follow which is being discussed at times eg page 4 line 77 "satisfaction" is mentioned, but there's no mention of satisfaction again. Line 141 "residents perception of biostatistics" versus Table 1's "impression of biostatistics"...... etc

- Terms have been standardized (e.g., "satisfaction" and "impression" changed to "perception").
Specific comments

Abstract

7) Tone down the conclusion - see my comments in Conclusion below
   - Addressed under the comments below.

Methods

8) Was number of errors reported an outcome measure? If no, why not?
   - The number of errors were only introduced as a way to standardize the manipulation of these manuscripts. We did not inform the residents that there were introduced errors, we did not use the EQUATOR Network reporting guidelines as part of our teaching curriculum, nor was there any plan to collect data on the number of errors that were "caught" by the residents.

9) There were nine study sites but it is not clear how many mentors there were? This number of sites and mentors for a small study is likely to have introduced some noise and inconsistencies in what was delivered. This should be discussed in the Discussion.
   - This has been added to the limitations section.

10) Page 6, line 108 - how were residents randomised?
    - We used a simple random number generator to allocate residents to two groups. We have added to the manuscript that the research assistant was the one who generated the random number sequence and allocated the participants to the two groups.

11) Pages 5-8 are quite difficult to follow. The use of more subheadings would be helpful and ensure that all the necessary detail is reported. I would suggest at least the following: intervention, blinding and allocation concealment, primary outcomes, secondary outcomes, sample size calculation, etc
    - Additional subheadings added. Sample size calculation excluded for reasons mentioned to Reviewer #1 above. The authors felt that all outcome measures could be described within a single paragraph due to the redundancy of this content.
12) Page 6, line 115 - the meetings were not mandatory (or they would have happened). Perhaps "highly encouraged" would be more appropriate?

- Meetings were presented as required as per protocol, so "highly encouraged" would not be precise. However, compliance was not enforced. This has been clarified in the text.

13) Page 6 - the five manuscripts included a range of study designs and as such any learning effect many have been diluted than if they had all been randomised controlled trials. This should be acknowledged in the discussion.

- The authors respectfully disagree, as the primary outcome measure was knowledge of epidemiology, biostatistics, and research methodology. Thus, a diverse range of study designs among the standardized manuscripts was appropriate as an intervention.

14) Page 6, lines 126-7 - Manuscripts 1 and 5 were both of the same design not just to allow comparison of RQI scores.

- The primary reason we chose manuscripts representing a single study design for Manuscripts 1 and 5 was to be able to make comparisons across similar study designs for the RQI.

15) Page 7 line 129 - state how many deliberate errors were introduced into the manuscripts.

- Done.

16) Were mentors encouraged to use reporting guidelines in their sessions with residents?

- No, the entirety of the "curriculum" is detailed in the manuscript and included as "Additional file 1." Reporting guidelines were felt to be beyond the scope of this early introduction to peer review.

17) Lines 134 to 155 - split this into different paragraphs for each assessment. Some of this will have occurred already eg the pre-test and post-test questions will be reported under primary outcome section - see my comments above.

- Done.
18) Line 141 - "residents perception of biostatistics" - I'm not sure what this means? Do you mean perceived knowledge? In fact lines 141 to 144 are not clear.

- Clarified. “Residents’ perception of their understanding of biostatistics, level of confidence about use and interpretation of statistical concepts, and application of scientific study results to patient care were also assessed…”

19) Line 146 - RQI is not an "objective" instrument

- Deleted “objective.”

Lines 157 - 167:

20) It is not correct to perform a statistical tests on the baseline data. If randomised properly, any differences should have occurred by chance.

- See #28 below.

21) only now is the primary outcome mentioned as change in knowledge

- Outcome measures now clarified in a separate paragraph

22) You have not adjusted for baseline knowledge scores. It would have been more appropriate to use ANCOVA methods to do this

- We agreed with the reviewer. We compared the change scores as well as post-test knowledge scores adjusting for pre-test scores, and made this correction in the manuscript.

23) not all the secondary outcomes are listed (lines 161-3)

- The secondary outcomes listed in the “Outcome measures” section match the secondary outcomes reported in the Results section under “Secondary outcome measures.” We obtained a lot of data but not all results were formally named as secondary outcome measures.
24) line 166 - It is inappropriate to use Cronbach's alpha for inter-rater reliability. This is a test for reliability but it tests for internal consistency. You should use an Intraclass correlation coefficient (ICC) for inter-rater reliability.
- As stated above, we used ICC in order to test reliability between two raters, and found ICC of 0.7. We have made this changes in the manuscript accordingly.

Results

25) When reporting the results it would be clearer to report under the subheadings seen in the Methods eg primary and secondary outcomes.
- Primary and secondary outcome measure subheadings have been created in the Results section.

26) Line 172 - remove reference to statistical differences
- We changed the wording to emphasize that the groups were similar at baseline, and removed reference to statistical differences

27) Line 174 "having previously had a mentor" would be clearer
- Changed to “Both mentored and non-mentored residents were likely to report previously having a mentor…”

28) Table 1 - remove last column (p value)
- Removed.

29) Table 1 - not clear what is meant by "impression of biostatistics"? There are several constructs thrown in here and I would break them down
"Impression" has been changed to "perception" as per earlier comment above.
30) Scoring details for Likert scales should be stated in the text for clarity in the sections on outcome measures.

- This has been clarified further in the Methods section.

31) Line 185 "Baseline perceived knowledge and self-reported confidence in …." Might this be clearer terminology?

- This has been changed to "Baseline perceptions of level of knowledge and confidence in biostatistics were well matched between groups."

32) Line 206 page 12 - what is "mentorship impression data"?

- Clarified in text. Changed to “An additional table reporting barriers to review completion and mentor meetings, as well as desire for mentorship is available in a separate file…”

33) Lines 208 - 210 - why were mentored residents more indifferent? I don't recall an outcome measure assessing indifference?

- It is mentioned in the Methods section that "Residents were also asked to provide reasons for difficulty completing reviews and to evaluate their experience in the study." This was not an outcome measure. The results of mentored residents being more indifferent are in Additional file 3.

34) Page 13 Table 3 - add the relevant column headings for each section

Report primary outcomes first.

- Done.

35) Page 14 - Lines 219- 224 - this is the reporting of the primary outcome measure and should come much earlier not thrown in at the end.

- Edited.
36) Line 225 - report the ICC not Cronbach's alpha

- We agree. ICC was used to test reliability between two raters, and found ICC of 0.7. We have made this change in the manuscript accordingly.

37) Figure 2 - what does "active" manuscript review mean? Completed?

Indicate baseline review

- "Active" manuscript review means that this group of participants was actively reviewing manuscripts. Baseline review is noted in Figure 2 as "Manuscript 1" and the figure is clear on which study participants withdrew from the study at which time points.

38) Was intention to treat analysis done or per protocol?

- Intention to treat analysis was performed according to randomization assignment at the time of enrollment. In other words, students randomized to the mentor group were analyzed against students randomized to the non-mentored group regardless of the number of mentor meetings or completion of mentored manuscripts. This was the planned analysis as per study design and protocol.

39) How was missing data handled?

- All participants completed the pre-test questionnaire. During the study, 15 participants requested withdrawal and were not included in the final data analysis due to lack of data for any of the primary objectives (Manuscript 5 review quality, post-test knowledge score, and end of study satisfaction). Enrolled and evaluable participants who did not complete manuscript 5, the post-test, or the post-test questionnaire could not be evaluated for the primary outcomes and were excluded from analysis of the primary objectives including review quality, knowledge, and satisfaction. For the remainder of the questionnaire responses, missing data was rare and when encountered, these participants were excluded and not imputed.

- For the primary outcome of the change in knowledge, we tested whether pre-test score (n=78) is associated with missingness on post-test (n=51), and found no evidence of systematic missing pattern (p-value=0.85).
Discussion

40) Line 229 - I am not sure that it does demonstrate feasibility? See comments above.
   - We have removed references to feasibility.

41) Lines 230-233 - put the primary outcome results first not perceived knowledge.
   - Edited.

42) Line 251 - the sample size may not have been large enough to show an effect
   - Added to study limitations.

43) Line 273 - how long were residents given to review each paper?
   - Clarified in the Methods section.

44) Line 270 - not clear how adding several study designs allows for a more balanced evaluation of review quality. You have stated that review quality was assessed at Manuscripts 1 and 5 which were both RCTs. I suggest you delete this statement.

If anything the introduction of several designs may have weakened any mentoring effect rather than concentrating on just RCTs.

This is a good point. As this was a pilot study, we included several study designs. Future studies of mentored peer review may wish to focus on one type of study only.

Conclusion

45) Lines 289 - 291 - this is an over statement and needs modifying (and in the abstract). See earlier comments.
   - This statement has been modified to the more muted “In conclusion, we used mentored peer review of standardized manuscripts to introduce the concept of peer review and to teach the principles of biostatistics and research methodology to neurology residents.”
46) Lines 291 to 295 - there is no mention of the primary outcome which did not show an effect.
- Added the modifier “Although primary outcome measure of content knowledge did not increase…”