Author’s response to reviews

Title: Does mentoring new peer reviewers improve review quality? A randomized trial

Authors:

  Michael Callaham (michael.callaham@ucsf.edu)
  Steven Green (steve@viridissimo.com)
  Debra Houry (Dhoury@aol.com)

Version: 2 Date: 19 July 2012

Author's response to reviews: see over
Response to reviews

Manuscript 3531037427094556

The authors are to be commended ...As a negative study that might discourage others from pursuing a similar strategy I have several areas of clarification I would like the authors to address.

We thank the reviewer for the positive comments.

1. Major Compulsory Revisions:

Was there any consent process for participation? If so could you describe it and if not explain why it was not required.

Yes, there was a consent process. *Annals* notifies every reviewer once a year that data on them and their reviews may be used for QI and research purposes, but that all results will be anonymized and no one will be identifiable in the results. If they do not wish to consent, they must notify the editorial office. A similar notice is sent with each individual review.

In addition specific consent was obtained for this particular study from two groups. The top reviewers from the journal were asked if they would be willing to serve as mentors; it was strictly voluntary and they consented in their reply stating they wished to do so. Reviewers who were randomized to the Mentor group were informed at the time of review that mentors were available and encouraged to communicate with them by phone or email as most convenient, but contact and usage was not mandatory (and several reviewers declined and were excluded from the data).

A shorter version of this information has been added to the Methods section.

2. How were the mentors trained. Was there any standardized instruction of the type of feedback to provide?

Mentors were not specifically trained as mentors by *Annals*, although all had been trained as reviewers in the past. One reason was that we were seeking to identify reviewer training methods that would require the least extra effort on the part of journals (and thus be more likely to be implemented). The second was we already had a group who had objectively demonstrated their consistent skill in review quality over
the years, in the form of the Top Reviewers who were ranked by review quality and timeliness. Thus their ability to perform the kind of reviews we wanted was not in question, and since they volunteered, neither was their willingness to help others. We did not provide any additional training in mentoring itself (nor are we aware of a generally accepted standardized method for doing so), but almost all of these reviewers were academic faculty who probably had experience and/or training in mentoring in an academic environment.

We have added clarification on this point to the Discussion.

Was the control group also able to access editors or others for feedback on their reviews. According to the survey of the participants it appears that may have occurred in some cases.

Communication between reviewers and editors on a manuscript does occur in a minority of reviews in routine journal operation, but very seldom involves any feedback on review quality (and certainly not specific commentary on how to improve it). We have done previous studies that showed that that type of feedback does not improve review quality. Those in the control group were not given any encouragement to consult anyone nor was any such resource mentioned. Some may have consulted other faculty on their own but we do not have data on this.

Was there any evaluation of the performance of the mentors in providing the mentorship?

The only and best evaluation could be provided by the reviewers who were mentored, who were all asked that question and the results reported on a scale of 1 (poor) to 5 (excellent) (Table 1). There was no validated evaluation tool available for assessing mentors and since the goal of the study was to examine a practical system likely to be implemented, interventions that required extra staff time were avoided.

3. The scoring system for reviewers results in a clustering of scores at 3, 4, and 5. How does the distribution of scores and the number of participants impact the ability of the study to detect a difference in scores between the participants. Although a sample size calculation was not done could you provide an estimate of what size of effect would have been statistically significant based upon the sample size and the distribution of scores among all reviewers. What percentile of current reviewers would fall within the range that would have been needed for statistical significance?
The reviewer asks regarding the "size of effect" for our negative study in order to retrospectively size up the power of our comparison. The confidence intervals for effect size were referenced in the original submission, but we have now added a more specific calculation to our results, and it demonstrates the range of potential error in our primary outcome such that readers can independently judge the likelihood that we might have missed a meaningful difference due to our sample size. (This size of effect was 0.1, with 95% CI of -0.4 to 0.6). Our statisticians tell us that such reporting of "size of effect" is preferable to a retrospective power calculation, which yields a simple number that is difficult to interpret.

We have further added this “size of effect” statement to the discussion.

4. Has there been any analysis of the reviewer scoring system for reliability?

Yes, it has been well studied as is reported in the manuscript (and in more detail and with more references) in reference 2, 11, and 13. It is the best studied such system in scientific publishing, and produces results virtually identical to the only other validated system (reference 12). This is referenced in the paper.

Minor revisions:

Although you provide the mean review scores for the initial 3 reviews, you do not provide that for the subsequent scores except in the Table. I think it would be helpful to also include the mean and median review scores for the subsequent reviews in the results and perhaps in the abstract.

We did calculate and report both, but in the interests of brevity and avoiding clutter did not report them again in the text, as that is the benefit of having tables. If the editors would like them repeated again in the text we would be glad to comply.

Discretionary Revisions:

It would be helpful to have more discussion concerning whether the mentorship process might be improved, standardized, lengthened or in some other way improved before being abandoned as a failure.

Thank you. Comment added in the discussion.
Reviewer Jatoi

Callaham and other have submitted an important ....

Thank you for the kind comments.

MAJOR COMPULSORY REVISION

The sample size numbers in the abstract do not seem to correspond to the numbers in the body of the manuscript. The abstract numbers are the same in the consort diagram. This reviewer may be misunderstanding things but would ask that the authors double check the numbers in the body of the manuscript.

I am not sure I understand this point, since on re-review all the numbers that are reported both in the abstract and the Results section are the same. Perhaps the reviewer was confused by the fact that the Results also report the findings of the survey of their mentoring experience filled out by the mentees, in which the numbers do (accurately) differ.

DISCRETIONARY REVISIONS

Some selection bias may have been at play to explain these negative findings. Any younger reviewer who decided to participate in the study was likely interested in critical thinking and the peer review process and hence that those reviewers who populated both groups were likely going to provide good reviews. Perhaps the authors would want to further expound on this point, or perhaps not.

First, reviewers did not choose whether to participate; every new reviewer during the study period was randomized to either control or mentee group. Once in the mentee group they could decline to participate, but this almost never occurred.

As our other published studies show, it is probably true that new (and usually younger) reviewers might be more motivated and produce better reviews. However our goal was to study ways to improve new reviewers (or had these results been positive, we could also have later tried applying them to longer term reviewers.) One could equally argue that new reviewers would be most likely to be amenable to mentoring and thus would gain more from the experience.

The manuscript provides a nice description of what was supposed to
happen between the mentored reviewer and the senior one. It falls short of telling us what really happened. This seems to be a shortcoming or at least a point of interest. Again, it may merit some comment.

Yes, this is a common failing of so many educational intervention studies. Unfortunately we had no way of capturing this information other than our survey asking the mentees about it (see Table). Further detail would have required some method of capturing and quantifying all the interactions, a difficult logistic task.

The authors describe that they had no baseline data for calculating sample size, but they did settle on a sample size. It would be meaningful if, at the very least, they would share their rationale for choosing the 4 year duration of the settled upon number of recruited subjects.

Please see similar comments from first reviewer.