Rebuttal Letter

Dear Peter:

Thank you for yours and the reviewers' feedback. We have now revised the manuscript following the feedback whenever possible. Our detailed responses are below.

Cheers,
Bob

REVIEWER 1

Re: Point 1 -- a question as to "How over hundred thousand students could have ended up providing less than 15 thousand evaluations"

We thought we were clear that our unit of analyses were course evaluations rather than evaluation provided by individual students. However, clearly, we were not. We have now revised the manuscript in several places to highlight that the unit of analysis were class based course evaluation rather than individual student course evaluations.

Re: Point 2 -- Not enough detail is provided for calculations (e.g., smoothing).

The revised manuscript now includes the details of the precise smoothing method used in Figure 1. The p-values attached to specific tests (t, r, RR tests) are calculated usual way and such calculations normally do not form part of the manuscripts.

Re: Point 3 -- "The authors confuse "probability" with "frequency"... and the claim that our data includes "a fixed census of data, no random sample, no experiment"

We do not confuse probability with frequency and we do not have "a fixed census of data" if that embodies a notion that we have evaluation of every single professor. We have a sample of course evaluation obtained by a sample of professors teaching these kinds of courses. The professors in our sample were sampled usual way for these type of studies: they applied, went through interviews, and some were selected from those who applied to teach at this institution and those who taught there at those times, and got their evaluations done, formed our sample. Clearly, we worked with a sample of professors rather than all professors even if the sample was selected by others (not by us). Equally clearly, our sample is not random. However, no other sample of participants in psychology or educational studies is random. Also, many studies rely on non-experimental designs.

Re: Point 4 -- "This is a story, not science"

There are several separate issues here:

1) Our study shows that in this sample of professors, those teaching math do more poorly on SET than those teaching say English. This replicates findings cited in the introduction, the findings obtained in different universities. In fact, one of them is based on hundreds of institutions being evaluated by ETS student evaluation system. Thus, we replicated the previous findings that math professors do more poorly than others. This is science just as any other science that uses non-experimental designs.

2) Question as to why math profs to more poorly on SETs than English profs is not answered directly by our data nor by any other prior data which we know of. Nor did we ever attempted to answer it. Our conclusions that math profs are far less likely to be labeled unsatisfactory, fired, etc. does not hinge in any way on the reasons for the difference in the SET ratings.

3) The crux of this article is that claims that factors such as teaching math vs. English explain only 1% of variance, and thus are "not important" as was concluded in the previous research is based on inappropriate analyses and irrelevant effect sizes. When the same data are analyzed appropriately and when appropriate effect sizes are used, the difference between math and English professors is substantial and has real, huge impact on these professors being labeled satisfactory vs. non-satisfactory. Moreover, such impact depends on the specific standard used by any given institution.

4) The issue of fairness. The reviewer argues that the differences could "be due to real differences in teaching, not to subject matter per se." Agreed, it could be, even if we personally believe this is extremely unlikely given a well-established evidence that students do not like to take quantitative course, that numerical literacy is on a steep decline, etc. etc.. Moreover, the basic principles of fairness (and various codes of testing) dictate that validity of a measure as well as validity of the measure in different context (quantitative vs. non-quantitative courses) must be established before the measure is used. If it was not done, the common standards should not be used. And this goes to the next, most critical issue:

5) The validity of SETs. The reviewer then goes on and says that it is impossible "to measure teaching effectiveness" and so we cannot figure out if these two groups are engaged in equally good teaching. Well, if we cannot figure it out, and especially if we have no evidence that we can actually measure the quality of the teaching, perhaps the ONLY fair solution is to evaluate a person teaching a particular

subject against those person teaching that same subject. That way we know where they stand relative to those who try to do the same job. And corollary is that to evaluate math teachers against English teachers is unfair UNLESS we can show that math profs are really truly engaging in “bad teaching”. The revised manuscript now includes more extensive discussion of these issues.

REVIEWER 2

Re: Points 2 and 6: Different ways to analyze the data.
We do not have access to data that would allow us to do meaningful analyses using multi-level modeling.

Point 5: There could be differences in a number of students, class sizes, type of course, that could account partially for the differences in the ratings among disciplines.
Indeed this is a possibility. However, even if this was the case, the fact remains that many universities compare one professor’s SET to averages for the department and/or university, regardless of class sizes, response rates, etc..