Dear Editor,

Please see below our responses to the changes requested.

Sincerely,

The authors

Editor Comments

Thank you very much for your Stage 2 submission of “Evaluating the pedagogical effectiveness of study preregistration in the undergraduate dissertation: A Registered Report”. I have read your submission with great interest and I have a few revisions I would like to see before I share this with the reviewers.

My main comment is about how to interpret the results given the smaller sample size where the effect sizes fall below the detectable range. In the Stage 1, it was determined that there would be the ability to detect an effect if the effect size of np-squared was 0.04 or greater (for the two-way interaction between group and time) and a d of 0.40 or greater (for the focal pairwise comparison between preregistration and control groups at time 2). In the Stage 2, the sample size was smaller than originally planned and the sensitivity analysis showed that the detectable effect sizes increased from an np-squared of 0.04 to 0.10 and the d changed from 0.40 to 0.66 or greater. The effect sizes you found (np-squared: 0.001-0.05) appear to be lower than the minimum detectable effect size (note that I was not able to find any d effect sizes reported in the results). Therefore, the probability of your detecting any effects is very low. As such, your findings of no correlations or a correlation should be discussed in terms of your inability to have the power to detect a difference (or not) and not that no difference/a difference was found. I suggest adding a sentence about this each time you report a np-squared effect in the Results section, as well as discussing this specific issue in the Limitations section (note that it is a distinct issue from what is currently discussed in the Limitations). You could add that, with this smaller sample size, you were not able to detect moderate effects, only stronger effects, of which, none were found. It could be useful to translate the effect sizes into actual differences in the measured variables to give readers an idea of what kinds of differences between the groups equate to what kinds of differences in the effect sizes (e.g., there would need to be a difference in 2 points on a 5 point scale for a moderate effect, and a difference of 3 or more points on a 5 point scale for a strong effect).

Response: Thank you for this – we agree, and have now gone through and added sentences throughout the Results to make the point that null results could be due to an inability to detect a significant result rather than the absence of one: (e.g., “Note that given our smaller sample than anticipated and the sensitivity power analysis, the null results here may reflect an inability to detect differences rather than the absence of an effect (see *Limitations*).”)

We also discuss this explicitly in the limitations section too (p. 13)

. This means that we were only able to detect stronger effects rather than moderate effects, of which none were found. Therefore, it is possible that null results reported here were owing to an inability for us to *detect* significant effects with our smaller than planned sample size, rather than the absence of a true effect. Therefore, future research should aim to conceptually replicate our findings with larger sample sizes that are better equipped to detect smaller effect sizes.

Please put the study design table back in the main manuscript document and add an optional column on the right stating the outcome. The full tracked changes version of the manuscript needs to be available at the link provided in question 2 of the Report Survey. Please also add line numbers to make commenting during the review process more efficient and clear. I include a few other minor comments below.

Response: We have added the study design table back and included a column of outcomes. We also now include all tracked changes and have added line numbers throughout.

Minor comments:

1) Perhaps the title should be updated to delete “registered report” because it is now a Stage 2?

Response: This has been removed.

2) Starting in the Abstract and Intro, you changed the term from “statistics anxiety” to “attitudes toward statistics”. I think the former is more informative to the reader about what the term means. However, if you consider competence, value and difficulty as attitudes around statistics, then the broader term is more fitting. If you define “attitudes toward statistics” on your first use, then it would be clearer what the term means.

Response: We have added a brief definition of statistics attitudes in the introduction. We feel this term is clearer because some of the constructs e.g., “value” do not necessarily align with anxiety, and so the term ‘attitudes’ feels more appropriate.

3) Page 14: “we could reliably detect an effect size of np2= .10 for the Group*Time interaction and pairwise comparisons of d=/> 0.66 with 80% statistical power” add “, which was higher than planned.” per your Stage 1 manuscript.

Response: This has been added.

4) Please provide a justification for why you changed the 11 point Likert scale to 5 points, as well as evaluating 3 rather than 6 components (page 19).

Response: The COM-B was measured using a 5-point scale, but at Stage 1 we proposed to use an 11-point scale. This was due to an oversight in the building of the survey. The number of components did not change – we measured 3 components of the COM-B (i.e., capability, opportunity, and motivation) and each component had two items, as per the materials registered in Stage 1. This has been clarified

transparently in text in the Stage 2 manuscript (p. 17) and does not impact the interpretation of the COM-B results.

This measure contains 6 items, where two items address each of the three components of the COM-B on a 5-point Likert scale ranging from 0 (*Strongly disagree*) to 5 (*Strongly agree*). Note that the 5-point scale is a deviation from our Stage 1 Registered Report, which proposed to use an 11-point Likert scale. This deviation was due to researcher oversight in the building of the Qualtrics survey.

We have also transparently noted this in the reporting of the COM-B exploratory results (p. 5-6).

Note that we proposed to measure the COM-B on an 11-point Likert scale at Stage 1 and deviated to a 5-point scale at Stage 2. This does not impact the interpretation of the results but does mean that variation (i.e., the standard deviations reported here) is likely to be lower than if we had used a broader scale.