

Dear Corina,

Thank you again for your comments. We have made the two changes requested. We hope you agree that the manuscript is strengthened by this.

Kind regards,

The authors

### **Response to Comments**

Thank you for providing clarifications on how your Stage 1 and previous Stage 2 version show that the reduced sample size was not a deviation. However, the issue of making it clear which of your results were below or above your power threshold is still there. I think the confusion comes down to two things:

- 1) “This means that we were only able to detect stronger effects rather than moderate effects, of which none were found.” It is unclear whether “of which none were found” indicates that you found no strong effects in your results, therefore you were unable to detect any effects in your study, or whether it indicates that you only had strong effects and no moderate effects were found, which would indicate that you had the power to detect your strong effects. Clarifying this sentence would help resolve the issue.

**Response:** Thank you – we understand the confusion here. We have amended the text as follows:

This meant that instead of being able to detect effect sizes of approximately  $d = .40$  for the pairwise comparisons of interest, we were able to detect effect sizes of  $d \geq .66$  with 80% power (i.e., medium-large effect sizes). This means that we were only powered enough to be able to detect medium-large effects. Therefore, it is possible that null results reported here were owing to an inability for us to detect smaller significant effects with our smaller than planned sample size, rather than the absence of a true effect.

- 2) It would make it much clearer if, in the Results section, you added text after the presentation of each result to note whether the effect size was larger or smaller than the threshold for which you were able to detect effects at 80% power.

**Response:** Thank you for this suggestion. You will see that we have now gone through and added text to all results which indicates this 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*). Note that all of the effect sizes here are small, and thus all fall below the threshold for which we were able to detect an effect size based on our sensitivity power analyses at 80% power (see *Limitations*).~~

...

However, beyond the NHST results, the Bayes factor here also lends support for the null result. Note also that the effect sizes here all fall below the threshold that were able to detect according to our sensitivity analysis (i.e., moderate-to-large effects with 80% power).

...

The Bayes factor here indicates a substantial difference, which lends strong support for the hypothesis. This effect size is also larger than the threshold effect size that we were able to detect based on our sample size and sensitivity analysis. We did not, however, detect a significant main effect of preregistration Group  $F(1, 87) = 1.726, p = .192, \eta p^2 = .019, BF_{10} = .587$ , but a significant Time\*Group interaction  $F(1,87) = 4.663, p = .034, \eta p^2 = .051, BF_{10} = 1.751$ . The effect size of the interaction was also larger than our threshold effect size according to our sensitivity analysis (at 80% power).

...

The effect sizes for opportunity and motivation were both comfortably beyond the effect size threshold that we were powered to detect, according to our sensitivity analysis (at 80% power). Students who preregistered also reported significantly higher capability to preregister ( $M = 4.09, SD = 1.042$ ) compared with those who did not ( $M = 3.51, SD = .96$ ),  $t(87) = 2.64, p = .009, d = .57, BF_{10} = 4.466$ . Although, this effect size was smaller than the effect size we were powered to detect.