Geosci. Commun. Discuss.,
https://doi.org/10.5194/gc-2020-36-AC2, 2020

Interactive comment on ““Thanks for helping me find my enthusiasm for physics!” The lasting impacts “research in schools” projects can have on students, teachers, and schools” *by* Martin O. Archer and Jennifer DeWitt

Martin O. Archer and Jennifer DeWitt

m.archer10@imperial.ac.uk

Received and published: 22 November 2020

This is an excellent paper, and describes important, and thorough research into the effects of an engagement programme - something which is notoriously hard to do. My comments are largely very minor.

We thank the reviewer for their time and comments.

Section 3: In point 5, the authors mention reliability. It would be interesting to know how reliable they found the codes (i.e. what were different between the first and second coder, and was it significant)

Overall there was 92% agreement between the two coders, which corresponds to a Cohen's kappa of 0.836 (Cohen's kappa is unity minus the ratio of observed disagreement to that expected by chance, hence ranges from 0 to 1). Disagreements were resolved by discussion to arrive at the final coding presented in the paper. We will add these points to the paper.

4.1.1: (sentence 2) "Additionally they were asked to reassess their confidence before having undertaken the project." I found this ambiguous - reassessment implies a second assessment, when the text mentions that there was no pre-assessment. Perhaps "retrospectively assess" would be more accurate?

We thank the reviewer for this suggested wording, which we will now use.

Figure 1: I very much like the figures in the article, and they are all clear. With this figure, it may be worth considering applying some transparency to the points as some are overlapping. I am, however, willing to believe that this makes it too confusing, but it is something the authors should consider.

We have tried the reviewer's suggestion, but found it makes things less clear. The main point of Figure 1 is that almost all the points lie in the upper triangle, indicating a positive effect, which is clear. The precise locations of each datapoint are not so crucial in this context.

Table 4: (caption) if 11 weren't placed in dimensions, it may be clearer to say using n=52 of 63 responses?

We thank the reviewer for this suggestion, which we have adopted.

Figure 4: what intervals are the error bars (1-sigma, 95%?). It is worth noting that these intervals are not reliable with small n or low probabilities of success. I suggest at least an acknowledgement that these should be treated at indicative given the sample size.

The error bars represent the standard (1 sigma) confidence interval using the Clopper and Pearson method. This was mentioned in the caption and discussed further in the methods (lines 111-114), explaining that they are a conservative estimate based on the exact binomial distribution. Therefore, the error bars do not rely on the normal approximation, which is known to be unreliable for small n or low probabilities. We will further clarify in the captions of figures that the word “standard” refers to 1 sigma.

Section 4.3: in the final paragraph there are details of (as yet) unpublished works which name the "anonymised" schools. Does this risk de-anonymising the schools used here, if the results are published in the future?

The reviewer raises a good point. The reason the schools’ pseudonyms were added in this section were so readers could check the type of schools (i.e. independent or state, high Free School Meals etc.) to show that these outcomes were not biasing to privileged schools. However, to further protect anonymity we will comment on these aspects within the text rather than providing the pseudonyms.

Section 5.2: The authors use a null hypothesis of 2. Would a better quantitative test to simply be to code positive vs negative, without the division into planned and definite? (i.e. give definite a score of 2 as well for this purpose, with a null hypothesis of 1.5)? At present the null hypothesis is that 2/3 of respondents claim a positive impact.

The reviewer’s suggested test is less strict than the one adopted in the paper that they refer to. The one-sample Wilcoxon signed-rank test tests whether the median is significantly different from a hypothetical value, as explained on lines 117-120. Applying the reviewer’s suggested test gives unilaterally smaller p-values.

We also note that a test of positives vs. negatives was also already performed (lines 502-503). The reason behind a null hypothesis of 2 in the later tests was that the “I will” response might be construed by some as neutral, therefore potentially biasing the positive results. Our analysis has thus taken both interpretations into account. We will

add this reasoning to the paper as follows:

> We acknowledge some may consider the “I will” response as neutral and thus our analysis takes both interpretations into account.

Figure 5: I found the grey error bars hard to spot, as they are narrow and overlap the black error bars. Perhaps thicker lines and/or offset horizontally with respect to the black error bars?

We have made the grey error bars thicker.

Interactive comment on Geosci. Commun. Discuss., https://doi.org/10.5194/gc-2020-36, 2020.