Author’s response to reviews

Title: Testing an Active Intervention to Deter Researchers’ Use of Questionable Research Practices

Authors:

Samuel Bruton (samuel.bruton@usm.edu)
Mitch Brown (MitchellBrown@usm.edu)
Donald Sacco (Donald.Sacco@usm.edu)
Ralph Didlake (rdidlake@umc.edu)

Version: 2 Date: 14 Oct 2019

Author’s response to reviews:

Dear Dr. Kowalczuk,

We would like to thank you and the two reviewers for your helpful and constructive feedback regarding our manuscript, "Testing an Active Intervention to Deter Researchers’ Use of Questionable Research Practices" (RIPR-D-19-00021R1). We are especially pleased that you found merit in our initial submission and have invited us to submit a revised version of the manuscript that addresses the reviewers’ concerns. We believe the result is a much improved paper.

Most of the reviewers’ suggestions have been incorporated into the revised manuscript. However, some points we thought better addressed in this letter. Nonetheless, we remain open to revising our manuscript further to accommodate changes you or the reviewers request. Below, we provide our responses to the original reviewers’ feedback and indicate where changes can be found in the revised manuscript. Since the reviewers provided comments in paragraph-form, we have detailed our responses to each of their concerns in list-form, which we hope simplifies the task for you and the reviewers.

Reviewer 1

This is an interesting randomized intervention study that assesses the effect to a writing exercise of 3-5 minutes (aimed at consistency motives activation versus control) on the defensibility of and the willingness to engage in 15 questionable research practices (QRPs). Although the text is well-written the methods and results sections are difficult to follow due to a number of issues that need to be solved.

RESPONSE: We appreciate the reviewer’s initial support for our manuscript and hope to have addressed each of these legitimate concerns sufficiently.
It takes the reader a lot of effort to sort out what the primary outcomes are and how these are measured.

I was confused about the number of main outcomes: two (defensibility and willingness to engage) or five (additionally: impact, rationalization and risk). The abstract and method section made me think two while in the results section five outcomes are presented. Only when consulting S1 and S2 it became clear that the first two mentioned are both primary outcomes and that the additional three are (probably meant to be) secondary outcomes. Please clarify.

RESPONSE: We apologize for the lack of clarity. The reviewer’s assumption that defensibility and willingness were the primary outcomes of interest is correct. The latter three, impact, rationalization and risk, are peripheral outcomes we found prudent to consider, based on previous findings indicating an association between defensibility and the impact, rationalization, and risk outcomes (Sacco, Bruton, & Brown, 2018). Given constraints of the word limit in the abstract, we found it most appropriate to report the two primary outcomes in the abstract because of their centrality to the hypotheses. Nonetheless, we have provided a more detailed description of primary and secondary variables in the manuscript (pp. 7-8). We had envisioned our consistency manipulation potentially reducing the ethical defensibility of QRPs by making them more difficult to rationalize. However, because the outcome measure of defensibility was not impacted by the consistency manipulation, we were unable to test for this relation.

S1 and S2 are also the only source that explain what these five outcome variables are: single items with 7 answer options on a Likert scale. How the five outcomes variables are calculated I could not find. My guess is that the numerical value of defensibility and willingness to engage are the means of the scores of these items over the 15 QRPs presented. And the numerical values of impact, rationalization and risk are probably the mean scores on the 3, 3 and 6 questions that together form these three (secondary?) outcome measures. Please clarify.

RESPONSE: This reviewer is, once again, correct in assuming this is how we calculated these scores. We have revised the manuscript to explain the aggregation process (pp. 7-8).

No information is provided on the psychometric properties (reliability, validity, responsiveness, comprehensibility) of these outcome measures, with the exception of some reliability estimates from earlier work (study 1 and 2) on the three outcomes contained in S2 (although it's unclear for which of the three these Cronbach's alpha's (is that what they are?) refer). Please clarify and expand.

RESPONSE: The defensibility measure is a previously validated measure to assess perceptions of QRPs in research scientists (Sacco, Bruton, & Brown, 2018), which was additionally correlated with the willingness measures and the three peripheral outcomes to identify a level of convergent/divergent validity with these items in that original study. Additionally, the peripheral outcome measures were designed to be face-valid items to assess the construct to which they were intended. We did not collect data on responsiveness and comprehensibility measures in this study, though. The alphas we reported in these studies were indeed Cronbach’s, thus providing the reader a reliability analysis to confirm the consistency between items. We have reported more information about the psychometric properties of these measures in the revised manuscript and indicate the alphas reported are Cronbach’s (pp. 7-8).
The main message (‘this is a negative RCT’) is buried in an confusing avalanche of subgroup analyses.

The confusion arises from the fact that the effect of the randomized contrast on the (3 or 5) outcomes is not clearly presented.

RESPONSE: We apologize for presenting our results in a confusing capacity and have reworded our manuscript to ensure better clarity in our findings.

The lengthy results section instead presents an enormous amount of subgroup analyses and makes too much fuss about some borderline cases of interaction, many of which have little or nothing to do with the randomized contrast that is the primary determinant studied. Please move all this to a digital supplement. That would make the main text much more clear and substantially shorter.

RESPONSE: We agree with moving these subsidiary analyses that are not germane to the intervention-related hypothesis to an online supplemental file (S4). The revised manuscript reflects this change. In the spirit of full disclosure and transparency, we were hesitant to omit this information from the original draft, even though we agree that it made the manuscript cumbersome. However, moving the recommended material to a supplementary file is a sensible means of ultimately retaining all content without cluttering the manuscript and its primary message.

In fact the RCT is mainly analysed as an observational (non-randomized) study. Please stick to the superior straightforward approach:

RESPONSE: We have made our analyses more straightforward so as not to have too many subsidiary moving parts that are largely irrelevant to the hypotheses.

- Confounders (other determinants (like gender) other than the primary determinant studied (the contrast of two writing tasks)) are only interesting for adjusting the effects of the primary determinants if needed

RESPONSE: We argue that consideration of gender itself remains interesting and could inform future interventions that could resonate better for men and women if necessary. Our inclusion of a gender analysis was admittedly exploratory, to some degree, but this was in the service of identifying how men and women may react to interventions differently as a reflection for future research. This all being said, we recognize that the central message of this paper is the intervention and have therefore moved the gender analyses online.

- Effect modifiers (interaction), if convincingly present, is only important to decide whether it's necessary to report the overall effect of the primary determinant for subgroups.

RESPONSE: We have removed unjustified subgroup analyses from the manuscript proper to the online supplemental material for the sake of a transparent reporting. We additionally utilized more cautious language in this reporting.
I believe that the results should be presented as Mean Differences with corresponding 95% Confidence Intervals. That's much clearer than Means and SD per group followed by 3 statistics. Furthermore, I would recommend to present the results as much as possible in a table and to only highlight the main findings in the text.

RESPONSE: We have reported Mean Differences 95% confidence intervals in-text alongside our inferential statistics. We attempted to streamline the reading process as much as possible, which makes the necessity for tables in the revised manuscript less of a concern, as each outcome variable is now presented in a clearer capacity.

Please refrain from statistical spin by labelling non-significant findings as being marginally significant.

RESPONSE: We have refrained from describing effects as marginally significant. In certain places, we have retained the language of “directionally consistent” and “trending” with hypotheses to help readers determine the extent to which the pattern of results was consistent/inconsistent with study hypotheses, even if not conventionally significant.

Some other major issues are:
Major study limitations are not acknowledged, e.g.:

- Attitude is a notorious bad predictor of behaviour (actually engaging in QRPs).

RESPONSE: We have addressed this limitation in the manuscript (pp. 17-18).

- The power at subgroup level is (very) low and it's highly likely that a substantial proportion of the reported statistically significant findings are false-positives.

RESPONSE: Given that we have revised what was reported in the manuscript, this concern is attenuated to some extent; much of the potentially underpowered findings were likely moved to the supplemental materials online and are explicitly described as exploratory in that capacity. For the exploratory analyses with narratives, we agree that several of these findings could be underpowered. To this end, we urge caution to readers in the revised manuscript in their interpretations (pp. 17-18).

- The difference between the two writing tasks may be too small and maybe it's a bit unrealistic to expect an effect from 3-5 minutes of writing.

RESPONSE: This is a fair point, but this decision was ultimately informed by empirical research previously demonstrating that such amounts of time to write are normally sufficient to render salient a given motivational state (e.g., Bernstein, Sacco, Young, Brown, & Claypool, 2010; Keefer & Landau, 2015); we have clarified this point in the revised manuscript. Nonetheless, we realize that the immersion time could have been insufficient and suggest future research considers longer manipulations that are more immersive (e.g., Murray, Kerry, & Gervais, 2019).
The absence of blinding (according to the study protocol in the digital supplements no blinding was attempted) may introduce bias that further diminishes the eventual effect. Maybe a randomized consent design would have been a better alternative.

RESPONSE: Blinding was not much of a concern in the current studies, given that the online link we provided participants utilized a randomizer divorced from researcher interference to remove the possibility of any experimenter effects on who received which condition(s). We explain this more clearly in the manuscript (p. 8).

The only limitation you list is not really important: the beauty of RCTs is that the selectivity you mention will hit both randomized groups equally. Consequently it's not a threat to internal validity, although it can compromise extrapolation (external validity or generalizability).

RESPONSE: We have removed this limitation from the manuscript.

The conclusions drawn (as reported both in the Abstract and the Conclusion section have very little to do with what was actually studied. I strongly recommend to stay closer to your data: no effect was found and it may be that the writing task aimed at consistency motives activation has the unwanted side effect of rationalizing the engagement in QRPs. A comment may be added that this might explain why RCR courses have disappointing little effect and sometimes even an effect in the wrong direction.

RESPONSE: We have adjusted our manuscript to report these findings first and foremost.

I took me a while to understand what you mean by themes 1 - 4. My impression is that you divided the respondents in subgroups based on a typology (the themes) you developed in the qualitative analysis of the texts that resulted from the writing tasks. Under the heading 'Exploratory Narrative Analyses' you present some quantitative results for 2 of these 4 subgroups. I could not find HOW you came to these 4 themes, how many participants ended up in these themes, and why you left out 2 of them in this section. Please clarify and explain and add information or the methods and results of you qualitative data-analysis. You could also consider to move this to the digital supplement or a future article.

RESPONSE: We have provided a more detailed description of what we mean by themes by describing the narrative research process in greater detail, which then leads to us discussing how came to the find these specific themes in our own work (pp. 9-10). We additionally have included the number of participants in each theme (p. 10). To the point of why we left two themes out of extensive analyses, the basic effects did not yield anything significant and were therefore not elucidated upon in the detail the other two were. Although we originally provided this rationale in the first manuscript, this was stated more clearly in the revised manuscript (pp. 10, 11).

Some minor issues are:
The links to OSF don't work.

RESPONSE: We have fixed this issue.
It's rather puzzling why two sources are utilized to recruit participants. Please explain and reflect on the potential differences between these subgroups.

RESPONSE: We utilized two sources in the service of optimizing the generalizability of our findings to even more research scientists beyond what could have been accomplished by only sampling from one source. Statistically, these two sources of participants did not differ from each other and it would therefore be inappropriate for us to reflect too much on this effect; we indicate this clearly in the revised manuscript (p. 6). Indeed, sampling on a single population would have unnecessarily limited the generalizability of our findings. Because we had access to both samples, utilizing a multiple populations strategy and finding that population does not moderate the findings increases the generalizability of our findings.

It's unclear how many were approached to get the 200 participants. Furthermore, no information is given on compliance with the writing task and completion of the survey items. Please add this information, preferably in a flow chart.

RESPONSE: We sent out mass emails every other day over the course of a month until we reached 200 participants in which there were approximately 200-300 PIs emailed in each individual wave. This point was clarified in the manuscript. Additionally, all participants who completed the study complied with the writing task by writing an honest answer to the prompt, thus making issues with compliance seen in other priming studies a non-issue (this point is clarified in the manuscript, p. 6).

QRP 4 suggests that rounding off a p value from .044 to .4 is wrong, but what you probably mean is then when the p-value is .44 it's wrong to say it's <.4.

RESPONSE: There may have been a miscommunication on this point and we apologize if this is our doing. The item was rounding a p-value of .044 to .04. This item was designed to reflect journal changes for authors to report more precise p-values than what was acceptable in previous years (e.g., Journal of Experimental Social Psychology). That is, rather than reporting out to the hundredth, it is more appropriate to report out to the thousandth to disambiguate effects that could have been barely significant using NHST standards.

Reviewer 2
The paper clearly touches upon an interesting avenue for empirical research, namely the question how to deter questionable research practices among scientists.

To begin with, I think you are underselling the problem of QRP's in your introduction. While some consider the replicability crisis as overblown, others argue that especially the use of QRP's (as opposed to outright fraudulent behaviour) is extremely damaging to the professions. As such, I think you could make a much stronger case as to why we should find measures to reduce the use of QRP's. In fact, in part the use of QRP’s might be related to inadvertent mistakes and misunderstandings, which cause a high prevalence and a systematic problem.

Fanelli, D., Costas, R., & Larivière, V. (2015). Misconduct policies, academic culture and career stage, not gender or pressures to publish, affect scientific integrity. PloS one, 10(6), e0127556.

RESPONSE: We agree with this reviewer about bolstering our arguments on the problems of QRPs. Both papers the reviewer mentions were cited in the original version of the manuscript; we have cited the John article more prominently in the revised introduction. Rather than the 2015 Fanelli article, the Fanelli 2009 (also cited in the original manuscript) better makes the reviewer’s case, and we have cited it additionally in the revised introduction. (pp. 2-3).

It would be great to discuss the shortcomings of the study in more detail. You study researchers with NIH/NSF funding. Yet, incentives to cheat might be higher for early stage researchers (those without funding) to advance their careers. The question is, how would success as a researcher and/or quality of the researcher affect your findings and the efficacy of the intervention?

RESPONSE: This is an interesting point and one we discuss in the revised document. Specifically, we have addressed this point in our own work by considering the age of respondents in an intervention study. The relevant paper is currently under peer review elsewhere, but we have discussed this point clearly in this revised manuscript (pp. 17-18).

Also, I am wondering about the theoretical implications of your study. Related research has shown similar results as people try to defend their self-concept. It would be interesting to discuss your results in light of these prior findings:


RESPONSE: This is an excellent point and we have addressed it in the revised manuscript. The Mazar et al. article referenced by the reviewer was already cited in the originally submitted manuscript; we have referenced it again in connection with the results (pp. 16-17).

Evidently, the intervention did not work in your context. Given the underlying focus of the narrative it might be interesting to extend the research along these lines. Rather than asking individuals to think about how they could or should incorporate ethical behavior, it could be interesting to let them think about the consequences of potential unethical behavior. For example one might speculate that imagining detrimental societal effects might hinder QRP’s or thinking about the potential risk of being detecting and exposed as a "bad scientist" might deter QRPs. Speculating how other interventions could look like would strongly benefit the research here. It is interesting to view the discussion in light of your findings, but the failed intervention comes up a bit short in this section.

RESPONSE: This is a valuable suggestion we now consider in the section on limitations, as it holds promise for a follow-up study. (p. 17)
I do believe that the author(s) should continue with this important line of research and with appropriate changes in the introduction, the embeddedness into the related literature, and an extended discussion this might result in an interesting research article. I wish the author(s) good luck in pursuing their research further and hope they find my comments constructive in doing so.

RESPONSE: We appreciate the vote of confidence this reviewer gave our findings and we hope to have sufficiently addressed these legitimate concerns.