# Editor Feedback

From: **Advances in Methods and Practices in Psychological Science** <onbehalfof@manuscriptcentral.com>

Date: Fri, Oct 21, 2022, 3:32 PM

Subject: AMPPS - Decision on Manuscript ID AMPPS-18-0074.R4

To: <Hallbf23@gmail.com>, <hmoshontz@gmail.com>

21-Oct-2022

Dear Braeden, Hannah, and Team:

Thank you for submitting your completed RRR Stage 2 (post-data analysis) manuscript ID AMPPS-18-0074.R4 entitled “Registered Replication Report: A Large Multilab Cross-Cultural Conceptual Replication of Turri, Buckwalter, & Blouw (2015)” to Advances in Methods and Practices in Psychological Science (AMPPS). We sent the final paper to the original author and to expert referees for input, and their comments appear at the end of this email. Their comments and suggestions, together with my own read, necessitate some revisions to the manuscript.

This is a unique Registered Report for a number of reasons: it involves the CREP, it took place over a long period of time (including pandemic lockdown related disruptions), and the editorial team has now changed hands (so I am not the person who accepted the Stage 1 version of this paper). These factors make decision making more complex than we might like at this stage, but we will try to muddle through and do our best.

An initial concern is that many changes have been made to the manuscript in the interim between Stage 1 and Stage 2. The authors are thanked for their transparency in disclosing and explaining the changes. Broadly, (as you did) we can lump the changes into (1) changes to the introduction/manuscript, (2) design-based changes, and (3) analysis changes. Let’s consider each set in turn.

First, regarding changes to the introduction, I appreciated that you included the original text as an appendix for readers to consult (which is great for transparency), we generally discourage such changes for a few reasons. Certainly, we want to avoid changing the framing of the paper after results are known (i.e., HARKing). Although the changes here are extensive, they do not seem to ultimately change the framing of the paper dramatically. However, a related concern is that the version of the paper that reviewers signed off on is not the same as the one they are reading now. Additional new issues can (and have) been raised. You will see in the reviews (in particular from R1 and R4) that there are some disagreements and points of clarification raised. These issues raise a bit of a conundrum. Ordinarily, I would not ask authors to revise their introduction to address reviewer concerns at Stage 2 of a Registered Report. However, given that there have already now been changes to the introduction, it is difficult for me to assess at this point which additional changes (if any) are most warranted. Therefore, as a general point of guidance, I would like you to address each reviewer's concern in your response letter, whether or not you decide to make a particular change, justifying your decision. I would like you to refrain from making changes that revise the spirit of the original introduction and introduce HARKing. I would like you to make changes that increase the clarity and transparency of the manuscript. I will call out one of them raised by most of the reviewers, which involved clarity in your description of what constitutes a Gettier case. As an outsider to this research area, I do not feel qualified to judge your definition of Gettier cases, but it seems there is some consensus among the reviewers that the paper is muddy on this point. Therefore, please pay particular attention to this area.

|  |
| --- |
| **Response**: The intended changes to the introduction from Stage 1 to Stage 2 were made to increase the clarity of the writing. Unfortunately, several paragraphs were deleted accidentally during the paper formatting process and the complete introduction was not included in our Stage 2 submission. We apologize for this error. This omitted material has been added back in as part of our edits (see highlighted text in manuscript on pages 7-19).  Adding this material back in has addressed some of the reviewer comments and made the scope and content of the introduction very similar to the accepted Stage 1 version of the manuscript.  We have also carefully considered the reviewers feedback regarding what constitutes a Gettier case; however, we have decided not to change our original definition that was reviewed and approved during Stage 1. Unfortunately, this framing from Stage 1 was part of the introduction text that was accidentally deleted from the submitted Stage 2 manuscript. Hopefully, the reinclusion of this vital information will clear up this point for reviewers, even if they may ultimately disagree with how we defined the term during Stage 1. |

Second, regarding changes to the design, you point out five changes in your cover letter. These issues are especially tricky, because we cannot go back in time and undo them. The first three do not seem like they would have likely affected results and can be mostly ignored. I may have missed it, but I mostly do not see reviewers commenting on the fourth change (changes to measurement, removal of covariates), so perhaps these changes are not as drastic as they might seem. To me, these look like quite substantial changes (e.g., splitting the luck vs. ability question into two parts). If there is not already discussion of these changes in the manuscript (e.g., in discussion section), please add reflection on these changes there. If it is there, please point it out for me (please). For the fifth change (exclusion criteria), I will discuss that issue under analysis plan.

|  |
| --- |
| **Response**: We have added clarifying information about these particular decisions throughout the manuscript.  We chose to drop two test setting site level covariates (online vs. in person and individual vs. group testing) variables because the COVID-19 pandemic changed (and significantly limited) how students could carry out their replication studies. After it began, our data collection was shifted to almost entirely online (and individual) participation. As can be seen in Table 1 of the manuscript, most sites had online and individual sessions, some of the sites had both session types for one or both of the two variables, and some sites were missing documentation. Thus, using the covariates as intended would have been impossible. No other planned covariates were removed from the analyses.  The luck vs. ability measure was originally a single question that required two responses. We altered how the question was displayed to alleviate participant confusion. We changed the question presentation so that participant responding to the first part (i.e., “Darrel got the [right/wrong] answer”) could determine wording and response scale for the second part (e.g. “Darrel got the right answer because of his: ability < — > good luck”) rather than relying on difficult to understand response anchors (i.e., the original second response used “ability/inability” < — > “good/bad luck”). Notably, the measure itself and tests involving it were planned and reported as exploratory. Thus, any changes to it are arguably acceptable.  As requested, discussion of dropping the two test-setting covariates and changing the luck vs ability measure have been added to the Limitations section on page 80. |

The third and final set of changes involved changes to the analysis plan. Again, the changes here were nontrivial. In particular, a continuous variable was dichotomized, and the resulting analysis strategy was changed. The data exclusion plan was not followed. As a general approach, I would like you to complete the analysis as originally conceptualized, to the best of your ability. You note that there were some factual errors in the original plan (e.g., regarding specification of the MLM), and you can/should correct those errors. However, beyond those changes to correct errors, please present a few versions of the analysis, which you can think of as varying on two dimensions: logistic vs. linear and original vs. new exclusions. Combining these two dimensions yields four models: (a) the version that is as close as possible to the preregistered plan (linear model, continuous measurement), and uses the original exclusion strategy, (b) your ideal version (logistic model, dichotomized measurement), but using the original exclusion strategy, (c) the version that is as close as possible to the preregistered plan (linear model, continuous measurement), and uses the new exclusion strategy, and (d) your ideal version (logistic model, dichotomized measurement), including using the new exclusion strategy [i.e., d is what is in your current draft]. If there are other variations on the analysis that you feel would be informative, please feel free to include them. To be clear, the purpose of including these many variations is to assess whether semi-arbitrary choices affect the outcome of the study. I think it makes sense to interpret the test that is closest to the one proposed as the ”confirmatory” test, and the others as additional robustness checks. Please label them as such, and if necessary, please move details of robustness checks to supplementary materials. As a final comment on the analysis strategy, please pay particular attention to Reviewer 3's reflections on these issues, as I suspect they will prove valuable.

|  |
| --- |
| **Response**: In terms of data exclusions, only two deviations occurred. First, we decided to exclude people from the analysis if they reported an age over 100, assuming that these entries were made in error and should thus be considered missing rather than accurate. Only 78 out of the 9,440 participants who took the survey were excluded due to this change. The other 2,040 participants who were excluded for age were excluded because of missingness - which is not technically a deviation from our original analysis plan. We would have always needed to remove participants from any regression analyses that were missing data on variables included in the planned models on a model by model basis, and since age was entered first, this effectively removed them from all analyses. However, 2,070 of those 2,118 participants who were excluded because of age would have also met at least one of the other exclusion criteria.  The only other deviation that occurred was that we originally said that we would *not* include data from student teams that did not complete all of the pedagogical steps in the Collaborative Replications and Education Project (CREP) process (e.g., submitting a final report of their site’s findings). All teams completed all steps with the exception of submitting the final report. Since we do not report on or discuss those site level analyses in this paper, we decided that we would still include data from labs as long as they were approved to run participants (i.e., submitted video of protocols, ethics approval, etc.) and used the SoSciSurvey survey that we had 100% control over. This way we could ensure that participants from those sites were exposed to the correct study protocols, despite the teams not finishing all the pedagogical tasks required of them to complete the CREP process. We have added clarifying information regarding these details on page 35.    The specification analysis you requested further supports our inclusion of the data from these teams. As shown in <https://osf.io/nvfbm>, the overall model decisions were the same, regardless of inclusion criteria. The only difference is that one of the four proposed model versions (i.e., the vignette by condition interaction for the reasonableness dv using linear modeling) did not show the interaction due to exclusion criteria. This model had less power because the exclusion of CREP teams that did not finish the final analysis step dropped the sample size by 1/4th.  As for the modeling linking function choice, we examined the data assumptions required for linear modeling (additivity, linearity, homogeneity, normality). No multicollinearity was present. The rest of the assumption checks show serious violations of all tests, indicating non-linearity, heteroscedasticity, and non-normality. This result is highlighted by showing all assumption checks for all four models together in <https://osf.io/nvfbm>. Given the most serious violation of linearity, we chose to dichotomize the data, as the residuals suggested a sigmoid function akin to logistic regression. The previous methodology on this research is bimodal in nature, which further supported these decisions. The logistic models did not violate these assumptions, and met assumptions for logistic regression. The most discernible differences when examining linear versus logistic models include the effects for interactions in the sample source analysis (MTurk vs. not MTurk) for the reasonableness and knowledge dependent measures, wherein there was slightly more sensitivity to detect an interaction. Examination of the patterns of results indicates the same general results with slightly weaker effects for MTurk data.  We appreciate the idea of “sticking to the plan” but have followed the spirit of a pre-registration - we noted where and why the plan didn’t work in our manuscript (with updates per your suggestions). Given the new supplemental analyses, readers can review if they think these changes mattered. The manuscript presents the model we believe is most appropriate given the statistical assumptions and data inclusion practicalities. See pages 38-40 for the inclusion of this information and justification of analyses choices. |

Reviewers 1 and 4 both raise an important point about whether your findings “fail” to replicate Turri et al.’s null finding or not. You carefully address this issue in the discussion section, which I appreciated, and I am inclined to agree with both the reviewers and with you that the findings are not substantially discrepant from Turri et al. In spite of the difference in statistical significance (significant vs. not), the effect sizes you obtained are remarkably similar. The issue is how to frame or address this issue in the paper more broadly. You might recall that debates on this very issue came up surrounding the Reproducibility Project. Some papers on this topic might prove helpful in considering the issue further (<https://doi.org/10.1371/journal.pone.0149794>, <https://doi.org/10.1126/science.aad91>; <http://datacolada.org/47>). The crux of the issue is that original studies, because they are often extremely underpowered, are well-nigh unfalsifiable. Their confidence intervals are so wide as to be consistent with nearly any follow up result. This is a tricky issue to navigate, but you may be better off describing your findings as support for Turri et al.’s hypotheses (or other scholars’ hypotheses, if more appropriate), rather than as a failed replication of their null result. I would advise you to revise the discussion (and abstract, and elsewhere as you find it) to reflect this conclusion, which does seem to more accurately describe your findings. Do your best to highlight the similarity in effect size across the original and replication studies.

|  |
| --- |
| **Response**: We appreciate your recognition of our efforts to address the issue of whether our findings replicate those of Turri et al. We have carefully considered your suggestion and have revised the discussion, abstract, and other relevant sections of the manuscript accordingly to more accurately describe the nuances of our findings.  We have further emphasized the similarity in effect size across the original and replication studies in the revised manuscript to highlight the consistency of our findings with Turri et al.'s (see pages 66-67, 70).The low power of the original study likely explains the differences in our statistical conclusions, though you are correct that larger or smaller effects would likewise have been consistent with the original effect confidence intervals. We also point out that our findings do contradict their claims that knowledge attributions were insensitive to Gettier-Knowledge condition differences; from this standpoint, our results better align with other studies that have demonstrated Gettier intuitions. We believe that this framing of the results best captures our findings and their implications.  See also our responses to the reviewer comments below. |

When you submit the revised manuscript, again, please explain how you have addressed each of the reviewer comments (or why you've chosen not to address them) and describe the changes you made to the manuscript. I reserve the right to send the paper out for review again, though I will strive to avoid it if possible.

|  |
| --- |
| **Response**: Please see below for our inline responses to the reviewers and descriptions of any corresponding edits.  We thank you and the reviewers for your thorough feedback which has improved the manuscript considerably. As the original Stage 2 submission was erroneously missing several sections from the introduction, we completely understand if our revision needs to be sent back out to reviewers. |

Please also remember to check the style of your manuscript and make sure that it complies fully with AMPPS guidelines: <https://www.psychologicalscience.org/publications/ampps/ampps-submission-guidelines>.

To submit your revision, log into <https://mc.manuscriptcentral.com/ampps> and enter your Author Center, where you will find your proposal title listed under "Manuscripts with Decisions." Under "Actions," click on "Create a Revision." Your manuscript ID number will be appended to denote a revision.

IMPORTANT: Your original files are available to you when you upload your revision. Please delete any redundant or outdated files before completing the submission.

I look forward to receiving your revised manuscript.

Sincerely,

Katie Corker

Associate Editor, Advances in Methods and Practices in Psychological Science

Reviewer comments:

# Reviewer: 4

Comments to the Author

This paper reports an attempt to replicate Turri et al. 2015. Before saying anything else, I should note that I reviewed the earlier registration of this MS. My overall take is that this research is technically sound. The deviations from the original registration do not undermine my confidence in the conclusions, and I think the findings are credible despite below requesting some additional details and analyses. At the same time, my sense is that the contribution to knowledge provided by this research is not substantial and that the claim to not fully replicate Turri et al is misleading. I'll elaborate on these matters below.

|  |
| --- |
| **Response**: Thank you once again for providing us with this helpful feedback in improving this paper. We have appreciated your contributions during both stages of this project. Please see below for our responses to your points. |

1. Overall, I think the Introduction does not establish the importance of this research. I commented on related points when reviewing the original registrations, and if anything the case for the present research may be weaker than before (in the registration). The review of previous findings spanning pp. 3-4 is inadequate. Findings are more complicated than this review suggests, particularly because there are some kinds of Gettier problems where people generally do attribute knowledge. Also, this paper investigates judgments about "barn country" cases, and although some researchers and philosophers view theses cases as a kind of "Gettier case", my sense is that this isn't universal or even the dominant view among philosophers. So although the paper frames itself as concerning Gettier cases, this isn't clear cut, and at best it's only true in a limited way.

|  |
| --- |
| **Response**: The intended changes to the introduction from Stage 1 to Stage 2 were made to increase the clarity of the writing. Unfortunately, several paragraphs were deleted accidentally during the paper formatting process, and the complete introduction was not included in our Stage 2 submission. We apologize for this error. This omitted material has been added back in as part of our edits (highlighted for your convenience on pages 7-19). We agree that the version of the paper you received did not clearly establish the importance of this research, but we hope that you feel differently about the corrected version of the introduction. We also understand that what constitutes a Gettier case is not well agreed upon. We acknowledged this problem and were intentional in how we defined the term in the Stage 1 manuscript, but this section was unfortunately part of the one that was deleted in error during the submission process. We hope that, despite perhaps not agreeing with how we operationalized Gettier cases (i.e., counterfeit-object cases like the “barn country” variety), you find our argumentation clearer with the reinclusion of this information. |

2. The description under “Measurement considerations” is not very clearly written. The reader has no idea what the difference was between the Gettier story used by Turri et al. and the control. Also, although the Introduction describes a follow-up study as finding a difference where Turri et al. found none, this in itself isn’t meaningful unless (a) the effect size of the follow-up study is given and (b) the meaning of the difference is discussed. Given enough participants almost any two conditions may come to diverge. The important question isn’t whether they differ at all, but whether the magnitude and direction of the difference tells us something interesting. The discussion on pp. 13-14 doesn’t tell us anything about this.

|  |
| --- |
| **Response**: Information about the conditions and specific findings in the original study was also part of the section that got deleted (see pages 10-11). We hope the reinclusion of this section resolves this concern. We have also edited the “Measurement considerations” section to increase clarity, including your recommendation to provide more details about the referenced follow-up study’s effect size and relevance to the present work (see page 16). |

Also, given mixed findings in the field, it’s difficult to see why findings diverging from those of Turri et al. are important. It's not even clear from the present discussion whether the original Turri et al findings are representative of the overall pattern from other studies. And it makes little sense to broadly refer to previous papers on Gettier cases, since again, the cases considered here and in Turri et al. are of just one specific sub-variety. As was true with the original registration, I'm puzzled about why this project was pursued at all (besides as a potential pedagogical tool). It's never been clear how this project might advance our understanding of people's conceptions of knowledge.

|  |
| --- |
| **Response**: We appreciate your highlighting this problem. We believe this concern has been addressed by adding back in the missing pieces in the introduction that make it clearer that we are focusing on this one specific sub-variety of Gettier-type cases. Similarly, we have added specifying language in the discussion to ensure that we are not over-generalizing to all Gettier-type cases and to better situate our findings in context to the literature regarding counterfeit-object Gettier-type cases (see pages 66-70).  This study was chosen through the Collaborative Replications and Education Project (CREP) study selection process, which relies on students’ ratings of interest in and feasibility of studies that were published 3 years prior to the selection process. The CREP’s goal is to engage students in real world research in a way that gives them valuable experience with the scientific process while providing the field with much needed replication studies. While the original Turri et al. (2015) paper may not have been ground-breaking, students chose this study to be the most interesting and most feasible study to replicate out of a list of dozens of studies that were published in 2015 across ten different subfields of psychology. Given that CREP replication projects are designed for and carried out by students, this project could not exist without making important pedagogical trade-offs between student engagement/accessibility and scientific rigor. |

3. Moving to the Methods, almost all sites recruited undergraduate participants which was not the sample tested by Turri et al. To know what the present findings mean, it is important to know what proportion of these participants had recently taken philosophy courses where they might have learned about Gettier cases. This would be an especially serious concern if the students were philosophy majors. It would also be helpful to know if the findings hold up if analysis is restricted to community samples, again because this group will more closely resemble the sample tested by Turri et al.

|  |
| --- |
| **Response**:  In a direct replication of Turri et al.’s procedure, participants completed a funnel debriefing in which they answered questions about the study’s purpose and their prior participation in similar studies. As part of our exclusion criteria, we removed participants who correctly guessed the study purpose as part of our exclusion criteria. Notably, this number was very low (*n* = 203, *n* = 16 in which this was the only exclusion). We also removed participants who indicated that they had previously participated in a similar study. While we did not ask participants about their familiarity with Gettier cases or whether they have taken philosophy courses, we believe that these exclusions would have removed participants who recognized that the study was related to Gettier cases.  Our study included an MTurk sample comparable to that collected in the original study. We tested to see if the sample source interacted with condition in predicting both reasonable and knowledge judgments. We did not find evidence of a difference in results between the community sample (MTurk) and the mostly student sample. See page 53 in the revised manuscript for a full description of these analyses. |

4. The exclusion rate, particularly resulting from failure of Comprehension questions is somewhat concerning. Whereas Turri et al. excluded only about 11% of participants for failing comprehension question (a bit low for MTurk in my experience) the current study excludes approximately \*half\* their sample for the same reason if I understand things. This raises the question of how many included participants scraped by via mere luck. A question: were the “n =1490 missing” participants among those included?

|  |
| --- |
| **Response**: We agree that the number of participants excluded due to comprehension failure was unfortunately and unexpectedly high. As discussed in the paper, a few features of the research may explain this difference in part. First, Turri et al. exposed participants to only a single vignette in a between-subjects design. In our study, participants each responded to three vignettes in a within-subjects design. Given that responding to multiple vignettes is more fatiguing than responding to one, participants may have been more prone to make errors in comprehension than in prior studies. Second, as pre-registered in the approved protocol, we excluded participants if they missed any of the three comprehension questions (i.e., one per vignette). This decision increased the overall exclusion rate. However, as can be found in <https://osf.io/nvfbm>, only excluding observations when the comprehension check for that given vignette was incorrect did not substantively change our results (more on this in our response to 11 below).  As we state in the paper, our exclusion rates are consistent with some cross-cultural Gettier intuition research (e.g., rates between 21% [Machery et al., 2017b] and 47% [Machery et al., 2017a]).  As to participants passing comprehension checks due to luck, our methods reduced the likelihood that participants were included in analyses because they responded correctly to the check questions by chance. The probability of passing one comprehension check by chance (50%) is much higher than the probability of passing three comprehension checks by chance (12.5%). So, if anything, our high exclusion rates raise the question of how many participants in past studies using these same vignettes were included in analyses by chance despite their lack of comprehension.  1490 missing participants were excluded because they had missing data on one or more of the comprehension questions; we report the *n* missing for each exclusion on pages 28-29. |

5. Page 6 notes that some data, contrary to the approved protocol, was gathered via surveys programmed by students. This section should note the total number of participants contributed by these sites (not just the number of sites). It should also state whether the results are the same if excluding these sites, and any differences in the student-programmed surveys than the main one (i.e., this shouldn’t have to be something that readers dig into themselves).

|  |
| --- |
| **Response**: We have added further clarity to the main manuscript about these sites by including the sample size for the final data from Qualtrics and a footnote explaining our test for differences (see page 31). A set of multilevel models examined if the data source (Qualtrics versus SoSociSurvey) interacted with the experimental condition in predicting knowledge, reasonableness, and luck judgments. No interactions were found in these analyses, which can be viewed at<https://osf.io/nvfbm>. Therefore, our decision to combine the data into one large dataset after matching variables did not impact results. |

6. Table 1 already packs in a lot of information, but it’d be good to know how many included (and possibly excluded) participants came from each vignette and condition. It's important to know if some vignettes, or some conditions from vignettes, were more likely to yield exclusions than others. Perhaps an additional Table could provide this information. Also, given that different vignettes can give different results, was this incorporated into the power analysis? That is, was there sufficient power to pick up on differences across vignettes.

|  |
| --- |
| **Response**: Thank you for this suggestion. We have added the number of correct comprehension question responses for each vignette and condition combination to page 30 as Table 4. However, it is important to note that exclusions were generated at the participant level, not at the condition/vignette level. If a participant missed one item, they were excluded from all analyses; thus, this table simply illustrates whether different combinations created different correct response rates.  The power analysis conducted as part of stage 1 included the assumption of complete data for each participant across vignettes to correspond to our exclusion criteria. |

7. The Methods aren't sufficiently clear about the order in which vignettes were presented. I thought this was random (based on the registered report from before), but when we later read about the number of participants who completed the Darrell vignette first it doesn't look so random. Also, in the survey, were questions presented in a fixed order, or was this randomized or varied? Also, did the questions appear on the same screen as one another, or were they on different screens. These details are extremely important for interpreting the results. I believe Turri et al. used a fixed order with different questions on different screens, with the Knowledge question presented first (removing the possibility of order effects).

|  |
| --- |
| **Response**: As footnoted on page 55, the Darrel vignette analysis only excluded participants who answered the Darrel comprehension check incorrectly. The number of participants per vignette-condition combination in the primary analyses was approximately equal before and after exclusions (see Table 4).  Please refer to the procedure section on page 36, where we have made some minor edits for clarity about randomization and question order to increase clarity. Participants were randomly assigned to both vignette and knowledge condition, and all questions were presented in the same order for each vignette (i.e., the order as was used by the original authors, followed by exploratory questions). Questions were also presented on separate screens, just as in the original study. |

8. Page. 12 reports that contrary to what was outlined in the original report it was decided to not exclude data from teams “who did not submit their work for final CREP review”. More information is necessary here (i.e., how much certainty do we have over what participants actually saw and did), and it may be important to know if the findings replicate with data from these teams excluded. (It’s also hard to understand the relation between the site-specific analyses, and the main analyses reported in the MS).

|  |
| --- |
| **Response**: We have added clarifying edits in regards to including data from student teams that did not complete all of the pedagogical steps in the CREP process (e.g., submitting a final report of their site’s findings; see page 37). We decided that we would still include data from labs who did not receive a completion certificate as long as they were approved to run participants (i.e., submitted video of protocols, ethics approval, etc.) and used the centrally programmed SoSciSurvey survey. Thus, we are confident that participants from those sites were exposed to the correct study protocols, despite not completing the more pedagogical tasks required of them to complete the CREP approval process.  As per the request from the editor, we have examined the differences in our results given the original exclusion criteria as shown in <https://osf.io/nvfbm>. The exclusion criteria did not change overall model decisions, and only one difference in results was found: for the reasonable dependent variable, the vignette by condition interaction did not appear in one tested model, as the number of participants was dropped by 1/4th by using this exclusion criteria (i.e., must finish CREP requirements). We now include this information on page 37 and discuss this change as a potential limitation on page 81. |

9. It’s surprising that the continuous data had to be treated as binomial. As an alternative, wouldn’t it have been possible to treat the data as ordinal?

|  |
| --- |
| **Response**: Ordinal level analyses would require that we create specific thresholds to the data for the rankings (i.e., 0-9, 10-19, etc.) in order to analyze vignettes, conditions, and continuous predictors showing different thresholds in the data. Unfortunately, any choice as to where to apply these cut points would have been arbitrary. The decision to use a binary outcome was suggested by the data and previous research that used a forced binary choice. |

10. Although it’s claimed that the results from Table 6 do not replicate Turri et. al., this is based on Turri et al., failed to find a significant effect. But as the paper later acknowledges, this is probably just a matter of power, since Turri et al. had a relatively small sample size. When we look at the actual results of the two experiments, they're basically identical: In Turri et al., the difference in knowledge attribution between Knows and threat is 14% (i.e., 81- 67). In Table 1, the corresponding difference is also 14% (i.e.,13.88 = 56.59-42.71). That is, the magnitude of the difference is the same. So, I think that if anything, the present findings \*do\* replicate Turri et al. The conclusion that they do not rests on focusing on differences in patterns of statistical significance. But this is trivial given that (again) Turri et al. used a small underpowered sample. It would be better to based this conclusion on replication-versus-failure on statistical comparisons: Do the original findings substantially differ from the present findings if they results are directly compared?

|  |
| --- |
| **Response**: Thank you for bringing up this important point. You are correct that our conclusion that we did not replicate the findings of Turri et al. is based on differences in statistical significance rather than differences in the magnitude of the effect. As explained in the discussion already, we likely found a significant difference where Turri et al. did not because of our study’s increased power. In fact, we note the similarity in the odds ratios of the original study to those in our replication (see page 67). However, we agree that this point should be more clearly highlighted in the paper. We have revised the manuscript accordingly (see pages 66-70).  Notably, while the condition differences in our research were similar in size to those found by Turri et al., our conclusions are clearly discrepant. Our results do not support their claim that, “a salient but failed threat to the truth of a judgment does not significantly affect whether it is viewed as knowledge”(p. 381). |

11. The analysis that focus on participants who got the Darrell vignette first only exclude participants who failed comprehension questions for that vignette. Could this approach be taken in the main analysis? i.e., use data for vignettes where participants passed comprehension question.

|  |
| --- |
| **Response**: We included the exclusion criteria of removing participants who answered any check question wrong for a reason related to your previous comment about participants being able to pass the comprehension questions because of luck. In the original paper, participants had a 50% chance of guessing correctly on the comprehension question. Because we employ multiple vignettes, we were able to decrease the chances of participants being able to guess correctly from 50% to 12.5% (50% x 50% x 50%). While participants may have failed to comprehend one vignette while attending to and understanding the other two, we preregistered the exclusion of participants who failed any comprehension check, assuming that this would capture more guessing behavior than the original study.  In response to this question, we repeated our analyses using the data for all vignettes with correct comprehension question responses with the other implemented exclusion rules (i.e., minimum/maximum age, language, previous experience, etc.). These results can be viewed at <https://osf.io/nvfbm>. The results were the same across analyses except in the models examining the effects of sample source (MTurk vs. not MTurk). For the knowledge and reasonableness analyses, the condition by data source interaction was significant. These interaction effects likely became significant because of the increase in power with more observations. However, the pattern of condition differences was the same for both sample sources–the size of the Gettier-ignorance condition difference just differed slightly according to source. |

Discussion

12. The paper suggests that order effects did not matter, "Further, given that participants were presented with vignette-condition combinations in random order, any contextual order effects would have been minimal." I'm unconvinced. To know whether order matters, this needs to be tested. Alternatively, just as the paper reports tests for participants who completed the Darrell vignette first, it could do the same for participants who completed the other two vignettes first.

|  |
| --- |
| **Response**: Thank you for these suggestions.  We tested for order effects (see <https://osf.io/uz8te>: Exploratory Analyses > Order Effects), and we found significant effects of vignette order and condition order individually. There were no interactions between condition and vignette order or condition and condition order; thus, the condition differences did not vary based on order. The main effects appear to be driven by a large sample size with small effects in the difference choice rates of the dependent variable, as shown below. We’ve edited our discussion of potential order effects to include these findings (see page 68).    As suggested, we additionally examined the Gerald and Emma vignettes when they were the first responded to by participants. Page 57 indicates the following results: these effects were similar for the Gerald vignette when presented as the first vignette (same effect size and pattern) and the Emma vignette (same pattern, half the effect size).  Thank you again for your helpful comments and feedback. |

# Reviewer: 1

Comments to the Author

I applaud the research team on all of their hard work in completing this complex international multisite investigation and registered replication of knowledge attribution in lucky Gettier-style scenarios. Below I raise several questions, concerns, and suggestions for additional analyses before publication that might potentially strengthen the manuscript.

|  |
| --- |
| **Response**: We appreciate your helpful feedback and have provided responses to each of your questions, concerns, and suggestions below. |

(1) Concern regarding the central finding and its framing. I think that it is incorrect to describe the central finding as demonstrating Gettier intuitions. The reason is because knowledge attributions in Gettier cases should be interpreted with respect to relevant control conditions and that doing so reveals a very different picture that for all intents and purposes does replicate the original finding.

To review, the Knowledge condition is a positive control meant to describe paradigmatic knowledge in a standard JTB (justified true belief) case, or what philosophers, at least, consider a clear case of knowledge. The Ignorance condition is a negative control that represents clear ignorance in a FB (false belief) case. Historically, philosophers consider Gettier subjects ignorant. If we want to know whether ordinary intuitions about Gettier subjects resemble paradigmatic knowledge or ignorance, then we need to compare them to these controls. One important reason for this is that it accounts for baseline skepticism in the sample by comparing those responses to how likely participants ever are to attribute knowledge in JTB. Afterall, there could be many reasons why people are generally skeptical of knowledge in classic JTB cases that have nothing to do with Gettier.

Presently, the research team observed only 56.59% attribution in all Knowledge control conditions (as reported in Table 6, although 55.6% when I analyze the data posted, so please double check reporting). This suggests that participants in this study are indeed very skeptical of standard JTB cases for unknown reasons (and indeed much more generally skeptical than Turri et al 2015’s participants). Thus, assessing the presence of a “Gettier intuition” in the present study, and arguably, making more informative comparisons to the original finding requires us to evaluate relative patterns in knowledge attribution. This was not suggested originally because it was unanticipated that baseline skepticism would differ so drastically between samples.

|  |
| --- |
| **Response**: Thank you for bringing up this interesting point. The skepticism in the knowledge control condition was unexpected and we agree that examining and addressing it would strengthen our contribution (see more on this below).  We have checked the reporting of the proportions in Table 6 (now Table 7). Note that the estimates reported are derived from the model rather than the raw numbers, which are slightly different. |

In order to accomplish this, I recommend a very simple procedure established by prior researchers (Starmans & Freidman 2020) be included in the manuscript. These researchers created “a Gettier score, generated by dividing the percentage of participants attributing knowledge in the Gettier condition by the percentage attributing knowledge in the JTB condition; and a False Belief Score, generated by dividing the percentage of participants attributing knowledge in the False Belief condition by the percentage attributing knowledge in the JTB condition” (p. 12). This tells us how people are evaluating knowledge in Gettier cases with respect to their inclination to ever attribute knowledge.

Doing this with data supplied in Table 6, the Gettier score across all vignettes is 75.47 and the false belief score is 31.70. This tells us that participants attributed knowledge across Gettier conditions on average about 75% as often as they do paradigmatic JTB knowledge. (By comparison, the Gettier score from Turri et al. Experiment 1 was 82.72; in Starmans & Friedman 2020 it ranged from 82 to 93). Just looking at Darrel vignette data posted by the authors, which was the only stimuli the authors used that were actually also used by Turri et al. (see below), 61.5% attribute knowledge in the Gettier condition as opposed to 69.6% in the Knowledge condition, resulting in a Gettier score of 88.36. This is arguably even stronger evidence against the Gettier intuition than was found by the original study.

|  |
| --- |
| **Response**: Thank you for this suggestion. We have added an examination of Gettier scores as described in Starmans & Friedman (2020) to our analyses and report the findings as exploratory, as we did not pre-register this approach. See page 66 for description and results and page 70 for their discussion.  Note that we calculated the “direct replication” Gettier score using the data from participants who completed the Darrel vignette first and found a value of 80.98 (72.40% knowledge condition; 58.63% Gettier condition). |

Summing up, I infer from this that the authors may have failed to replicate a null result, in that they were able to find a difference that Turri et al. did not detect between Knowledge and Gettier conditions in a much larger sample. As the authors note, the difference here is very likely due to power. But more to the heart of the matter, the data do not support the claim that Gettier intuitions are common. According to philosophers, the “Gettier intuition” is supposed to be that protagonists in Gettier cases are ignorant. But the authors found, consistent with Turri et al. and others, that participants attribute knowledge to a Gettier case subject up to 88% of the time as they would JTB. The fact that similar or perhaps even stronger results for Gettier knowledge are observed despite incredibly large differences in the stimuli, probes, methods, survey flow, samples, education, languages, etc., in studies is a testament to the degree to which people all over the world do this and is a valuable contribution to our understanding of knowledge in these cases.

|  |
| --- |
| **Response**: Thank you for the thoughtful feedback on our results and their interpretation.  While we see merit in your argument, prior research (including the study we replicated) have routinely tested Gettier intuitions based on condition differences or comparing rates of knowledge attribution to chance. We observed a difference between the Gettier and knowledge control conditions and found that participants attributed knowledge to Gettier protagonists less than they would by chance.  Since we pre-registered what counts as evidence in support of Gettier intuitions, changing how we interpret these results now would run counter to our predetermined criteria. So, while your argument is sound, we will continue to interpret these results using the criteria we defined during Stage 1 of the study - such that statistically significant differences in rates of knowledge attributions between Gettier and knowledge conditions constitutes evidence in favor of Gettier intuitions.  However, to address your important point that Gettier intuitions were not common, we have altered the claim in the abstract (see page 6) and reframed some of the discussion to better reflect the nuanced nature of the results (see pages 68-70). |

(2) Questions about stimuli materials and their relation to the original study. As I mentioned in my evaluation of the initial submission before data were collected, I had concerns about vignette adaptations used in the study, such as the Emma vignette. As it turns out, Emma might be an outlier from the other two vignettes. Only 34% attribute knowledge in the positive knowledge control representing paradigmatic knowledge in JTB, which naturally leads one to doubt whether it is a case of JTB. Also, looking at the Gettier condition more closely (pp. 104-5), the Emma vignette differs in several ways from original materials that deemphasize important epistemic features of Gettier cases. For example, adaptations deemphasize or remove the justification that the protagonist has (e.g. the store and its employees are made out to be completely unreliable; it is directly stated that “Emma could not tell” there is a diamond). The belief component is also deemphasized by using language that encourages the perception that the protagonist is guessing (e.g. “the one she chose happened to be real”). Without belief or justification, the case does not qualify as a Gettier case. Additionally, this possibility, if true, would nicely explain the authors’ findings for the effect of cover story and the interesting additional follow-up result that knowledge attribution increases with vignette protagonist’s expertise. It would explain this assuming that expertise confers the epistemic components obscured by the Emma case that the authors constructed.

|  |
| --- |
| **Response**: Thank you for this important observation about the Emma vignette. We were also intrigued by the differences in responding to the Emma vignette and had briefly explored this in the discussion. However, your additional points are helpful in highlighting how the Emma vignette was meaningfully different, so we have added further coverage of this issue in the Discussion section (see pages 75-77). |

I invite the authors to reflect on these points and whether they impact any of their stated conclusions. At the very least, I think this warrants clarification that collapsing across vignettes and presenting knowledge attribution percentages as the authors do in Table 6 does not qualify as a replication attempt of Turri et al., since Turri et al. did not use 2 of the 3 vignettes presented. Better would be to present descriptive statistics for each vignette and condition.

|  |
| --- |
| **Response**: Given the unique findings that we saw with the Emma vignette, we also considered that the Gettier intuition effect size estimated by collapsing across vignettes was not directly comparable to the results observed in the original study. We had this issue in mind when we decided to report an exploratory direct replication analysis of the Darrel vignette (from Turri et al., 2015) that included only those participants who responded to that vignette first. We believe that this exploratory analysis constitutes the closest we can get to directly replicating the original researchers’ analysis. We believe that the inclusion and discussion of this exploratory analysis, in addition to the analysis and discussion of vignette differences, sufficiently examines and describes how including the two additional vignettes may have affected our results.  As requested, we have added descriptive statistics for each vignette and condition to what is now Table 7 (see page 47). |

(3) Data exclusion and resulting analyses. The high rates of data exclusion of nearly half of the total participants are very concerning. Could the authors verify how exclusions impacted the central analyses? Related to this point, were any adjustments made to chance rates of knowledge attribution considering the exclusion criteria used? For example, participants had to answer at least three comprehension test questions correctly before they had the opportunity to answer knowledge questions, which objectively decreases the chance rate of observing a knowledge attribution to 12.5% chance at most (p. 31). If no corrections were made to chance rates given the data exclusions, then this emphasizes the need to look at relative knowledge attribution to control conditions, in which the same exclusions were also made.

|  |
| --- |
| **Response**: Participants who answered any one (out of three) comprehension question incorrectly were excluded from the full analysis. While this resulted in very high exclusion rates, we did preregister this exclusion rule beforehand, and we believe that this strict criteria helped prevent the inclusion of participants who did not understand or pay attention to the materials.  We carried out a specification analysis to further investigate the impact that our strict exclusion criteria may have had on our findings shown at <https://osf.io/nvfbm>. If vignettes are only excluded when their corresponding attention question is missed, we generally find the same results as reported in the manuscript. See response to reviewer 4 (comment 11) for exception.  No adjustments were made to the knowledge attribution variable (or any other variables) to account for the possibility of chance, inattentive, or errant responding. |

(4) Understanding high rates of skepticism. It is surprising how generally skeptical participants were of JTB in these studies, and it suggests further analyses to figure out why. My first thought was that it made me curious to learn more about the field sites and the differences in knowledge judgments between sites. Could the authors elaborate on rates of knowledge attribution across sites? My second thought was whether order effects could be contributing to skepticism. The basic idea here is that sequences of cases (and perhaps questions) might make people more skeptical, which would then drive down attribution rates in cases they would have otherwise attributed knowledge. For this reason, it would be very informative to see separate analyses of results where each case was only first presented.

|  |
| --- |
| **Response**: Knowledge attribution did not vary according to site. All of our models included random intercepts of site; this variable accounts for the between site variability in outcomes. Notably, the random intercept of site did not contribute meaningfully to any of our models (variance < .0001). We added an interpretation of this overall finding on page 42. As shown in <https://osf.io/nvfbm> (see also response to reviewer 2), the condition differences in knowledge attribution also appear largely consistent by country (except for differences in the confidence interval of the effect size due to country sample size differences).  We had also considered the possibility of order effects, but we found no meaningful order effects in our analysis as shown in (<https://osf.io/uz8te>; Exploratory Analyses > Order Effects). See also our response to Reviewer 4 (comment 12).  As suggested, we additionally examined knowledge attributions in response to the Gerald and Emma vignettes when they were the first responded to by participants. See page 57 and our response to Reviewer 4 (comment 12). |

(5) A rhetorical and conceptual point. This point regards the use of the phrases “Gettier case”, “Gettier intuition” etc., which are both common in psychology and are repeated often by the authors. The fact is that while “Gettier-style cases” is a convenient label used to study knowledge attribution in the face of luck (or often involve what is sometimes called a “double-luck” structure), the cases originally used by Edmund Gettier in the 1960s have almost no other similarity to the cases presently used by the authors. One of Turri et al. 2015’s original points (see also Blouw et al. 2018) was to argue that the label is used to describe a range of cases that likely involve distinct phenomenon. The authors may have discovered evidence of this as well, given the vignette effects they have observed. In the absence of work establishing validity or uniformity of this category, I suggest caution about the use of “Gettier case” or “the Gettier intuition” in describing responses to cases that have not been normed, or in drawing comparisons between studies using different cases. The risk is that lumping these things together might accidentally collapse distinct theoretical categories about how knowledge and luck are related to each other and introduce confusion to the research record with respect to their underlying psychological constructs.

|  |
| --- |
| **Response**: Thank you for raising this very important point. We addressed this issue briefly in the introduction; however, several paragraphs were unfortunately deleted by accident during the paper formatting process and the complete introduction was not included in our Stage 2 submission. We apologize for this error. This omitted material has been added back in as part of our edits (highlighted for your convenience on pages 8-20). We understand that what constitutes a Gettier case is not well agreed upon. We acknowledged this problem and were intentional in how we defined the term in the Stage 1 manuscript, but this section was unfortunately part of the writing that was deleted in error. Attempting to redefine or rename our primary construct at this point would radically change the framing of the research approved in our Stage 1 manuscript and go against the spirit of a Registered Report. We hope that, despite perhaps not agreeing with how we defined or operationalized Gettier cases, you find our argumentation clearer now. We have also added some clarifying language to the discussion section to avoid the problem of unintentionally collapsing across distinct theoretical categories (see pages 68-77).  Thank you again for your helpful comments and feedback. |

References

Blouw, Peter, Wesley Buckwalter, and John Turri. (2018). Gettier cases: A taxonomy. In Borges, R., C. de Almeida, and P. Klein (Eds.), Explaining knowledge: New essays on the gettier problem (242–252). Oxford: Oxford University Press. <https://academic.oup.com/book/25890/chapter-abstract/193598859>

Starmans, Christina, and Ori Friedman. (2020). Expert or esoteric? Philosophers attribute knowledge differently than all other academics. 44(7), e12850. doi:<https://doi.org/10.1111/cogs.12850>

# Reviewer: 2

Comments to the Author

I have two main suggestions for improving the final version of this article:

1. The discussion of contextualism on p. 48 of current ms in the general discussion could be deleted. This is a very intricate matter and I really doubt it can be usefully engaged with in such a quick fashion.

to put it simply, not evidence of variation is evidence for contextualism. Many factors would be expected to influence knowledge ascription, whether or not contextualism is true. The Emma vignette differs from the two other vignettes by commenting on the lack of discriminatory skills of the character, which is a relevant feature.

|  |
| --- |
| **Response**: Thank you for this suggestion. We have removed the relevant text from the discussion as recommended. |

2. Since the study is cross-cultural, it would be great to have a graph representing the results site by site or at least region by region, perhaps in the supplementary materials.

|  |
| --- |
| **Response**: Thank you for this suggestion. Please see graphs for the knowledge outcome measure by region in the supplemental materials at <https://osf.io/nvfbm> (Forest Plot of Effects) and below.      Thank you again for your helpful comments and feedback. |

# Reviewer: 3

Comments to the Author

Overall, it was a pleasure to see the results of the Stage 1 Registered Report. As you will see from my comments, I found the whole endeavor to provide interesting information and have a notably thoughtful discussion. I am also impressed that the team was able to do all this, given that we had a global pandemic in between Stages 1 and 2.

I will note that I reviewed and considered each change to the preregistered protocol. Therefore, if I do not mention a change, it can be assumed that I feel the authors justified the change satisfactorily. I do have one recommendation that pertains to the writing of the introduction (Comment #1), although I recognize that this is a registered report (i.e., where I typically avoid commenting on what was reviewed in Stage 1). So, please take Comment #1 with a grain of salt. My other comments surround changes to the analytic plan from the preregistration or are general comments.

1. This is a stylistic comment, mostly, but I found one paragraph in the introduction to be somewhat unclear. The paragraph that begins on manuscript page 3 and ends on manuscript page 4 mostly describes the empirical support for and implications of Guttier-type cases. However, in the middle of the paragraph is the sentence, “However, past results have been mixed (e.g., Powell et al., 2015).” After that, studies are described that further support Gettier cases, but none that show the opposite (i.e., what I would consider to be “mixed” results). The Turri et al. study obviously didn’t support the Gettier-type case, but that isn’t mentioned in that paragraph. It is mentioned in the following paragraph which is a different section. As a result, the sentence in the middle and the lack of any examples of studies other than Turri et al. (2015) that did not find support for Gettier-type cases in that paragraph made me, as a non-expert in this substantive research area, wonder whether the Turri et al. finding is the only published work that did not find support for Gettier-type cases or whether it is just a notable and recent one. My recommendation here would be to either add a description of and citation to one of the studies that you later review in the “Cultural Considerations” section, where there are some results that are consistent with Turri et al (e.g., Weinberg et al., 2001) to clarify that Turri et al. was not the only piece of evidence against the existing scientific understanding of knowledge attribution in Gettier-type cases.

|  |
| --- |
| **Response**: Thank you for this feedback. Unfortunately, sections of the introduction were accidentally deleted during the formatting of the manuscript. Most relevantly, the paragraph following the one you highlight was deleted, which went on to describe the counterevidence (see what is now page 9). We have carefully reviewed this writing and believe it resolves the issues that you have identified. |

2. I am glad that the authors switched to a VAS in order to determine whether the binary or continuous nature of the measurement could explain the different result by Turri et al. However, of course, the decision was later made to recode the values into binary variables. I was not convinced that this was the best choice. So, my first recommendation is to provide more justification for why this choice was made, perhaps relative to other choices that could be made. However, I will admit that I would have thought it was fine to consider the DV as a binary variable if there was a comparison between the binary and continuous quantification. When I look at the distributions provided in Figure 2, I do not conclude that the data are “dramatically bimodal” (p. 13). I certainly see the ceiling and floor effects but I also see a range of responses along the continuous axis in between the two poles (0 and 100). Especially given the personal communication with Turri that is highlighted, I think it would be nice to do some type of quantitative comparison to explore the question of whether a binary or continuous measure affects the effect size. There are a number of alternative analytic approaches that could be used (e.g., see Buntin & Zaslavsky, 2004, <https://pubmed.ncbi.nlm.nih.gov/15120469/>, for a review). Ultimately, I do not think it is absolutely necessary for the authors to reanalyze the data, but I am interested in the variation that exists in between the two poles. It would be sufficient for the authors to better justify this choice. To me, it was not self-evident that these responses were best considered as a binomial random variable.

|  |
| --- |
| **Response**:  We have expanded our discussion of the assumption violations to include more clear justification of our decision to dichotomize the outcome measures than we previously provided (see page 40).  To see if our decision impacted the research findings, we repeated our analyses using the continuous variables in their original form as suggested. As we explain on page XX and in <https://osf.io/nvfbm>, these analyses revealed the same general pattern of results. Note that the effect sizes from the linear and logistic models are not directly comparable (i.e., they are different metrics), but they generally appear similar in magnitude.  Further, on page 64-65, we detail an exploratory analysis that examined the interaction between original measurement type (binary vs. continuous) and condition in predicting outcomes. No interaction effects were found for knowledge or reasonableness, our primary dependent measures.  Please see our response to the editor regarding this change above for more details. |

3. This will only require a sentence to address, but it would be good to more directly state the criteria that would get a non-missing responses dropped during the transformation from continuous to binary (see bottom of manuscript p. 14). That is, I think it would be good for clarity’s sake to state that non-missing responses between 40 - 60 were dropped, even though that can be inferred from the description of the values that were coded as 0 and 1. Continuing from my previous comment, though, if those responses were dropped because they were midpoint values (i.e., 50), then that would be further reason to choose another analytic approach than the binary transformation. If the authors determine that it is best to stick with the binary transformation, then it would be good to additionally provide a strong justification for omitting these responses (e.g., instead of doing a midpoint split).

|  |
| --- |
| **Response**: Thank you for this suggestion. As suggested, we have added clarifying information about the criteria we used to split the data. We have also added justification for why this criteria was used (see pages 40-41).  As we state in the manuscript, a small percentage of values were deleted during the transformation procedure. For example, only 2.87% of responses that met all inclusion criteria for the knowledge attribution analyses were excluded due to dichotomization.  To further justify our choice, we examined a direct midpoint split in an exploratory analysis. It is difficult to argue for a midpoint split given that 49 and 50 are treated as separate categories when they are very close, which is why we choose points that were clearly delineated. However, as shown in <https://osf.io/nvfbm>, we do not find any differences in our results when you use this split rather than the 40/60 one in the manuscript. |

4. I appreciated the thoughtful and rich discussion. The authors went beyond simply evaluating the success of the replication. They considered how their work could be truly integrated with the findings of Turri et al. (2015) and used that to identify new questions. Altogether, the discussion was well written and interesting.

|  |
| --- |
| **Response**: Thank you for this comment. |

5. I really appreciated the transparency and comprehensiveness of the changes reported to the Stage 1 Registered Report/Preregistration. I felt that it was useful to me to consider all these challenges of such a large and multi-site collaboration.

|  |
| --- |
| **Response**: Thank you for this feedback. |

6. I also appreciated the “pedagogical goals” section, because it provides important context for the reader, given that these goals were central to the project. I also appreciate the training provided to burgeoning scientists that occurred through this work.

|  |
| --- |
| **Response**: We really appreciate that feedback! |

Thank you all for your great work!

Elizabeth Page-Gould

|  |
| --- |
| **Response**: We have really appreciated your feedback throughout the process! You have been a big help in making this paper more rigorous, thank you! |