Dear Kathleen, Priya, and Chris,

Thank you for the opportunity to review your manuscript. Research on how well researchers can predict research findings is always fascinating to me. Results often cast somewhat of a shadow on our self-image of supposed expertise, and I think this is an incredibly useful check and balance on our collective work.

This is an interesting, worthwhile, and thoughtfully planned study and I have no doubt that it should be conducted. I have several comments, but only one of these refers to something I feel is genuinely absent from the current manuscript (the first point below). I think all my points are easy or fairly easy to address, and none would stand in the way of progression towards study conduction and eventual publication.

**Note on priority labels**

I myself sometimes struggle to know how strongly a given reviewer feels about how authors should respond to their comments in order to satisfy them. I have therefore annotated each point based on how what sort of response I am hoping for. This ranges from a simple [invitation] to consider something (and possibly discard it as not an interesting or useful point in your opinion), to a [request] that something be revised (but perhaps there are good reasons not to, I would not die on any of these hills), to a [strong request] where I feel that something might be a barrier to publication if it was not revised, discussed, qualified or directly defended in some way (where authors may need to either dedicate the most effort to either making suitable changes or defending why I have misunderstood, or they disagree about the importance of the point if not the validity of it, etc.). I’m still experimenting with this way of writing reviews, please feel free to DM me on twitter with feedback about whether you think it’s useful or not.

**Comments**

(doesn’t refer to any single quote in the current manuscript)

* [strong request] I would like to see more reflexivity in the application of the concept of generalizability.

Your opening statements in your abstract etc state that you wish to make conclusions about all of psych science: From your abstract, “The proposed research will examine researcher predictions regarding the generalizability of psychological effects”. However, this will be done via four studies, the domain and nature of three of which is not currently known. The representativeness of these studies is therefore important to the generalizability of your findings. You also have acknowledged some constraints on the selection of these studies, however you don’t seem to have balanced this with a prior acknowledgement of how it may necessarily temper your conclusions. Perhaps this is in part because the limitations section, usually part of the discussion, is absent from the stage 1 RR process.

I would like to see deeper consideration of whether some domains of study are a priori more likely to vary between contexts, e.g., things like the social norms and the distribution of attachment styles on one extreme vs. low level effects like the stroop effect, principles of learning/conditioning effects, just noticeable differences effects etc. on the other extreme. I recognise that I don’t have data for any of these myself! But it is my prior belief here, and I think may readers may share it. From which end of the spectrum (if it exists) are your studies likely to come, and will they be randomly chosen from it or are there likely constraints to the domains etc?

More importantly, I would like to see some discussion of how the nature of the studies that are eventually chosen, via convenience sampling, may influence not only the conclusions that may eventually be drawn but even the questions that are asked. Do the authors conceptualise the studies as coming from the same population of “psychology studies”, and that the reader can and should generalize their results from this sample to the population, as you imply in your abstract? What factors influence which type of research questions and studies make it across the desk of the psych science accelerator? Do they tend to come from research domains that, at least in your estimation, are a priori more or less likely to involve heterogeneity between sites? Of course, it is an empirical question as to whether they DO show heterogeneity, but I am asking more about your conceptualisation of the population that they are drawn from than what heterogeneity exists. At a very basic level, my understanding is that these studies are more likely to be related to social and personality psych than, for example, clinical psych, comparative psychology, or perception, right?

On p.12 the manuscript states: “Research questions will be related to the funding agency’s strategic priorities, which include the dynamics of religious change, intellectual humility, religious cognition, the science of character virtue, and health, religion, and spirituality. Projects will be selected in accordance with PSA policies and procedures based on the feasibility, quality, and appropriateness for the call.” There are therefore multiple filters that constrain which studies are run vs not run, including both their domain/topic which bias towards religious belief (due to the funders’ interests) and other less well specified feasibility constraints. What bearing might these and other factors have on the generalizability of this research? Given the topic of your paper is generalizability, I think your paper would be best received if you are seen to be reflexive in the application of the concept to your own results where you can.

(p.5) “In multi-laboratory investigations of replicability, effects have either consistently generalized or failed to replicate across sites (Ebersole et al., 2016, 2020; Klein et al., 2014, 2018, 2022; Olsson-Collentine et al., 2020).”

* [invitation] Perhaps this point could be fleshed out further to mention the proportion of studies that do indeed replicate, and the implications of this for the study of any variables associated with it. I.e., studying such factors is difficult when there is little variation to work with. This was the bane of Many Labs 5 for example - as you are of course well aware of as co-authors of Ebersole et al. (2020). In the last year, I’ve heard talks from many groups interested in understanding heterogeneity in replications and this point keeps coming up: there is often insufficient variance to work with, possibly due to the over representation of what are likely to be true-null effects among these studies. However this point seems to be less prominent represented in published research. Perhaps the authors would like to speak to it in their manuscript?

(p.7) “Peters et al. (2022) argued that scientists demonstrate a generalization bias in which they generalize their results to broader populations than is warranted.”

* [invitation] I think this point has been made by people before Peters et al. (2022), including in a preprint I wrote (i.e., Hussey, 2020, General claims require generalized effects, <https://psyarxiv.com/83z2y/>). I hesitate to do that annoying thing where reviewers suggest their own papers, so please don’t feel that I am strongarming you to cite me, especially as it’s not a published piece of work, that’s not my goal here. Reviewers of my preprint said that this point had been made before, although I couldn’t seem to get them to suggest by whom. Peters et (2022) is a good citation and I wasn’t aware of it, but feedback that this point had been made before and lacked novelty was one reason why I abandoned trying to publish my preprint. Perhaps you could track down earlier work that makes this point, assuming my reviewers were right? The Peters et al citation is appropriate to this point though, and should stay of course.

(p.8) “Overall, the effects of expertise and experience on predicting research outcomes are unclear, and whether these previous findings extend to generalizability prediction is unknown.”

* [invitation] Perhaps it would be useful to be reflexive/reflective here in the manuscript’s consideration of the replicability of research findings. It may be the case that the effects in the literature on the prediction of research outcomes is not merely mixed/unclear but that some or all of these effects are themselves not replicable. Injecting some consideration of replicability into the discussion of the meta-scientific research in addition to the scientific research might could be useful. “Who will watch the watchmen” is a topic I see more and more when it comes to the replicability of meta-science findings. Just a thought; not looking to derail your existing narrative.

(p.8-9) “However, prior research has found some evidence for individual differences predicting performance in other forecasting or prediction contexts. For instance, Haran et al. (2013) examined how predictions under uncertainty were related to individual differences, including actively open-minded thinking (AOT; Baron, 1993; Stanovich & West, 2007) and need for cognition (Cacioppo et al., 1984). Of the examined variables, only AOT was associated with accuracy, though this positive relationship depended on the usefulness of the available information. Researchers investigating so-called “superforecasters” predicting future geopolitical events found that superforecasters score higher than other forecasters on AOT and other individual differences in cognitive style and ability (Mandel & Barnes, 2014). These same variables positively correlated with prediction accuracy among forecasters more generally (see also Mellers et al., 2015).

Intellectual humility, or the willingness to recognize the limitations of personal knowledge, has several potential social and personal benefits (for a review, see Porter et al., 2022) that may also be relevant to prediction. Researchers have found relationships between intellectual humility and AOT (Beebe & Matheson, 2022; Krumrei-Mancuso et al., 2020; Krumrei-Mancuso & Rouse, 2016), curiosity (Krumrei-Mancuso et al., 2020; Leary et al., 2017), cognitive flexibility (Zmigrod et al., 2019), and the ability to identify argument strength (Leary et al., 2017). Intellectual humility among scientists may even improve research quality and credibility (Hoekstra & Vazire, 2021; Nosek et al., 2019) and drive scientific progress (Porter et al., 2022). For instance, intellectual humility predicts how much psychology researchers update their beliefs about effects in response to new evidence (McDiarmid et al., 2021). Thus, intellectual humility may be another feature of researchers that relates to their ability to predict research generalizability.”

* [request] Prior to this point, the manuscript often mentions previous works’ effect sizes, but not here. Could the authors flesh out this point with some mention of the strengths of association/prediction accuracy etc?

“All sample size determinations, data exclusions, manipulations, and measures will be reported (Simmons et al., 2012).”

* [invitation] Perhaps change this statement to the 21 words used by Simmons et al? They state that one of the benefits of the 21 word solution is that it is standardized, both in its meaning and its reporting format (standard wording makes it more machine-readable). Deviations from it may be merely stylistic or may represent important qualifications that may also be subtle. If you mean to invoke the 21 word solution wholesale, perhaps its best to use it verbatim?

(p.13) “A single focal effect will be chosen from each study based on input from the proposing authors. The effect will be the result of a single inferential statistical test that answers a central research question from the project. Priority will be given to effects that are grounded in theory and supported by previous research. For example, the focal effect for Moral Experiences will be that experiences of moral events will produce higher momentary happiness than experiences of immoral events.”

* [request] Chosen by whom, the authors of the current manuscript? Will the authors of the component study have any input into the selection of this? The answer could well be no, I’m just looking for a few more details here on whether anyone will be able to post hoc say they didn’t think that question was central etc. E.g., will it be taken from the study’s prereg, or will you seek assent (not necessarily approval) of the first author of the component study, etc? The answer could be no, again, but more details on the selection process would be useful.

(p.14) “Data from each study will be analyzed to produce a binary focal effect outcome (significance at p < .05)”

* [request] The PSA call for studies mentions that ethical peaking strategies that don’t inflate the false positive rate may be used by these studies. However, this manuscript states that all results will be computed analysed for the whole study with alpha = .05. It is therefore possible that this manuscript’s analyses may come to different conclusions to write ups of those component studies due to potential differences in alpha control. Is this design choice intentional? I get that it simplifies things, but also how it complicates in other ways. If there is reasonable potential for deviation here, perhaps this could be considered.

(p.14). “All effect sizes will be transformed to a common metric of Cohen's *d* before analyses. We chose this metric because it is unbounded and easily interpretable.”

* [strong request] Using what equations and/or R packages? Preregistering this would be desirable. Meta-analyses often suffer from non-reproducibility due to experimenter degrees of freedom in effect size conversions. Even preregistering the most common effect size types would cover a lot of bases. Perhaps even a reference to a specific article that contains equations for many of these would be useful (although I don’t know which R packages have good correspondence to single papers).

(p.14). “Five to ten potential moderators will be chosen for each study with input from the proposing authors.”

* [request] This is quite vague. The supplementary materials list examples of moderators in more detail than the current manuscript’s text. Could the manuscript be updated to represent this? (p.S4):

*● Age*

*● Gender*

*●  Education*

*●  Religiosity*

*●  Socioeconomic Status*

*●  Political Orientation*

*●  Individualism-Collectivism*

*●  Relational Mobility*

*●  Cultural Tightness-Looseness*

(p.14) “At the individual level, moderators will be tested as appropriately specified additions to the overall focal effect analyses.”

&

(p.14) “For the sample level tests, moderators will be tested as a predictor/moderator in random effects meta-regression models. The binary outcomes of these moderation analyses (i.e., their significance at p < .05) will serve as dependent measures in secondary analyses.”

* [request] Could you provide more information on which variables will be used as individual level vs sample level moderators? Some variables could refer to individual level if used natively or group level differences if aggregated. E.g., individual level cultural tightness-looseness could be an individual differences variable, but equally the estimation of the \*culture\* might be better determined by the mean cultural tightness-looseness in each sample, used as a sample level moderator. Perhaps I have misunderstood here, but if there is reasonable potential for lack of clarity about how each variable will be used (individual vs sample level) – and indeed I think there is likely to be given that these studies and variables have not yet been designed – I think it would be useful to put some guardrails / constraints on the EDOF here. Whatever logic you’re likely to follow in making these decisions could at least be sketched out here, IMHO.

(p.20) “Missing data and exclusions … No participants will be excluded from the analytic dataset. All participants with available data on the relevant variables will be included in a given analysis. Given our population of interest, PSA member researchers, we anticipate high participant engagement that produces good data quality.”

* [request] these exclusions seem to refer to the predictions data gathered from researchers, but not the focal effects gathered in the component studies. I assume that you’ll take the final analytic datasets from the component datasets with whatever exclusions they applied in their analyses, but if so perhaps better to state this here? And, assuming that those studies’ exclusion strategies will be preregistered, maybe useful to mention that too so that readers can gauge the researcher DOF in these datasets without having to fully read their core manuscripts, match up results, etc.

(p.20) “Relationships between Predicted and Actual Results … we will examine how the mean probability estimates of finding an effect in the subsamples relates to our binary outcome variable. We will also examine the relationship between the means of the predicted effect sizes for the subsamples and their observed effect sizes.”

* [invitation] Given the bounded nature of probabilities, means often don’t aggregate them well. E.g., Differences in confidence of finding an effect between .99 and .999 matter a lot more than the difference between .50 and .509; Bayes Factors are expressed as ratios rather than probabilities for this reason, etc. I understand the desire to simplify the question put to researchers so that it’s comprehendible for the participants, and the desire to simplify the aggregation metric and analysis so that it’s comprehendible for the reader, but is there a chance that this risks simplifying things to the point of inaccuracy? At the simplest level, is there any good reason to use means over say medians, given unknown skew in probabilities? At a more complex level, would it be more appropriate to either ask researchers about odds instead, or convert their probabilities to odds and calculate mean/median odds instead? Whatever metric of aggregated predictions is chosen, the same questions arise for the aggregated actual results (i.e., choice of metric of central tendency, choice to use probabilities vs odds).
* [request] Separately, I completely understand the need to summarize across sites using a comprehendible dichotomous metric, such as p < .05, that is summarized in some way (see above). You are in effect treating all subsamples are mutually substitutable, when they are very likely not to be so: they will have different sample sizes, and therefore different power to detect the underlying effect. Aggregating over the dichotomous significance decisions like this would be a big no-no in meta-analysis, as of course you well know. However, this is not a reason not to do it here – your strategy is sensible for your goals, and there is no clearly superior alternative that I know of. However, perhaps a brief discussion of these assumptions and the pragmaticism/necessity of this metric in the absence of viable alternatives would be useful?

(p.21) “Because the outcome measures will not vary according to researcher, including random intercepts of researcher in the models is not appropriate (i.e., would produce a singular model fit). Instead, we will calculate “intercepts” for each researcher to include in the models as fixed effects by running an individual model for each researcher with their predictions predicting outcomes and extracting the model intercept.”

* [request] this is a tricky step in the analysis for the reader to grok without reading your code, which not everyone will be able to do. Could you unpack this point, both in terms of the prior analyses being doing and the rationale for them?

(p.21) “Calculated individual researcher prediction slopes (i.e., their model coefficients) and random slopes of prediction for study, sample source, and sample region will be tested to see if they contribute to the model and retained when they improve model fit. … If we observe relationships between predicted results and actual results in aggregate (i.e., the correlational analyses) or at the level of prediction (i.e., the multilevel model analyses), we will conclude that researchers are at least somewhat accurate in their predictions of the generalizability of psychological effects across regional subsamples.“

&

(p.22) “If we find an effect of a tested researcher characteristic on accuracy scores, we will conclude that prediction accuracy relates to that characteristic.”

&

(p.23) section on “Moderation Predictions”

* [request] Could you explicate in the text, and ideally also in the code, what metric(s) of model fit you will use to compare models. E.g., AIC will be used to compare model fits and the model with the lower value will be selected. Similarly, could you clarify your decision method for determining whether relationship are observed. I presume p < .05, but better to explicate this. Please also specify the method and implementation for calculating any such p values, e.g., Wald vs. Kenward-Rogers, using lmerTest, etc.
* [invitation] As a general point, I think there are several places in the current manuscript where I think that researcher degrees of freedom could be constrained further though more explication of methods and implementation in code. Of course, there’s always more one could do and at a certain point it’s overkill, but I think quotes like the above one could have their EDOF constrained quite easily, usefully enhancing the evidentiary weight of your results.

(p.21) “Absolute differences and Brier scores will serve as dependent variables in multilevel linear models including random intercepts of researcher, study, sample source, and sample region.”

* [invitation] Maybe note somewhere that Brier scores are the mean squared error as applied to predicted probabilities, it may save some readers a google search (I’m among them).

(p.22 and elsewhere) use of “exploratory analyses” for preregistered analyses

* [invitation] I wrapped myself in knots in a previous RRR trying to later explain what I meant by a preregistered yet exploratory analysis. Perhaps it will save you headaches to think about this. You’ve already used the primary vs secondary analysis distinction, so perhaps it’s the best available option, but it caused needless confusion in my previous project and perhaps you could avoid the same fate.

(p.23) “We will also use one-sample t-tests to compare the effect size predictions to the observed effect sizes. … To examine whether researchers tended to over- or under-generalize on average, we will compare the mean of the aggregated subsample predicted probabilities to the proportion of observed subsample effects across the four studies using a one-sample t-test.”

* [request] I assume you plan to do it, but it’s not currently stated: I would be very interested to see the unstandardized effect size and its 95% CIs here too, as well as the t test results that you mention. Perhaps even a standardized Cohen’s d too, although I haven’t thought enough about that. Metrics of the degree of under/over generalization would be very useful beyond the significance test; perhaps these could be added to the prereg.

**Things I have not done in my review**

It can be useful to explicate what I have not done or thought about as well as what I have. While I have examined your analysis RMarkdown file and though through its logic (indeed, I had to use the code to understand the analyses at times, see above comments), I did not perform a full code review. I did not inspect the RMarkdown file for the power analyses in any depth; I follow the logic you describe in the manuscript but made no attempt to do a deep dive on your logic or implementation of the power analyses in particular.

Best wishes and good luck with the project,

Ian Hussey