Response to Recommender and Reviewers

Recommender Chris Chambers

I have now received four very constructive and helpful evaluations of your Stage 1 submission. As you will see, the reviewers concur this is a valuable RR proposal that already comes close to meeting the Stage 1 criteria at PCI RR. Nevertheless I think you will find the reviews helpful in strengthening the work even more prior to in-principle acceptance.

Without summarising all of the insights, across the set of reviews the following issues struck me as particular headlines: (1) including a deeper reflection on the concept of generalisability and, in particular, the generalisability of this study itself, (2) addressing concerns about measurement validity, (3) elaborating a variety of additional key methodological details (including analysis plans and potential exclusion criteria), (4) tightening up of potential sources of bias resulting from some remaining researcher dfs, (5) justification of specific design decisions (including the participant pool), and (6) considering structural edits to improve the clarity of presentation.

All of these issues are readily addressable as part of the regular Stage 1 review process, therefore I am happy to invite a comprehensive revision and response.

Thank you for the positive decision on our manuscript. We have carefully considered all of the comments provided in preparing our revision.

Please see below for our inline responses to the reviewers and descriptions of any corresponding edits.

We are grateful to you and the reviewers for your well-considered feedback which has greatly improved our manuscript.

Ian Hussey

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 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.

Thank you for your review and feedback. We appreciate the labeling of your comments by priority.

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.

Thank you for this important feedback and request. We agree that the generalizability of our research findings will have substantive constraints, and, while we do intend to discuss this limitation fully in our Stage 2 discussion section, we have carefully considered how best to acknowledge it in the Stage 1 manuscript.

In the abstract, we have adjusted our language to provide more specificity to our claims.

page 2 (additions in bold), “The proposed research will examine researcher predictions regarding the generalizability of **four** psychological effects. … Our investigation will reveal whether researchers can accurately predict the generalizability of **these** psychological effects across cultural contexts while offering insight into what features of the researchers are related to their prediction accuracy.”

> **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?**

While we share your intuition that certain domains of study may produce more cultural variability than others, our focus is not on whether the effects themselves generalize, but rather if researchers can predict if and when they generalize. Intuitions about the generalizability of effects is precisely what we aim to investigate.

Any four projects, even if chosen randomly from the full population of research studies in psychological science, would not adequately represent the field. However, we acknowledge that our approach to selecting projects will restrict the types of effects that our research will examine. The effects themselves are unlikely to be “low level” given these restrictions. Please see below for further responses to this and related points.

> **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?**

We did not intend to imply that our results would generalize to all research studies. As explained above, we have adjusted the wording in our abstract accordingly.

Several factors influence what research types of proposals are submitted to the PSA and selected for implementation beyond any requirements included in the call for studies (we discuss this aspect below). First, the PSA membership is not evenly distributed across subfields of psychology. Second, the criteria for project selection include the feasibility of implementation across many data collection sites around the globe. Past and current PSA studies have generally fallen into the domain of social or cognitive psychology defined very broadly. As described in the manuscript (though they are cited generally, not specifically as PSA studies, see page 5), two of these studies found mixed evidence regarding generalizability (Bago et al., 2022; Jones et al., 2021). Others have found fairly strong evidence for generalizability (Wang et al., 2021) or consistent null effects (Chen et al., 2023).

Given the lack of systematic research on the cross-cultural generalizability of psychological effects in general, let alone across multiple subfields, we hesitate to make any claims about the likely generalizability of effects from a given subfield. Further, we believe that the variation within those subfields is likely much larger than the variation between them. Features of both the effects and the samples will contribute to the observed variation. As we state on page 6, “Effect heterogeneity, and failures to generalize, should emerge in multisite research to the extent that samples vary on cultural factors that produce or relate to the psychological phenomena.”

Thus, we do not have an a priori prediction as to the likely heterogeneity of the selected effects in our research. Importantly, however, the heterogeneity of the effects themselves is not the focus of our research–we are examining whether researchers can *predict* how outcomes and effects vary across cultural contexts. Still, generalizability prediction accuracy is likely to vary somewhat based on the features of the research outcomes. The heterogeneity of the results may be one such feature. For instance, researchers may be better at predicting the generalizability of large and highly generalizable effects than of small effects that are only inconsistently identified across sites. While we will discuss how the specific selected effects and their heterogeneity may have influenced our results, we are not testing their impact directly, and uncovering what characteristics of focal effects moderate prediction accuracy is outside of the scope of our research.

New Reference (not in manuscript):
Chen, S., Buchanan, E. M., Kekecs, Z., Miller, J. K., PhD, Szabelska, A., Aczel, B., … Chartier, C. R. (2023, May 31). Investigating Object Orientation Effects Across 18 Languages. https://doi.org/10.31219/osf.io/2qf6w

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.

Using projects from these special calls for studies does limit the types of research and research questions our focal effects will represent.

The first chosen project is within the domain of moral psychology. We now have more information about the potential research domains of the three remaining studies, as the deadline for project proposals has now passed. Author-defined research areas include moral psychology, cognitive psychology, and social psychology.

The strategic priorities may have biased the representation of the proposal research areas as you suspect. The feasibility constraints of the selection process, as noted above, also limits the types of research the PSA can implement–biopsychological or neuroscientific research that requires specialized equipment would not have been considered feasible.

We have clarified in the text that the projects need only be related to one of the priorities.

page 13 (addition in bold), "Research questions will be related to **one or more of** the funding agency's strategic priorities, which include..."

The call for studies also highlighted testing global generalizability of the phenomena of interest, and proposing authors have likely focused on phenomena for which cross-cultural generalizability is a particularly interesting and open question. Some authors may have been motivated to submit research proposals because they expect global variation whereas others may be confident in the generalizability of their proposed research across cultural contexts.

Notably, however, the effect chosen for each project may also not represent the average effect size within that project let alone that of the topic of research or domain of psychology.

As we noted in our prior response, researchers may be more or less accurate in their generalizability predictions due to specific features of the selected effects. We have no reason to expect that the filtering process will systematically influence accuracy, however.

We intend to discuss these and other constraints on generality extensively in the discussion section of the Stage 2 manuscript. However, we have added a brief preview of these concerns to the introduction.

page 12, “Given the focus of the proposed research on generalizability prediction, limitations on the generalizability of our results should be acknowledged. For instance, methodological features of the research, such as our chosen sample and how we will select the studies and their focal effects, will likely produce results that do not generalize to all researchers or all effects. We will discuss our findings with these constraints in mind.”

(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?**

Thank you for this suggestion. Replicability certainly is a prerequisite for generalizability across cultural contexts.

We have added a sentence describing the high proportion of failed replications in some of these investigations and the necessarily low heterogeneity of null results.

page 5, “However, the high proportion of failed replications in some of these investigations (e.g., 80% in Ebersole et al., 2020), likely contributed to low heterogeneity across sites because true null results have limited effect size variability.”

(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.

Thank you for this suggestion. We did some additional literature searching and could not find any references for this specific point (i.e., that overgeneralization is a cognitive bias); however, several of our pre-existing references and some others we found do make the point that unwarranted generalization is common and/or problematic. As we are focusing on generalization across cultural contexts specifically in this paragraph, we did not want to expand our discussion beyond that aspect of generalizability. However, we do think integrating other relevant references here is a worthwhile addition.

page 7, "For instance, Rad et al. (2018) found that most of the papers published in *Psychological Science* in 2014 relied on WEIRD samples and nevertheless made general claims, and DeJesus et al (2019) found that the majority of articles published in 11 psychology journals in 2015 and 2016 used unwarranted generic language to describe results."

(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.**

Thanks for this suggestion. We agree that the replicability of meta-scientific findings, including our own, is an important point to consider. We intend to discuss this in regards to our own research in our limitations section, and we have adjusted the text here slightly to indicate that the replicability of the findings mentioned is also unknown.

page 9 (addition in bold), “Overall, the effects of expertise and experience on predicting research outcomes are unclear, and whether these previous findings **are replicable or if they** extend to generalizability prediction is unknown.”

(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, effect size = XX). 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?**

We’ve added effect sizes or standardized coefficients for the results mentioned in these paragraphs when available. Unfortunately, such statistics were not available in some cases. Note that we also made a few other minor edits in this section to fix errors or clarify the writing.

page 9-10, “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 (NFC; Cacioppo et al., 1984). Of the examined variables, only AOT was associated with accuracy (β = .209), though this positive relationship depended on the usefulness of the available information. Researchers investigating so-called "superforecasters" predicting future geopolitical events found that superforecasters scored higher than other forecasters on AOT and other positive individual differences in cognitive style and ability (e.g., NFC; Mellers, Stone, Murray et al., 2015). These same variables positively correlated with prediction accuracy among forecasters more generally (e.g., AOT, *r* = -.12; see also Mellers, Stone, Atanasov 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 (*r* = .32, Beebe & Matheson, 2022; β = .56, Krumrei-Mancuso et al., 2020; *r* = .56, Krumrei-Mancuso & Rouse, 2016), curiosity (β = .22, Krumrei-Mancuso et al., 2020; *r* = .27, Leary et al., 2017), and cognitive flexibility (*r* = .35, Zmigrod et al., 2019). 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 (β = .086, McDiarmid et al., 2021). Thus, intellectual humility may be another feature of researchers that relates to their ability to predict research generalizability."

We've also added such statistics to other areas of the introduction when possible (see pages 8-9)

"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?**

Thank you for this suggestion. The deviation was merely stylistic, and we have updated the phrasing to match the suggested wording:

page 12, "We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study (Simmons et al., 2012)."

(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.**

Thank you for this suggestion. We have clarified the details of the effect selection process as requested.

page 14, "A single focal effect will be selected from each project based on input from the proposing authors. We will ask the proposing authors to identify effects from their project that meet the following criteria: 1) answers a central research question, 2) results from an inferential statistical test, and 3) is grounded in theory and supported by previous research. They will be told to prioritize simple and easily described effects tested at α = .05 if multiple effects meet this criteria. If the proposing authors suggest more than one focal effect, we will choose from among these randomly. We selected the following focal effect for the Moral Experiences project based on this procedure: Experiences of moral events will be associated with higher momentary happiness than experiences of immoral events."

(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 peeking 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 and 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.**

Thank you for this feedback. Our overall focal effect analyses will be based on the planned analysis described in each project's preregistered research protocol. Based on your feedback, we will now ask the proposing authors to consider the planned alpha in suggesting focal effects. The focal effect for the *Moral Experiences* project will already be tested at α = .05.

page 14, “They will be told to prioritize simple and easily described effects tested at α = .05 if multiple effects meet this criteria.”

However, the overall predictions are not the focus of our primary analyses. The subsample effects are unlikely to be examined by the authors of the individual studies. Further, the meta-analyses of the focal effects described in our analysis plan are also unlikely to be included in their reports.

Thus, we will disclose the differences between our analytic strategy and that reported by the authors and acknowledge any discrepancies in our conclusions.

We have added this detail as a footnote in the manuscript accordingly.

page 24, “10. While we will model our focal effect analysis after those planned for each project, discrepancies between our results and those reported by the researchers may occur. For instance, the researchers may not report a meta-analysis of the focal effect, or they may report a meta-analysis with different specifications. We will disclose and explain any such discrepancies in the analytic strategy or statistical conclusions for each effect.

(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).**

Thank you for this suggestion. Effect sizes will be converted using the *effectsize* package (Ben-Shachar et al., 2020).

We have added this detail to this manuscript.

page 16, “All effect sizes will be transformed to a common metric of Cohen's *d* before analyses using the *effectsize* package in R (Ben-Shachar et al., 2020).[5]”

As you suspected, the sources of the specific equations seem to vary based on the specific conversion function. From what we can tell from the package documentation, most seem to be derived from Borenstein et al. (2009). We have added a footnote indicating this source.

page 16, “5. The formulas for the effect size conversion functions in this package were primarily derived from Borenstein et al. (2009).”

We have also added code to the planned analyses script that will be used to perform the most likely effect size conversions.

New References (added to manuscript):
Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. John Wiley & Sons. https://doi.org/10.1002/9780470743386

(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**

Thank you for this feedback. The examples we provided in the supplement are common demographic or cultural difference measures that we anticipate may be included in the projects. We will be asking the proposing authors to suggest 5 to 10 potential moderators from among variables they intend to measure in their research.

We’ve added some details about this request along with examples to the manuscript (see page 16 or the quoted text in our response below).

We have also updated the supplement to indicate what moderators have been chosen for the Moral Experiences project (see page S5 of the supplement).

(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.**

Thank you for this suggestion. The same scales or items will be used for both levels of analysis. Each participant will predict the moderation analysis result for the same five variables at both levels of analysis in a total of 10 predictions. The moderator variable will be aggregated for the subsample level analyses. Continuous measures will be averaged (e.g., the mean age for each subsample), and categorical variables will be examined as proportions (e.g., the proportion of male participants in the subsample).

The details of the moderation analyses will be explained to the participants as part of the survey.

We have revised our description to clarify these aspects of the research.

page 16, “Five to ten potential moderators will be chosen for each study with input from the proposing authors. They will be asked to suggest demographic and individual difference measures included in their project that they believe may moderate the focal effect at the participant and/or subsample level. All moderators will be tested at both levels of analysis. At the individual level, moderators will be tested as appropriately specified additions to the overall focal effect analyses. For the subsample level tests, moderators will be aggregated (i.e., continuous measures will be averaged, while proportions will be calculated for categorical variables) then 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. For the *Moral Experiences* project, we will examine the following moderators: gender, religiosity, religious affiliation, belief in a personal afterlife, relational mobility, thriving, and moral identity internalization. For the individual level analyses, each variable and its interaction with experience valence (moral vs. immoral) will be added to the focal effect model. The interaction effect will be interpreted as evidence of moderation.”

(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.**

Thank you for this suggestion. We do intend to implement any preregistered exclusion criteria before calculating the actual effects and outcomes.

We have added this detail in our Actual Research Results subsection of the Methods.

page 15, “For each project, we will implement any preregistered data cleaning procedures and exclusion criteria prior to our analyses.”

(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 comprehensible for the participants, and the desire to simplify the aggregation metric and analysis so that it’s comprehensible 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).

While using means to aggregate predictions can reduce the extremity of the predictions, which may decrease their accuracy (Baron et al., 2014), prior research on predictions of research outcomes has most typically used means for aggregation. This method actually appears to produce less error than medians or other alternatives (Gordon et al., 2021).

We have added our justification of using means in our Measures section:

page 17-18, “Though aggregation of predicted probabilities using means can reduce the extremity of predictions, which may decrease their accuracy (Baron et al., 2014), prior research on predictions of research outcomes has most typically used means for aggregation (e.g., Benjamin et al., 2017; Camerer et al., 2016, 2018; Delios et al., 2022; DellaVigna & Pope, 2018; Dreber et al., 2015; Forsell et al., 2019; Hoogeveen et al., 2020; Viganola et al., 2021). This method may produce less error in aggregated replication outcome predictions than medians or other alternatives (Gordon et al., 2021).”

We likewise chose to ask about probabilities because of the use of probability estimates in prior research. Though we do not have evidence to support our intuition, we also believe that reasoning about odds would be difficult and confusing for participants.

We have added our justification of this choice in our Measures section:

page 17, “We chose to ask participants to estimate probabilities rather than alternatives (e.g., odds) in the interest of task ease and to allow for easier comparison with prior research on predictions of research outcomes (e.g., Benjamin et al., 2017; Camerer et al., 2016, 2018; Delios et al., 2022; Dreber et al., 2015; Forsell et al., 2019; Viganola et al., 2021).”

The actual results will not be aggregated. The binary outcome and effect size for each subsample will be determined based on the analysis of the subsample data; these values will be the same for the prediction level and aggregate analyses.

New Reference (added to the manuscript):
Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two Reasons to Make Aggregated Probability Forecasts More Extreme. Decision Analysis, 11(2), 133–145. https://doi.org/10.1287/deca.2014.0293

- **[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?

Thank you for this suggestion. We agree that the dichotomous measure has some important limitations, namely that variance in the subsample sizes will affect this outcome. However, it is the most common means of classifying research outcomes in prediction studies. We have also included effect size predictions in the research so we can examine predictions regarding both the presence and magnitude of the focal effect.

We have edited our Measures section to include a discussion of this choice. Further, we now acknowledge how variability in subsample sizes will impact power.

page 15, "Then, data from each project will be analyzed to produce a binary focal effect outcome both overall and within each subsample–a combination of country and sample source (i.e., university vs. community). The binary outcome measures will be the significance of the focal effect analysis at p < .05 in line with prior empirical outcome prediction research (e.g., Benjamin et al., 2017; Camerer et al., 2016, 2018; Delios et al., 2022; Dreber et al., 2015; Hoogeveen et al., 2020). We will also calculate an effect size for the focal effect both overall and within each subsample. While our dichotomous measure has limitations due to the impact of varying subsample sizes on the outcome, including both types of measures will allow us to examine predictions regarding both the presence and magnitude of the focal effects. Only subsamples (e.g., university students in Colombia) with 100 or more valid participants will be used in analyses. This number was chosen to ensure power to detect at least a medium sized effect in each subsample; however, the minimum detectable effect in each subsample will depend on the focal effect analysis as well as the subsample size."

Finally, we also intend to address this limitation in the discussion section of the Stage 2 manuscript.

(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?**

Thank you for this suggestion. We have edited and expanded on this text to further explain the rationale for and details of this analytic approach.

page 25, “Typically, random intercepts of participant would also be included in multilevel models such as these because of the repeated measures design. Random intercepts account for baseline differences in participant outcomes and are necessary when observations are not independent. However, as our outcome measures (i.e., the actual research results) will not vary according to researcher, including random intercepts in these models would produce singular model fits. Thus, we will instead calculate prediction “intercepts” for each researcher individually to include in our models as fixed effects. Specifically, we will run separate models for each researcher (i.e., 400 total per outcome) with their predictions predicting outcomes and extract the model intercepts. These values will then be included in the models to account for baseline differences in researcher predictions.”

(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.**

Thank you for this suggestion. As requested, we have added descriptions of the model comparisons to the text. We have also specified the order in which potential model additions will be tested.

page 25, “Calculated individual researcher prediction slopes (i.e., their model coefficients) and random slopes of prediction for study, sample country, and sample source will be tested to see if they contribute to the model. They will be tested one at a time in the order listed and retained when they improve model fit. For each addition, we will compare the new model’s Akaike information criterion (AIC) to the previous model’s AIC and select the model with the lower value.”

page 28, “Additions will be tested one at a time in the order listed; we will compare the new model’s AIC to the previous model’s AIC and select the model with the lower value.”

We have also added these details as annotations in the analysis code.

In the manuscript, we have added clarifying details regarding our alpha for all tests ($\alpha = .05$) and the methods for calculating *p*-values in our models.

page 23, “We will employ $\alpha = .05$ for all analyses. The logistic multilevel models will be fit using the *glmer* function in *lme4* (Bates et al., 2015) with *p*-values calculated using Wald tests. The linear multilevel models will be fit with the *lmer* function in *lmerTest* (Kuznetsova et al., 2017) with *p*-values calculated using Satterthwaite's degrees of freedom method.”

New Reference (added to manuscript):
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

- **[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.**

Thank you for this feedback. We have carefully reviewed the Methods and Analysis Plan sections, as well as the analysis code, in regard to potential researcher degrees of freedom. We believe the changes we have described in our responses to you and the other reviewers have sufficiently limited our methodological and analytic degrees of freedom.

Further, while the details of every planned analysis cannot not be described in the manuscript in the interest of length and practicality, the planned analysis code specifies their implementation to ensure transparency and reproducibility.

(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).**

We define Brier scores as operationalized in our Methods section when they are first introduced (see page 18): "Measures of accuracy will also be computed for every prediction. For binary outcomes, Brier scores (i.e., the squared prediction error; Brier, 1950) will be calculated for each predicted probability."

(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.**

We appreciate the potential confusion that including planned but exploratory analyses may create. However, we would like to note our intention to perform a series of analyses that are relevant to our research questions while acknowledging that our results will be limited because of the nature of the tests.

Our exploratory analyses currently comprise tests of several additional potential predictors of generalizability prediction accuracy and tests examining what researcher characteristics relate to moderation prediction accuracy. We are not preregistering any specific hypotheses for these analyses, and the analyses will be presented as exploratory in our results. Still, these analyses may yield findings that motivate future confirmatory research.

(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.**

Thank you for this suggestion. We have added in these details of how the tests will be reported.

page 27 (addition in bold), "We will also use one-sample *t*-tests to compare the effect size predictions to the observed effect sizes. **For these tests, both standardized effect sizes and unstandardized effect sizes with 95% confidence intervals will be reported.**"

page 27 (addition in bold), "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. **We will report both the standardized effect size and unstandardized effect size with 95% confidence intervals for this test**."

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

Thank you for the well-wishes and your helpful comments about our work.

Jim Grange

I think that the study addresses an interesting and timely topic, and an answer to the question posed by the research would definitely be of value. Although whether empirical effects generalise across different samples is of more importance than

whether researchers can predict such generalisation, the current proposed research certainly fills an interesting gap in the literature and I am looking forward to seeing the outcome of this study.

Thank you for your review and feedback.

I only have a few relatively minor points I hope the authors find of some use. I provide them in chronological order in which they appear in the manuscript.

Page 3 - "Psychology's WEIRDness problem": Perhaps spell out what this initialism stands for on first use for readers unfamiliar with the term.

Thank you for this suggestion. We have added a footnote fully explaining the term:

page 3, “1. WEIRD is an acronym for Western, Educated, Industrialized, Rich, and Democratic, all common characteristics of research participants in the psychological and behavioral sciences (Henrich et al., 2010).”

Page 4, second paragraph - When discussing generalisability of an effect across methods & measures, it might be of use to cite the following paper which proposes a method to deal with this issue: Baribault, B., et al. (2018). Metastudies for robust tests of theory. Proceedings of the National Academy of Sciences, 115(11), 2607–2612. https://doi.org/10.1073/pnas.1708285114

Thank you for this suggestion. We have integrated this reference.

page 4, “Researchers often make general claims based on specific operationalizations and fail to account for important features of the research like stimulus variation in their models. While some research methods, such as radical randomization (Baribault et al., 2018) or integrative experiment design (Almaatouq et al., 2022), may produce comparatively comprehensive results that consider variability across methods and measures, few researchers actively examine generalizability or address its limitations in their own work (e.g., Simons et al., 2017; Yarkoni, 2022).”

New Reference (added to the manuscript in addition to Baribault et al.):
Almaatouq, A., Griffiths, T. L., Suchow, J. W., Whiting, M. E., Evans, J., & Watts, D. J. (2022). Beyond playing 20 questions with nature: integrative experiment design in the social

and behavioral sciences. Behavioral and Brain Sciences. Advance online publication. https://doi.org/10.1017/S0140525X22002874

Page 13 - The supplementary material were very clear and provided an excellsent overview of what the participants would experience. Thank you for including this.

Thank you for this feedback!

Page 14 - As you are providing reminders of effect size interpretations, is there a concern that predictions will "cluster" around the effect size boundaries of small, medium, and large? Is there a statistical consequence to such clustering if indeed it occurs (e.g., reduction in variance)? Is there any way to encourage full use of the scale rather so this is minimised if there is a statistical concern?

Thank you for this comment. The concern about clustering around the given effect size anchors is a valid one. We likewise considered that our participants might anchor on these values when making their predictions. We chose to provide these rules of thumb to help participants complete what will likely be a difficult task.

Based on your comment, we have decided to add information about the range of the effect size metric to those prediction items to encourage the full use of the scale (see supplemental materials page S3). For instance, the moral experiences effect size predictions will include the following text, “As a reminder, partial eta squared ranges between 0 and 1 and is typically interpreted as follows: small effect ~ .01, medium effect ~ .06, large effect ~ .14.”

Though responses may be impacted by the interpretations we provide, we believe this addition will increase the range of responses.

As for statistical consequences, our analyses do have distributional assumptions that clustering may impact. We will examine the distribution of the effect size predictions before determining whether to use a Pearson or Spearman correlation for the aggregate subsample level analysis. Regarding our prediction level analyses, multilevel linear models are fairly robust to violations of the assumptions that residuals and random effects are normally distributed (e.g., Schielzeth et al., 2020). Thus, we should still be able to identify relationships between predicted effect sizes and actual effect sizes at both levels of analysis if they exist.

Finally, we intend to discuss the limitations of our chosen measures in the Stage 2 discussion of the manuscript.

New Reference (not in manuscript): Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., ... & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141-1152. https://doi.org/10.1111/2041-210X.13434.

Related to this, as you are converting the effect sizes to Cohen's d before analyses, might it be worth using this metric when asking participants to estimate the effect size of the focal effect rather than using the original effect size metric (e.g., odds ratios)? As you mention in the paper researchers (perhaps!) have a better feel for Cohen's d so might lead to more accurate estimates.

We likewise had considered this approach, but worried that it would increase difficulty and confusion in some cases. For instance, asking for a Cohen’s *d* estimate for a correlational effect would increase the difficulty of the prediction task. Further, as we provide participants with the details of how the focal effect analysis will be conducted, using a metric that does not correspond to the planned analysis may cause confusion. In any of the four studies where the project is a replication, such as the *Moral Experiences* project, we will additionally provide the original effect size in the same metric.

Thus, while we do think Cohen’s *d* is the best metric for analytic purposes, we will ask for the predictions using the metric that coheres with the focal effect analysis.

Page 18 - Thank you for including the R markdown html for the power analyses. This was very clear and comprehensive and sets the standard for how power analyses should be reported in studies.

Thank you for this feedback!

Page 20 - "No participants will be excluded from the analytic dataset". Will participants with missing data (e.g., dropping out halfway through) still be included? If so, how will missing data be handled (e.g., imputation?)

We will include all available data in each analysis and missing data will not be replaced.

Due to the way we have ordered, randomized, and scored our tasks, participants completing any percentage of the study will have relevant data for one or more analyses. Because of the setup of our primary analyses (i.e., examining predictions at the level of subsample or prediction rather than participant), we will be able to use all the available data in those tests, even if participants do not complete all the predictions, without the use of multiple imputation or other data replacement methods.

We have clarified our handling of missing data in the manuscript.

page 23 (addition in bold), “All participants with available data on the relevant variables will be included in a given analysis. **Missing data will not be replaced.**”

Signed: Jim Grange.

Thank you again for your review and helpful feedback.

Michèle Nuijten

The aim of this Registered Report is to 1) investigate whether researchers can accurately predict generalizability of effects across global regions and 2) if certain researcher characteristics are related to prediction accuracy.

I think this is an interesting and relevant research question and overall, the proposed research plan is solid. I think this is an innovative and relevant way to make use of the extremely rich data of the PSA.

I do still have some questions/remarks I would like the researchers to address. I will copy them in a numbered list below. I think the majority of these points could be addressed relatively easily.

I am very curious to see the results of this project.

Signed,

Michèle Nuijten

Thank you for your review and feedback.

1. I miss any remarks on the generalizability of the results of *this* study. It seems to me that although PSA researchers are likely from all over the world, they might not be representative for “psychology researchers” in general. I think it is important that the authors “practice what they preach” in a way, and make some sort of statement about this.

Thank you for this suggestion. The generalizability of our results will be limited in several ways, including, as you mention, by our participant sampling strategy. However, we agree that members of the PSA may not be representative of psychology researchers generally. For instance, their interest in or knowledge about the generalizability of psychological effects may be higher than the average researcher. We will acknowledge such possibilities and how they may have impacted our results in the discussion section of the Stage 2 manuscript.

While the main discussion of these limitations of the research will be reserved for the Stage 2 manuscript and discussion section (e.g., we will include an explicit Constraints on Generality statement), we have added a brief preview of these concerns to the introduction.

page 12, “Given the focus of the proposed research on generalizability prediction, limitations on the generalizability of our results should be acknowledged. For instance, methodological features of the research, such as our chosen sample and how we will select the studies and their focal effects, will likely produce results that do not generalize to all researchers or all effects. We will discuss our findings with these constraints in mind.”

2. In the introduction, the authors clearly show a gap in the literature: there is research on prediction of effects, there is some research about predicting generalizability over time, but no prediction of effects across global samples. I think it would strengthen the argument if the authors also indicate what the main benefit of answering this question would be. Is the main point to assess potential generalization bias (as mentioned on p. 7)? Or could we use the results from this study in some way to come up with guidelines or advice for researchers along the lines of the Constraints on Generalizability statement?

Thank you for this suggestion. We have added a brief description of some potential benefits of the research.

page 12, "Nevertheless, taken together, our results will provide insight into how researchers understand the generalizability of psychological effects across cultural contexts. Our findings may inform recommendations for researchers discussing the constraints on generality of their research or help determine whether predictions should be used to prioritize effects for future research on generalizability.”

3. The authors mention a host of different measures for researcher characteristics. I wonder about the validity of these measures. I would like to see some validity information about the scales that are planned to be used (open-minded thinking, need for cognition, and any other I'm forgetting here).

Thank you for this suggestion. We will use two scales, the Comprehensive Intellectual Humility Scale and a measure of Actively Open-Minded Thinking. We have added validity information for these two measures to the manuscript.

page 19, "The Comprehensive Intellectual Humility Scale has high internal consistency (α = .82 - .89) and shows evidence of convergent, discriminant, and predictive validity (Krumrei-Mancuso & Rouse, 2016). The specific measure of AOT has not been formally validated, but researchers have previously demonstrated its relationship with prediction accuracy (Haran et al., 2013; Mellers, Stone, Murray et al., 2014)."

4. I think the paper could use an additional sentence or two on what it means to be a "member of the PSA", since this describes the participant pool.

Thank you for this suggestion. We have added information about PSA membership accordingly.

page 13, "PSA Membership requires agreement to support the mission and core principles of the PSA and adherence to the PSA's code of conduct. All contributors to PSA projects must first become members."

5. Three of the four PSA projects still need to be selected. What is the sampling plan for this? Are there any predetermined criteria that a project needs to meet before it is selected? Is there a set timeline (e.g. the first three projects that have a finished protocol)?

The remaining three studies will be selected according to the same requirements and procedures as the first study. We describe this process in the manuscript (see page 13).

As for timeline, study selection for the second call for studies is underway, and we anticipate that the remaining three studies will be provisionally accepted by the end of August, 2023. The research protocols should be finalized by early 2024.

We have chosen not to add our anticipated timeline to the manuscript as these dates are only estimates, and we feel that it is irrelevant to understanding and evaluating our methods or results.

---

6. Again, I may have missed it (if so my apologies and please ignore this comment), but I don't remember seeing a clear explanation of the different "sources" of data from each of the regions. From the context I'm deducing that each region/country will collect a university sample and a non-university sample? Please make sure that this is clarified in the text.

The projects will aim to collect both types of samples within each country. We have added "in countries" to the following sentence in our introduction of the research to help clarify our intent.

page 10-11, "Selected projects will test the generalizability of psychological phenomena across university and community samples in countries around the world."

We further define the sample source variable, "sample source (i.e., university vs. community)", in both the Methods (see page 15) and Analysis Plan. We have moved this reminder to earlier in the Analysis Plan because of other edits (see page 24 or the quoted text in our response to your point 9 below).

---

7. The authors plan to measure many additional variables at the researcher level. I may have missed it, but for some of them, I can't seem to find any hypotheses or analysis plans (task difficulty, demographics). For the ones that are planned to be included in the accuracy analyses (p. 21-22), I do not completely understand the rationale behind the analysis plan. If I understand the authors correctly, they first intend to correlate all measured researcher characteristics to prediction accuracy, and when a correlation is found (is there some sort of cut-off here?), the characteristic will be added to a regression model? I'm not familiar with this analysis strategy, and I wondered if it would not be more straightforward to simply run a regression model including all the measured researcher characteristics and look at the resulting regression coefficients to judge which characteristics are important. Finally, there are *a lot* of characteristics added; did the authors take into account the risk of Type I error inflation due to the large number of tests?

While we will measure many researcher characteristics, only six will be tested as predictors of generalizability prediction accuracy in our confirmatory analyses. These characteristics were chosen based on prior research suggesting they may relate to prediction accuracy. As described on page 26, these are “prediction confidence, involvement in the project, highest degree, self-rated expertise in the project subfield, intellectual humility, and actively open-minded thinking”.

The raw correlation coefficients will not be examined. We apologize for the confusing use of the phrase “potential correlates” to describe the variables we would be testing, and we have changed the wording accordingly. We will test these characteristics in separate models as we are interested in whether they relate to prediction accuracy.

Our hypotheses are simply that these six characteristics will predict accuracy in their individual models. In the design table, we outline how the research question, hypotheses, and analyses align (see page 45-47).

However, your point about the importance of the characteristics in predicting accuracy is an interesting one. Thus, we have chosen to add models examining these predictors simultaneously to see if they independently predict accuracy.

Here is our updated analysis description:

page 26, “These models will serve as the base models for our analyses examining what researcher characteristics relate to prediction accuracy. Tested characteristics will include prediction confidence, involvement in the project, highest degree, self-rated expertise in the project subfield, intellectual humility, and actively open-minded thinking. These characteristics will be added as predictor variables and tested in separate models. We will also include all six variables in the same models to examine whether they independently predict accuracy.”

While we do intend to examine a number of additional other researcher characteristics in exploratory analyses (i.e., “prediction difficulty ratings, researcher beliefs, and the other measures of research involvement, experience, and expertise”; see page 26), we will interpret these findings cautiously due to the potential for type 1 errors.

---

8. On p. 20 the authors state they will calculate “a correlation” between mean probability estimates of finding an effect in the subsamples and the binary outcome variable. Please specify the type of correlation that will be used (considering that the probability estimates are likely non-normally distributed, and one of the variables is binary).

Thank you for this suggestion. In response to your concern about normality, we will now examine the variable distributions before determining our approach. If the probability estimates appear normally distributed, the relationship between the binary outcomes and probability estimates will be examined as a point-biserial correlation coefficient; otherwise, we will use a Spearman correlation. If the effect size outcomes and predictions are normally distributed, we will examine their relationship as a Pearson correlation coefficient; otherwise, we will use a Spearman correlation. We have updated the description of the analyses to include this information.

page 24-25 (addition in bold), “Specifically, 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. **If the continuous variables appear normally distributed according to quantile-quantile plots, we will use point biserial and Pearson correlations, respectively, for these tests. Otherwise, we will use Spearman correlations.**”

---

9. This is a complex project with data and predictions at many different levels. This can sometimes result in unclear sentences. E.g.: “We will compare the responses on the overall prediction items to the overall results within each study.” (p. 22). It is not clear to me which variables and items this refers to, exactly. Do the authors refer to all prediction items? At which levels? And what is meant by “overall results within each study”? Which results? Effect sizes? I had similar difficulties with wrapping my head around the distinction between the “aggregate-level analyses” and “prediction-level analyses”. It may help improve clarity if the authors also explain what the substantive difference/advantage/interpretation is of having these two levels in the analysis.

Thank you for this feedback.

Our research includes predictions of outcomes (significance at *p* < . 05) and effect sizes for the focal effects both overall and within each subsample. Thus, the “overall” items and results refer to the single item predictions for both effect sizes and outcomes and their corresponding results of the focal effect for each study. As these are not relevant to generalizability, we are focusing on the subsample items in our primary analyses.

We have clarified our meaning by editing the sentence you identified.

page 27, “To examine whether researchers accurately predicted the study-wide focal effect outcomes and effect sizes, we will compare the single-item overall predictions to their corresponding overall results within each study.”

To answer our primary research question, we will examine the relationships between predictions and results (both in terms of effect sizes and significance outcomes) at two levels of analyses. First, we will examine how aggregated participant predictions at the level of subsample relate to the subsample results using correlations. Second, we will examine how the raw predictions (participants make 20 predictions of each type) are related to the subsample results in multilevel models.

We have added clarifying language throughout the manuscript to better define the two levels of analysis. For example, we added “subsample level analyses” to the following sentence:

“We expect that these relationships will emerge in both aggregate subsample level analyses and prediction level analyses.” (see page 11).

We also now justify the use of both approaches as requested.

page 24, “We will examine the relationships between the predicted and actual results at two levels of analysis. Aggregate subsample level analyses will estimate the relationships between predictions and results on average, and prediction level analyses will estimate the relationships while examining and accounting for variability according to study, sample country, sample source (i.e., university vs. community), and participant researcher.”

Thank you again for your review and helpful comments.

Matthias Stefan

The study examines, in the field of psychology, how well researchers can predict the generalizability of psychological effects and whether and which researcher characteristics influence prediction accuracy.

My report is directly addressing the authors. My general assessment is that the research question is well defined, valid and timely; the hypotheses are well stated, coherent and precise; the procedure is feasible and the methodology proposed is sound. There is much to like about the registered report. Therefore, I have only few comments:

Thank you for your review and feedback.

We have split up some of your comments into multiple subpoints so we could respond to them more clearly.

The following points are major to me. I suggest to address them in a revised version of the registered report if you agree with my concerns and if I did not miss anything:

- One major point is on the measure of generalizability: you ask researchers to estimate the “probability that a statistically significant focal effect ($p < .05$) in the hypothesized direction will be observed”. In the paper you define this as “estimate [of] the probability that the expected effect will be observed”. If I understand correctly, this definition does not include the studies’ effect sizes. For example, if an effect is significant and has the same direction, but is substantially lower, you would still

define this outcome as generalizable. While I think this is a fair approach, it still merits some open discussion. If I am wrong, maybe you can clarify your measure.

We have included predictions of both effect sizes and outcomes (i.e., statistical significance in the expected direction) in our research to capture both aspects of generalizability. These two measures and approaches are described throughout the manuscript. Thus, we will examine researcher's ability to predict both the presence and magnitude of the effects in separate analyses. We will discuss both the generalizability of the focal effects, and the prediction accuracy of the researchers, considering both of these metrics.

We agree that the reasoning behind our choices of measures could be better explained. Thus, we have added justification in the Methods section.

page 15, "Then, data from each project will be analyzed to produce a binary focal effect outcome both overall and within each subsample–a combination country and sample source (i.e., university vs. community). The binary outcome measures will be the significance of the focal effect analysis at p < .05 in line with prior empirical outcome prediction research (e.g., Benjamin et al., 2017; Camerer et al., 2016; 2018; Delios et al., 2022; Dreber et al., 2015; Hoogeveen et al., 2020). We will also calculate an effect size for the focal effect both overall and within each subsample. While our dichotomous measure has limitations due to the impact of varying subsample sizes on the outcome, including both types of measures will allow us to examine predictions regarding both the presence and magnitude of the focal effects."

- On the point of effect sizes, it would be interesting to have some discussion of effect sizes in your study on researchers' prediction accuracy. I wonder what we can learn from (very) small effect sizes in your study. This is an important point, since your power analysis indicates more than 90% power to detect (very) small effects. If you find one, you should be able to determine if the effect is relevant or not in order not to be "overpowered" in your study. I wonder why you chose such high power? I was missing a discussion on this point.

Thank you for this feedback. Many journals that accept registered reports require at least 90% power – and some even require greater than 95% power – to detect relevant effects. Given the lack of previous research on generalizability predictions, we powered our research to detect small effects for most of our analyses to increase confidence in our results. While we believe that even very small relationships are interesting and important in the context of the topic, we do intend to discuss the size of any effects we identify to qualify our results.

We have added the following statement to our power analysis section to explain our choice of effect sizes:

page 21, “Given the lack of previous research on generalizability predictions, we aimed to ensure that we had sufficient power (90% with α = .05) to detect small effects in most of our analyses, thereby increasing confidence in our results.”

- One rather general comment is on the choice of the subject pool: your study is restricted to psychological researchers, i.e., experts. While this might be the most interesting pool, it restricts generalizability to other researchers and potentially focuses on a biased group. For example, such researchers might be overoptimistic regarding their own field – or, alternatively, overly critical. Just to be clear, I do not suggest to conduct a wider study, I just think it would be nice to see a discussion of the pool choice and the implications. One such implication is generalizability since your wording (e.g., “researchers”) could suggest a generalization that might be unjustified. The discussion of researcher groups and their replicability forecast in Gordon et al. (2020) could be helpful.

Thank you for this feedback.

Your point about the generalizability of our results is an important one. We intend to discuss constraints on generality, including our choice of sample, extensively in our discussion section of the Stage 2 manuscript. However, we have added a brief preview of this limitation in our introduction.

page 12, “Given the focus of the proposed research on generalizability prediction, limitations on the generalizability of our results should be acknowledged. For instance, methodological features of the research, such as our chosen sample and how we will select the studies and their focal effects, will likely produce results that do not generalize to all researchers or all effects. We will discuss our findings with these constraints in mind.”

We have also added a justification for our participant pool:

page 13, “We chose to target PSA members to have a clearly defined sampling strategy and enable examination of the relationship between researcher involvement and prediction accuracy.”

Thus, we do intend to examine what features of researchers, such as expertise and research involvement, are associated with generalizability prediction accuracy.

Thank you for the suggested reference. Gordon et al. (2020) reported research examining forecasts of replicability across different fields. However, and importantly, these forecasts were not compared to the actual results of the replications. While they did find that participant’s replicability estimates were slightly higher for topics within their field than those outside of it, whether those estimates were more or less accurate is unknown.

- I was a bit confused by the country/region choice: what exactly does "region" refer to.

Thank you for this feedback. Some geographic regions that may be represented in the studies are not universally recognized as "countries" (e.g., Taiwan). Generally, researchers reporting international collaborations need to be quite cautious with their terminology for such regions so that all contributors can be listed as authors on the manuscript. We had likewise avoided the term "country" despite intending to group samples by country (generally defined). However, as the terminology concern does not directly apply to authors of this particular manuscript, we have changed our wording to "country" throughout the manuscript to increase clarity.

How do you derive at 15 countries and why do you choose 10 out of those 15?

As stated on page 22, we chose to estimate power "assuming that 15 countries will have subsamples of at least 100 university participants and 100 community participants in each associated project. This number was based on minimum recruitment expectations." The actual number will likely be higher and will vary from project to project. Thus, we also estimated power assuming 30 countries per project.

We chose to randomly assign 10 countries for predictions to keep the study length reasonable for participants. We have added this justification as a footnote into the manuscript.

page 17, "7. As participants will complete four predictions per country, we chose to assign ten countries to limit study length and participant fatigue."

And why are prediction items presented in one out of eight possible orders instead of just using randomization?

We chose to keep the order of prediction items the same within each participant to reduce confusion. Eight possible orders were determined so that either the type of sample or the type of prediction would be paired together with the order of the other factor consistent within each type. We reasoned that a logical pattern presented consistently would make answering these questions easier for participants.

We have added this information to the manuscript and now list the specific orders to our supplementary materials.

page 17, “For ease of responding, the eight orders pair together either the type of sample or the type of prediction with the other factor ordered consistently within each type; item order will remain consistent within participants.”

page S3, “Participants will be randomly assigned to one of eight possible orders of the four items for each country, with E = effect size, O = significance outcome, U = university, and C = community:

EU - EC - OU - OC
EC - EU - OC - OU
OU - OC - EU - EC
OC - OU - EC - EU
EU - OU - EC - OC
OU - EU - OC - EC
EC - OC - EU - OU
OC - EC - OU - EU”

Similarly, why are you only including subsamples with 100 or more valid participants (page 14)? There must be a reason for this choice, but I did not find it. In the end, I am not sure I fully understood the details of data collection.

We wanted to ensure that each subsample would be powered to detect at least a medium sized effect. Due to likely differences in how the effects would be tested across the studies, the precise power sensitivity of a sample of 100 participants will vary somewhat.

We have added justification for this choice.

page 15, “Only subsamples (e.g., university students in Colombia) with 100 or more valid participants will be used in analyses. This number was chosen to ensure power to detect at least a medium sized effect in each subsample; however, the minimum detectable effect in each subsample will depend on the focal effect analysis as well as the subsample size.”

Further, in the discussion section of our Stage 2 report, we will discuss how the differences in power to detect effects between the subsamples within a study and between the four studies may have impacted our conclusions.

We hope that our responses and corresponding edits have improved your understanding of the data collection details.

- The following statements from your registered report are too general to be understood by the reader:

“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.”
and
“Single page project descriptions will be generated for each study and approved by the proposing authors of the project as a quality check.”
I think it would be helpful to give more details on your procedure.

Thank you for this feedback. We have expanded the description of how the focal effects will be chosen and further clarified how the project descriptions will be created.

page 14, “A single focal effect will be selected from each project based on input from the proposing authors. We will ask the proposing authors to identify effects from their project that meet the following criteria: 1) answers a central research question, 2) results from an inferential statistical test, and 3) is grounded in theory and supported by previous research. They will be told to prioritize simple and easily described effects tested at α = .05 if multiple effects meet this criteria. If the proposing authors suggest more than one focal effect, we will choose from among these randomly. We selected the following focal effect for the Moral Experiences project based on this procedure: Experiences of moral events will be associated with higher momentary happiness than experiences of immoral events.”

page 15, “We will compose single page project descriptions for each study that will be approved by the proposing authors of the project as a quality check. Descriptions will include the study title, a study summary, a statement of the focal effect, details of how the focal effect will be tested, and any necessary references.”

Here are some more minor points and I think they can be addressed in the main paper (i.e., after data collection), if you agree with them or consider them helpful:

- In general, it was not always easy for me to follow the manuscript. Some of my comments might be helpful for writing-up the paper:

o I found the discussion of concepts interesting. However, for me it was difficult to follow this discussion until the second half of page 4, where you define which concept you are focusing on (generalizability across cultural context). Before this, I was missing a clear definition of relevant concepts in the context of your study. For instance, on page 4 you state that generalizability refers not only to settings and samples, but also to methods and measures. In your study, however, you only focus on the former.

We believe the first paragraph of the introduction clearly sets up our focus on generalizability across samples and settings. We begin with the problem of researchers in psychology making general claims based on narrow samples and end with a statement of our research goals, “The present research seeks to fill this gap by investigating whether researchers can accurately predict when psychological effects will generalize across regions and sample sources and examining what researcher characteristics relate to their prediction accuracy.”

We would argue that this writing sufficiently sets up the present research before we turn to a brief discussion of the concept of generalizability more broadly. Providing a more general overview of generalizability is needed to situate our research within the topic.

Moreover, a common understanding that replicability is related to statistical factors (such as sampling error), while generalizability is related to samples (e.g. participants, time period, cultural factors, etc.). This understanding might be too narrow and definitions are not clear. Even more so, it would be helpful to clearly define the two concepts in the context of your study from the very beginning and maybe focus less on replicability.

We see replication and generalization as closely interrelated concepts. As we explain on page 3-4, “...replication studies have been described as tests of generalizability that can help identify the boundary conditions of an effect (Nosek & Errington, 2020). Failures to replicate an effect may indicate that it does not generalize to the conditions of the replication study...”

From this perspective, all replications are tests of generalizability, but tests of generalizability are not always replications (i.e., their results may not impact confidence in the original findings). Further, generalizability, as we explain, can refer to many aspects of the research beyond samples and settings. Given this close relationship among concepts, and the previous research on replication prediction we cite, we believe that our coverage of replication in the introduction is appropriate.

We have edited our writing to further clarify our definition of these concepts.

page 3-4 (additions in bold), “**Defined broadly, replications test the reliability of a research finding with different data (Nosek et al., 2022)**. Accordingly, replication studies have been described as tests of generalizability that can help identify the boundary conditions of an effect (Nosek & Errington, 2020). Failures to replicate an effect may indicate that it does not generalize to the conditions of the replication study**, such as its sample or setting.**”

page 4 (edited to increase clarity), “The generalizability of a research finding refers to its applicability to not only other samples or settings, but also to other methods and measures.”

Another example again can be found on page 4: you discuss that generalizability can be related to methods or interpretation, without clearly separating the two concepts. It would be helpful to the reader early on to understand what you are focusing on in your study and which concepts are (ir)relevant.

In this section, we reference an argument that widespread misalignment between verbal and statistical expression constrains generalizability and contributes to the low rates of replicability in psychology (Yarkoni, 2022). This mismatch between claims and methods, while perhaps not directly relevant to the present research, is an important consideration in the broader discussion of generalizability in psychology.

We have edited this paragraph to make the writing more clear.

page 4, "For a given study, such features likely contribute to whether a hypothesized effect is observed; failures to generalize may arise from methodological sources. Accordingly, Yarkoni (2022) argued that the low rates of replication in psychological research can be explained in part by the misalignment between verbal and statistical expressions. Researchers often make general claims based on specific operationalizations and fail to account for important features of the research like stimulus variation in their models."

o You are asking participants about their prediction of moderators. This part of the study came as a surprise as moderators are first mentioned on page 14. It has not been completely clear how it relates to your research questions.

Thank you for this feedback. The moderation predictions are not a primary focus of the research. However, we believe their examination will provide insight regarding researcher's understanding of generalizability. This aspect of the research is first introduced in the Present Research section (on page 11-12).

We are testing whether culturally relevant variables moderate the focal effects across individuals and subsamples. As moderators explain the variability of an effect, and systematic variability of an effect based on features of a sample is directly relevant to its generalizability, researchers' ability to predict moderation results could indicate an understanding of *why* effects do or do not generalize across cultural contexts.

We have added an explanation of the relevance of this secondary question to increase the clarity of its contribution to the research.

page 12, "We included these predictions in the research because generalizability can depend on systematic variability in effects based on participant and sample features (i.e., moderators). Thus, accurate moderation predictions may suggest that researchers understand why effects do or do not generalize across cultural contexts.”

o In the main paper, you could introduce the Psychological Science Accelerator, since readers from other fields (such as myself) might not be familiar with it.

Thank you for this suggestion. We have moved our description of the PSA from the Methods section to where the PSA is first mentioned to provide this introduction.

page 10, “Our investigation will focus on the four projects selected by the Psychological Science Accelerator (PSA) in response to two special calls for studies. The PSA is a globally distributed network of researchers in psychological science with members from all six populated continents that coordinates data collection for crowdsourced research projects (Moshontz et al., 2018).”

- Another minor point: Is the potential bias of predictions by desired results really motivated reasoning or not rather confirmation bias? Of course, the two concepts are closely related and might not even be clearly distinguishable.

We agree that these concepts are closely related and difficult to differentiate. In our understanding, the confirmation bias occurs when evidence is gathered (and perhaps evaluated or remembered, depending on the definition) in ways to confirm pre-existing beliefs. Motivated reasoning occurs when an individual’s goals or motivations affect cognitive processes of reasoning and judgment (Kunda, 1990). While the confirmation bias may be a mechanism of motivated reasoning, we believe that the latter concept better reflects the potential ways in which predictions may be influenced by involvement in the research (e.g., that desiring certain research outcomes impacts responding).

Reference:

Gordon, M., Bishop, M., Chen, Y., Dreber, A., Goldfedder, B., Holzmeister, F., Johannesson, M., Liu, Y., Tran, L., Twardy, C., Wang, J., & Pfeiffer, T. (2022). Forecasting the Publication and Citation Outcomes of Covid-19 Preprints. Royal Society Open Science, 9: 220440.

Thank you again for your review and helpful feedback.