Registered Report: Researcher Predictions of Effect Generalizability Across Global Samples

Kathleen Schmidt[1], Priya Silverstein[1,2], and Christopher R. Chartier[1]

[1] Department of Psychology, Ashland University

[2] Institute for Globally Distributed Open Research and Education

Author Note

Kathleen Schmidt https://orcid.org/0000-0002-9946-5953

Priya Silverstein https://orcid.org/0000-0003-0095-339X

Christopher R. Chartier https://orcid.org/0000-0002-4568-4827

This research is funded in part through a grant from the John Templeton Foundation (#61825). The funders have no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors thank Charles R. Ebersole for his work preparing the grant.

Correspondence regarding this article should be directed to Kathleen Schmidt, Department of Psychology, Ashland University, Ashland, OH, 44805, United States of America. Email: kschmi13@ashland.edu.

Abstract

The generalizability of effects is an increasing concern among researchers in psychological science. Traditionally, the field has relied on university samples from Europe and North America to make claims about humans writ large. The proposed research will examine researcher predictions regarding the generalizability of four psychological effects. Predictions of outcomes and effect sizes overall and within regional subsamples will be compared to the results of four large scale international collaborative studies. We will also examine relationships between researcher characteristics and prediction accuracy. Our investigation will reveal whether researchers can accurately predict the generalizability of these psychological effects across cultural contexts while offering insight ~~on~~into what features of the researchers are related to their prediction accuracy.

Keywords: generalizability, psychology, metascience, prediction, forecasting, context sensitivity, cross-cultural differences

Researcher Predictions of Effect Generalizability Across Global Samples

Psychological science has typically relied on convenience samples, such as undergraduate students at universities (Sears, 1986) or, more recently, individuals recruited online through services like Amazon's Mechanical Turk (MTurk; Anderson et al., 2019), located in Western and Northern regions of the globe. Described as psychology's WEIRDness problem (Henrich et al., 2010), evidence from these narrow samples is nonetheless used to make general claims about human psychology (e.g., DeJesus, et al., 2019; Rad et al., 2018).[1] The recognition of this and other limitations of research practices has led to an increasing concern about the generalizability of psychological effects (e.g., Haeffel & Cobb, 2022; Tiokhin et al., 2019; Yarkoni, 2022). While the field has begun grappling with how to address these issues (e.g., Apicella et al., 2020; Deffner et al., 2022; Hruschka et al., 2018; Medin et al., 2017; Rad et al., 2018; Syed & Kathawalla, 2022), researchers' understanding of generalizability in psychology has not been examined empirically. 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.

Generalizability in Psychology

Over the last decade, calls for reform have focused on changing research practices to increase the transparency and replicability of psychological science (e.g., Munafò et al., 2017; Vazire, 2018). ~~Within this discussion~~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

[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).

to the conditions of the replication study, such as its sample or setting. Indeed, context sensitivity is often invoked when replications fail (e.g., Crisp & Birtel, 2014; Schnall et al., 2014; Schwarz & Strack, 2014; Shih et al., 2014; Van Bavel et al., 2016). Specifically, sample differences between the original and replication studies may be cited when effects fail to replicate (e.g., Cesario, 2014; Dijksterhuis, 2018; Ferguson et al., 2014; Gilbert et al., 2016; Schnall et al., 2014). However, researchers often make unwarranted universal claims and do not suggest that their findings could be context dependent (e.g., Simons, 2014; Simons et al., 2017).

~~Importantly, the~~The generalizability of ~~an effect refers not only to other settings and samples~~a research finding refers to its applicability to not only other samples or settings, but also to other methods and measures. In a meta-analytic context, the heterogeneity of an effect is how it varies across any or all of these study features. For a given study, such ~~study~~ 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 ~~may be best described as a "generalizability crisis" that~~ can be explained in part by the misalignment between verbal and statistical expressions. Researchers often make general claims based on specific operationalizations~~, failing~~ 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).

Generalizability across Cultural Contexts

The concept of generalizability is broad and may even be conflated with representative sampling and other features of the research key to internal and external validity (e.g., Kukull & Ganguli, 2012). ~~Here~~In the present research, we focus on one facet of generalizability: the heterogeneity of effects across cultural contexts. While the lack of diversity in psychology's researchers and samples is becoming widely acknowledged (e.g., Arnett, 2008; Henrich et al., 2010; Medin et al., 2017; Nielsen et al., 2017), the field has made little progress in changing ~~this practice~~sampling practices (Apicella et al., 2020; Rad et al., 2018; Thalmayer et al., 2021). Researchers have argued for explicit discussion of constraints on generality (Simons et al., 2017) or increased specificity of population based claims (Rad et al., 2018) to mitigate the problem of overgeneralization.

Perhaps surprisingly, recent large-scale international collaborations have found little effect heterogeneity in their investigations. 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). ~~Notably, those~~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. Further, the samples were mainly undergraduates from high income countries, and thus, relatively "WEIRD". Potentially, without formal theory to guide cross-cultural investigations (e.g., Muthukrishna et al., 2020), international collaborations may fail to identify meaningful cultural variation that could be otherwise found with theoretically justified cultural sampling. Indeed, cultural variability of psychological phenomena has been observed in many investigations.[2] In one illustration of the need for strong generalizability tests, Tiokhin et al. (2019) failed to replicate social discounting effects in rural Bangladesh and Indonesia. Other large-scale global studies have found mixed evidence for generalizability; in research on face perception (Jones et al., 2021) and moral judgment (Bago et al., 2022), effects generalized across world regions for only a subset of outcomes and analyses. 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. Whether researchers can predict when and how such cultural heterogeneity will be found, however, remains unknown.

Predicting Research Results

Researchers have demonstrated the ability of scientists (e.g., Forsell et al., 2019) and even laypeople (e.g., Hoogeveen et al., 2020) to predict replication and other empirical outcomes through prediction markets and belief surveys. While prediction markets have theoretical advantages over other approaches (e.g., Plott & Sunder, 1988), and some researchers have demonstrated the superiority of prediction markets over survey beliefs in replication prediction accuracy (Dreber et al., 2015), others have found similar outcomes for both approaches (Camerer et al., 2016, 2018; Viganola et al., 2021). For relative effect size estimates, survey beliefs may even outperform markets (Forsell et al., 2019). In a recent meta-analysis of replication predictions, Gordon et al. (2021) found that prediction market prices and average survey responses were similarly correlated with outcomes (i.e., r = .581 and r = ~~0.564~~.564, respectively). Overall, research outcome predictions appear to have strong relationships with actual research results.

[2] For examples of reviews and perspectives on cultural differences in psychology, see Apicella and Barret (2016), Boer and Fischer, (2013), Henrich (2015), and Kline et al. (2018).

While research examining replicability predictions is a new but growing area (see also Fraser et al., 2023), research examining predictions of research generalizability is much more limited – we are aware of only one prior investigation in this area. Delios et al. (2022) asked researchers to predict the likelihood that findings from management research using archival data would generalize to data from different time periods. Predictions of generalizability were positively associated with outcomes at the individual and study level. The correlations between aggregated predictions and outcomes were modest (r[22] = .259) and the individual level effects appeared small.[3] They found no evidence that researchers over- or underestimated the generalizability of effects; the overall predicted generalizability rate (57%) was close to the observed generalizability rate (55%). While providing initial evidence for generalization prediction accuracy, this investigation examined predictions of generalizability across time periods rather than across geographic regions.[4] Thus, whether researchers can predict generalizability across sample regions and sources remains an open question. Further, whether researchers anticipate that research results will generalize across cultural contexts more or less than they do in reality remains unknown.

Peters et al. (2022) argued that scientists demonstrate a generalization bias in which they generalize their results to broader populations than is warranted. 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. If this generalization bias applies to the present research, researchers may predict more generalizability of psychological effects than is found. However,

[3] No prediction test statistic or effect size was reported, but the model explained little variance (R_{adj}^2 = .001).
[4] In their generalization analyses, Delios et al. (2022) included some tests of generalization to other geographies. However, they did not ask forecasters to predict the outcomes of these tests.

skepticism about the effect, focus on the potential for cross-cultural differences, or awareness of generalizability concerns in research could instead produce underestimations of generalizability.

Variations in Prediction Accuracy

Several researcher characteristics may relate to the accuracy of their generalizability predictions. For example, research expertise and experience may increase prediction accuracy. However, previous research has found that experts may be equally skilled as, or only nominally better than, non-experts at predicting outcomes in their domains of expertise (see Camerer & Johnson, 1991). Prediction market studies of accuracy in predicting research replications (e.g., Camerer et al., 2016, 2018; Dreber et al., 2015) generally examine aggregate effects rather than individual predictions. Thus, research on what characteristics of researchers increase their predictive accuracy is lacking, and existing evidence regarding whether expertise positively relates to accuracy is mixed.

Hoogeveen et al. (2020) found that, in aggregate, lay people were able to predict research replication outcomes with some accuracy (59%) but may be less accurate than experts (~~Hoogeveen et al., 2020~~65-72%). However, providing details about the original effects increased lay persons' accuracy~~, perhaps even to the level of experts~~ (67%). McBride et al. (2012) examined expert prediction accuracy for ecology research outcomes and found no consistent effects of self-assessed expertise, experience, or publication record. In their investigation of outcome predictions for behavioral economics experiments, DellaVigna and Pope (2018) found that researcher expertise did not positively correlate with accuracy. However, academic experts were more accurate than non-experts (e.g., MTurk workers), but only according to prediction error measures and not on rank order measures of accuracy. Benjamin et al. (2017) examined the accuracy of cancer researchers' predictions of replication outcomes from the reproducibility project: cancer biology. As a group, experts outperformed trainees on some metrics of accuracy but not others. Among experts, publication impact (*r* = - .15) but not topic expertise ~~predicted forecast accuracy. Interestingly, confidence in forecasts appeared to decrease prediction accuracy; however, confidence has been positively related to research outcome prediction accuracy in other research~~related to prediction errors. Interestingly, prediction confidence was positively related to error (*r* = .36); however, other researchers have found positive relationships between confidence and the *accuracy* of empirical outcome predictions (DellaVigna & Pope, 2018). 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.

Involvement in the research may also impact accuracy in generalizability prediction. Familiarity with nuances of the study (e.g., fine methodological details) could increase prediction accuracy. Alternatively, investment in the outcomes of the research could decrease accuracy; the predictions of researchers involved in a project may be less accurate than those not involved because they hope that specific outcomes (e.g., finding the effect) will occur. Predictions may be biased by desired results, perhaps due to motivated reasoning (see Kunda, 1990). However, the evidence for such desirability biases in predicting future outcomes is mixed (Krizan & Windschitl, 2007).

Other researcher characteristics, such as individual differences in cognitive styles, may also relate to prediction accuracy. To our knowledge, these factors have not been examined in regard to predicting research outcomes. 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 ~~score~~scored higher than other forecasters on AOT and other positive individual differences in cognitive style and ability (~~Mandel & Barnes, 2014~~e.g., NFC; Mellers, Stone, Murray et al., 2015). These same variables positively correlated with prediction accuracy among forecasters more generally (~~see also Mellers~~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 (~~Zmigrod et al., 2019), and the ability to identify argument strength (Leary~~*r* = .35, Zmigrod et al., ~~2017~~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.

Present Research

The proposed research will examine researcher intuitions regarding the generalizability of psychological effects across cultural contexts. Our investigation will focus on ~~the~~ four projects selected by the Psychological Science Accelerator (PSA) in response to two special calls for studies ~~that will be carried out by their global network of researchers. These~~. 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). Selected projects will test the generalizability of psychological phenomena across university and community samples ~~from~~in countries around the world. Previous PSA studies have included samples ranging from 11,570 to 25,718 participants from 45 to 89 geographical regions (Bago et al., 2022; Dorison et al., 2022; Jones et al., 2021; Psychological Science Accelerator Self-Determination Theory Collaboration, 2022; Wang et al., 2021). Similarly, the focal projects for the present research will be designed to be highly powered with considerable regional diversity.

For each project, we will survey researchers in the PSA who are project contributors and those who are not prior to the start of data collection. Researchers will complete a series of predictions about a single focal effect from the project. Specifically, researchers will estimate the probability that the expected effect will be observed both overall and within regional subsamples. They will also predict the size of the focal effect overall and within regional subsamples. We will examine researcher predictions about the effects in relation to the research results to determine whether researchers can accurately predict the generalizability of the studied psychological phenomena. Given previous research demonstrating the ability of researchers to accurately predict empirical outcomes, we anticipate positive relationships between predicted and actual outcomes and effect sizes. We expect that these relationships will emerge in both aggregate subsample level analyses and prediction level analyses.

The proposed research will also examine if researcher characteristics are associated with generalizability prediction accuracy. We will test hypotheses that researcher involvement, prediction confidence, expertise, experience, actively open minded thinking, and intellectual humility predict the accuracy of their outcome and effect size predictions. We will include several measures of researcher expertise and experience in the studies but focus on a subset for our confirmatory analyses. In secondary analyses, we will also test if researchers can predict when variables hypothesized to capture relevant cultural differences will moderate the focal effects. ~~Taken together, our results will provide insight into how researchers understand the generalizability of psychological effects across cultural contexts.¶~~

~~**Methods¶**~~

~~All sample size determinations, data exclusions, manipulations, and measures will be reported~~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.

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. 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.

Methods

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). See Table 1 for the Study Design Table. Our methods and analysis plan will be ~~pre-registered~~preregistered and data will be made publicly available. The proposed research will comply with all relevant ethical regulations; we will obtain informed consent from participants for each study. The research protocol has been reviewed and approved by Ashland University's Institutional Review Board (FWA #00014951).

Participants

The proposed research will include four studies linked to four ongoing PSA research projects. For each of these studies, participation will occur after the PSA project's protocol is finalized but before data collection for that project has commenced. Participants will be members of the PSA recruited through the network's newsletter, project updates, and social media. The sample will include researchers in psychology and related fields across educational and career stages from institutions across the globe; the only criteria for participation will be membership in the PSA network. 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. 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. For each investigation, we will recruit approximately equal samples of researchers who are involved in the ~~target~~focal study, such as members of the data collection teams, and researchers who are not. We aim to collect 100 participants per study, or 400 total. This number provides sufficient power for our main hypothesis tests according to power analyses (see below).

Project Selection

~~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).~~ Four projects will be selected from two special calls for studies released by the PSA. These calls invited proposals of research projects investigating generalizability of psychological phenomena across global samples. Selected projects will be funded in part by the John Templeton Foundation. Research questions will be related to one or more of 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. The proposals will undergo two rounds of review, overseen by the PSA's Study Selection Committee. These include an initial feasibility and quality review followed by a full review by experts from within and outside of the PSA and feedback from the PSA network. The Study Selection Committee will then synthesize this feedback to decide which projects to accept.

One project, provisionally titled, "~~Global Moral Codes: A cross-cultural experience sampling study of moral experiences~~A Cross-cultural Study of Everyday Moral Experiences," was accepted from the first call for studies and is in preparation. Referred to here as "Moral Experiences", this project will assess how moral experiences vary as a function of individual and cultural factors, utilizing experience sampling methodology. The remaining three projects will be selected from the second, ongoing, call for studies (see https://psysciacc.org/2022/12/06/second-special-call-for-studies-studying-generalizability-with-global-samples/) and, thus, are not yet known.

Materials

A single focal effect will be ~~chosen~~selected from each ~~study~~project 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~~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 ~~will be that experiences~~project based on this procedure: Experiences of moral events will ~~produce~~be associated with higher momentary happiness than experiences of immoral events.

~~Single page project descriptions will be generated~~We will compose single page project descriptions for each study ~~and~~that will be approved by the proposing authors of the project as a quality check. Descriptions will include ~~a brief overview of the research background and methods followed by the~~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. See the supplementary materials (https://osf.io/fu7dk/?view_only=348484e6e86442e5a43e75e0cf9aa310) for the Moral Experiences project description.

Measures

The full text of all items and scales can be found in the supplementary materials (https://osf.io/fu7dk/?view_only=348484e6e86442e5a43e75e0cf9aa310).

Actual Research Results

~~Data~~For each project, we will implement any preregistered data cleaning procedures and exclusion criteria prior to our analyses. Then, data from each ~~study~~project will be analyzed to produce a binary focal effect outcome ~~(significance at $p < .05$)~~ both overall and within each subsample–a combination of ~~world region (most typically country)~~country and ~~sampling~~sample source (i.e., university vs. community). ~~We will also calculate an effect size for the focal effect both overall and within each subsample~~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. 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] We chose this metric because it is unbounded and easily interpretable.

[5] The formulas for the effect size conversion functions in this package were primarily derived from Borenstein et al. (2009).

The effects of moderators hypothesized to capture relevant cultural differences will be tested separately at the level of individuals and ~~samples~~subsamples. 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 ~~sample~~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.¶

~~***Primary Results Predictions***~~¶

~~Participant researchers will predict outcomes and effect sizes both overall and within regional subsamples. For focal effect outcomes, they will predict the probability that a statistically significant effect (*p* < .05) in the hypothesized direction will be found on a 100-point scale from 0% to 100%. They~~ 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.[6] 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.

[6] Religious affiliation will not be included in the subsample level analyses because it is not easily dichotomized to create a single proportion.

Primary Results Predictions

Participant researchers will predict outcomes and effect sizes both overall and within regional subsamples. For focal effect outcomes, they will predict the probability that a statistically significant effect ($p < .05$) in the hypothesized direction will be found on a 100-point scale from 0% to 100%. 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). Participants will also estimate the effect size of the focal effect. They will be instructed to enter an effect size in an appropriate metric and provided with a reminder of effect size interpretations. Each participant will be randomly assigned to respond to subsample items for 10 ~~regions~~countries from a list of all ~~regions~~countries in which university and community samples are anticipated to be larger than 100.[7] For each ~~region~~country, they will make both outcome and effect size predictions for university and community subsamples separately. ~~Region~~Country presentation order will be randomized. The four prediction items for each ~~region~~country will be presented in one of eight possible orders randomly assigned to each participant. 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. All predicted effect sizes will be transformed to a common metric (i.e., Cohen's *d*) before analyses.

We will examine these predictions relative to actual results both at the prediction level and in aggregate at the level of subsample. For the aggregate analyses, we will compute

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

subsample means for both outcome and effect size predictions. 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).

Prediction Accuracy. 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. For effect sizes, absolute differences will be calculated by subtracting predicted effect sizes from actual effect sizes and taking the absolute value of the result. For both indices, lower values indicate more accuracy than higher values.

Task Difficulty and Prediction Confidence

After making their predictions, participant researchers will respond to four items ~~about the difficulty of the prediction tasks~~ (i.e., two for outcome probabilities and two for effect size estimations) about the difficulty of the prediction tasks and their confidence in their responses on the tasks in a random order. The items will be measured on 5-point Likert-type scales ranging from “Not at all difficult” to “Extremely difficult” or “Not at all confident” to “Extremely confident”.

Moderation Predictions

Participants will predict the probability that culturally relevant moderators will impact the focal effect at the individual and ~~sample~~subsample levels of analysis. After reading how the moderators will be tested, participants in each study will be randomly assigned to complete predictions for five of the potential moderators. ~~A brief definition of the moderator will be provided when necessary followed by the~~For each moderator, they will respond to two items (i.e., individual and sample level predictions), the order of which will be randomized between participants. A brief definition of the moderator will be provided as a parenthetical in the items when necessary. Moderator presentation order will also be randomized.

Researcher Beliefs

Participant researchers will report their holistic assessments of effect generalizability. Specifically, they will be asked to identify how generalizable types of effects are across cultural contexts on a 5-point Likert-type scale ranging from "Not at all generalizable" to "Completely generalizable". These three questions will target the focal effect, effects in the project's subfield of psychology (e.g., moral psychology), and effects in psychology as a field. Participants will also rate their confidence that the psychological phenomenon underlying the focal effect is real on a 5-point Likert-type scale ranging from "Not at all confident" to "Extremely confident".

Individual Differences

Participants will complete the Comprehensive Intellectual Humility Scale (Krumrei-Mancuso & Rouse, 2016) and a measure of Actively Open-Minded Thinking (Baron, 1993; Stanovich & West, 2007) as implemented by Haran et al. (2013).[8] 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., 2015).

[8] Haran et al. (2013) adapted Stanovich & West (2007)'s scale to shorten it from 41 items to 7 items.

The scales and items within each scale will be presented in a random order. For the Comprehensive Intellectual Humility Scale, participants will be asked to indicate their agreement with 22 items, such as "I have at times changed opinions that were important to me, when someone showed me I was wrong," on a 7-point Likert-type scale ranging from "Strongly disagree" to "Strongly agree". The Actively Open-Minded Thinking Scale will include 7 items, such as "Allowing oneself to be convinced by an opposing argument is a sign of good character," measured on a 7-point Likert-type scale ranging from "Completely disagree" to "Completely agree". For both measures, we will score items so that positive values indicate agreement and negative values indicate disagreement. After reverse scoring necessary items, we will compute mean index scores for analyses.

Researcher Involvement, Experience, and Expertise

Participant researchers will indicate their involvement in the particular project (i.e., a collaborator or not), the length of their membership in the PSA, and the number of PSA studies to which they have contributed. They will additionally report the number of large scale international research collaborations outside the PSA to which they have contributed and the number of peer-reviewed publications they have published both in total and as first author. Participants will self-report their level of expertise in the research areas of cross-cultural differences, generalizability, the project subfield (e.g., moral psychology), and the focal effect topic area (e.g., the effect of moral experiences on momentary happiness) on an 8-point Likert-type scale ranging from "No knowledge" to "Very high knowledge." They will also report whether ~~or not~~ they have previously published research in these same areas as an additional proxy for expertise.

Demographic, Educational, and Occupational Characteristics

Participants will be asked to report their age, gender identity, racial and ethnic identity, and country of residence. They will also provide their employment or educational institution type, current position if their institution is a university or college, highest degree obtained, year they completed their highest degree, primary field of study, and subfield of psychology if that was their primary field.

Procedure

Participant researchers will be recruited for each of the four studies separately. After providing consent, they will read a description of the research project and focal effect. Then, they will complete focal effect predictions overall and for their 10 assigned ~~regions~~countries; after which, they will report how difficult they found the prediction tasks and their confidence in their predictions. Next, they will be asked to make predictions about potential moderators and to answer items about their beliefs about generalizability. Then, they will complete the individual differences questionnaires. Finally, they will report their research experience and expertise and their demographic, educational, and occupational characteristics. Participants will create unique identification codes to track their participation across studies. After completing the study, participants will be redirected to a separate survey to enter their personal information for compensation purposes.

Power Analyses

Power analyses were conducted using the *pwr* (Champely, 2020) and *simr* (Green & MacLeod, 2016) packages in *R* (R Core Team, 2022). See https://osf.io/32m6h/?view_only=348484e6e86442e5a43e75e0cf9aa310 for code and output.

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.

First, we calculated power assuming that 15 ~~regions~~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. For aggregate analyses, we would have 120 observations (four projects x 15 ~~regions~~countries x two sample sources) and 90% power to detect moderate correlations (r = .290 with α = .05) between the mean predictions from our participants and the actual research results. For the multilevel analyses (see model specifications under "Analysis Plan"), we examined the power to detect very small relationships between predicted and actual results with 50 to 100 participants in each of our studies in steps of 10. The binary outcome model simulations revealed that 60 participants per study would provide more than 90% power to detect a very small effect of OR = ~~1.176~~1.18 with α = .05. The effect size model simulations revealed that 70 participants per study would provide more than 90% power to detect an very small effect of η^2 = .002 with α = .05.

For our tests examining what participant characteristics are related to prediction accuracy, we chose to focus on the potential effect of study involvement as it is most relevant to participant recruitment. The model simulations for both absolute effect size differences and Brier scores suggested that 70 participants per study would provide sufficient power (i.e., over 90% at α = .05) to detect very small effects of study involvement (η^2 = .003) with approximately equal group numbers.

We reran the simulations for our tests examining the relationships between predicted and actual results to see how power would be affected if each study included twice as many possible subsamples, or 30 ~~regions~~countries per study.[9] For aggregate analyses, this increase would provide 90% power for the detection of small to medium correlation effects ($r = .207$ with $\alpha = .05$). Power for the multilevel analyses should be less affected; the number of ~~regions~~countries represented in each study does not affect the total number of observations in the analyses (i.e., 20 per participant per study, or 8,000 total with 100 participants in each study). Still, for the models predicting actual results, the number of unique observations for each dependent measure will increase with the number of subsamples. Simulations revealed that this difference had minimal impact on power; 70 participants per study would still be sufficient to achieve at least 90% power to detect the same size effects as the previous analyses. Thus, our sampling goal of 100 participants per study, or 400 participants total, will provide enough power for all our primary multilevel analyses even with as much as 30% data loss due to incomplete responding.

Analysis Plan

All analyses will be conducted in *R* (R Core Team, 2022). ~~Formulas provided below adopt the *lme4* package notation (Bates et al., 2015).~~ Raw and clean data, and analysis code and output, will be shared publicly at https://osf.io/skx8d. The code for the planned analyses can be found at https://osf.io/32m6h/?view_only=348484e6e86442e5a43e75e0cf9aa310.

~~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. Missing data will not be replaced. Given our population of interest, PSA member researchers, we anticipate high participant engagement that produces good data quality.

[9] One of the largest PSA studies (Wang et al., 2021) included 37 countries/regions that had a minimum of 200 participants.

We will employ α = .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. The formulas provided below adopt the *lme4* package notation.

Focal Effects

For each of the four associated projects, we will briefly report the findings regarding the focal effect both overall (i.e., from a single test) and across subsamples (i.e., as determined by a meta-analysis).[10] After calculating an effect size for each subsample, we will report the weighted mean of the effect size based on a random-effects meta-analysis with sample ~~region~~country and sample ~~type~~source as random effects using the *metafor* package (Viechtbauer, 2010). We will report three heterogeneity estimates from each meta-analysis: Q, I^2, and τ^2. We will also report the percentage of ~~samples~~subsamples in which we found the expected effect.

Primary Analyses

Relationships between Predicted and Actual Results

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. First, we will examine the ~~correlation~~correlations between the aggregate subsample predictions of the researchers and the

[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.

actual subsample results using data from the four studies combined. 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. ¶

~~For~~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.

Second, for both binary outcomes and effect sizes, we will construct multilevel models with researcher predictions predicting actual results including random intercepts of study, sample country, and sample source (~~community~~university vs. ~~university), and sample region. Because the outcome measures~~community). 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 ~~of researcher~~ in ~~the~~these models ~~is not appropriate (i.e.,~~ would produce ~~a~~ singular model ~~fit). Instead~~fits. Thus, we will instead calculate prediction “intercepts” for each researcher individually to include in ~~the~~our models as fixed effects ~~by running an individual model for each researcher~~. Specifically, we will run separate models for each researcher (i.e., 400 total per outcome) with their predictions predicting outcomes and ~~extracting~~extract the model ~~intercept. Both linear (effect sizes) and logistic (binary outcomes~~intercepts. These values will then be included in the models to account for baseline differences in researcher predictions. Both logistic (binary outcomes) and linear (effect sizes) models have the following specification: result ~ prediction +

researcher_intercepts + (1|study) + (1|~~source~~country) + (1|~~region~~source). Calculated individual researcher prediction slopes (i.e., their model coefficients) and random slopes of prediction for study, sample ~~source~~country, and sample ~~region~~source will be tested to see if they contribute to the model ~~and retained when they improve model fit~~. 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.

If we observe relationships between predicted results and actual results in aggregate (i.e., the subsample level 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. The substance and degree of their prediction accuracy will be inferred by the presence and magnitude of these relationships across analyses.

Accuracy Measures

~~Absolute differences and Brier scores~~Brier scores and absolute effect size differences will serve as dependent variables in multilevel linear models including random intercepts of researcher, study, sample ~~source~~country, and sample ~~region~~source. The models' specification will be: score ~ 1+ (1|researcher) + (1|study) + (1|~~source~~country) + (1|~~region~~source). The fixed intercepts in these models provide estimates of the overall accuracy across predictions of each type. Random intercepts for sample, source, and study will be examined to assess variations in accuracy. These models will serve as the base models for our analyses examining what researcher characteristics relate to prediction accuracy. ~~Potential correlates will be added as predictor variables and tested in separate models.~~ 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.

If we find an effect of a tested researcher characteristic on accuracy scores, we will conclude that prediction accuracy relates to that characteristic. The impact of a given researcher characteristic on accuracy will be inferred by the presence and magnitude of its effect across analyses. Additional researcher characteristics will be tested as predictors of accuracy in a series of exploratory analyses. These will include prediction difficulty ratings, researcher beliefs, and the other measures of research involvement, experience, and expertise.

Secondary Analyses

Overall Result Predictions

~~We will compare the responses on the overall prediction items to the~~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. We will report the mean predicted probabilities relative to the effect outcomes. 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.

Under- or Over-generalization

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.

Moderation Predictions

We will also examine moderation predictions to determine how well researchers can predict the probability of individual and ~~sample~~subsample level moderation effects. First, we will fit multilevel generalized linear models for the moderation predictions across the four studies. Models for the individual level predictions and ~~sample~~subsample level predictions will be fit separately; each will include moderation predicted probabilities predicting binary moderation outcomes with random intercepts of ~~researcher, moderator variable~~study, researcher, and ~~study~~moderator variable. The model specification will be: result ~ prediction + (1|study) + (1|researcher) + (1|moderator)~~ + (1|study)~~. Random slopes of prediction for study, researcher, and moderator variable~~, and study~~ will be tested to see if they contribute to the model and retained when they improve model fit. 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.

Brier scores will also be computed for the moderation predictions as a measure of accuracy. In exploratory analyses, researcher characteristics will be tested as predictors of these scores in multilevel linear models with the following specification: score ~ characteristic + (1|researcher) + (1|moderator) + (1|study).

Researcher Beliefs

~~Researcher~~We will examine researcher beliefs about the focal effect and research generalizability ~~will be examined~~ for each study separately. Relationships among researcher characteristics, mean Brier scores, mean absolute effect size differences~~, mean Brier scores~~, and the researcher beliefs items will be examined and reported in correlation tables. Mixed ANOVAs will compare differences between the three generalizability belief items according to category (i.e., psychology overall, study subfield, and focal effect). We will use post hoc pairwise comparisons with Satherwaite adjusted degrees of freedom to examine simple effects.

Author Contributions

KS & CRC contributed to conceptualization.

CRC contributed to funding acquisition.

KS contributed to formal analysis and project administration.

All authors contributed to investigation, methodology, and writing.

References

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

Anderson, C. A., Allen, J. J., Plante, C., Quigley-McBride, A., Lovett, A., & Rokkum, J. N. (2019). The MTurkification of ~~Social and Personality Psychology~~social and personality psychology. *Personality and Social Psychology Bulletin*, *45*(6), 842–850. https://doi.org/10.1177/0146167218798821

Apicella, C., Norenzayan, A., & Henrich, J. (2020). Beyond WEIRD: A review of the last decade and a look ahead to the global laboratory of the future. *Evolution and Human Behavior*, *41*(5), 319–329. https://doi.org/10.1016/j.evolhumbehav.2020.07.015

Apicella, C. L., & Barrett, H. C. (2016). Cross-cultural evolutionary psychology. *Current Opinion in Psychology*, *7*, 92–97. https://doi.org/10.1016/j.copsyc.2015.08.015

Arnett, J. J. (2008). The neglected 95%: Why American psychology needs to become less American. *American Psychologist*, *63*(7), 602–614. https://doi.org/10.1037/0003-066X.63.7.602

Bago, B., Kovacs, M., Protzko, J., Nagy, T., Kekecs, Z., Palfi, B., Adamkovic, M., Adamus, S., Albalooshi, S., Albayrak-Aydemir, N., Alfian, I. N., Alper, S., Alvarez-Solas, S., Alves, S. G., Amaya, S., Andresen, P. K., Anjum, G., Ansari, D., Arriaga, P., … Aczel, B. (2022). Situational factors shape moral judgements in the trolley dilemma in Eastern, Southern and Western countries in a culturally diverse sample. *Nature Human Behaviour*, *6*, 880–895. https://doi.org/10.1038/s41562-022-01319-5

~~Baron, J. (1993). Why Teach Thinking?-An Essay. *Applied Psychology*, *42*(3), 191–214. https://doi.org/10.1111/j.1464-0597.1993.tb00731.x~~Baribault, B., Donkin, C., Little, D. R., Trueblood, J. S., Oravecz, Z., van Ravenzwaaij, D., White, C. N., De Boeck, P., & Vandekerckhove, J. (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

Baron, J. (1993). Why Teach Thinking?-An Essay. *Applied Psychology*, *42*(3), 191–214. https://doi.org/10.1111/j.1464-0597.1993.tb00731.x

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

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting ~~Linear Mixed-Effects Models Using~~linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beebe, J. R., & Matheson, J. (2022). Measuring ~~Virtuous Responses to Peer Disagreement: The Intellectual Humility and Actively Open-Minded Thinking of Conciliationists~~virtuous responses to peer disagreement: the intellectual humility and actively open-minded thinking of conciliationists. *Journal of the American Philosophical Association*, 1–24. https://doi.org/10.1017/apa.2022.8

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

Benjamin, D., Mandel, D. R., & Kimmelman, J. (2017). Can cancer researchers accurately judge whether preclinical reports will reproduce? *PLOS Biology*, *15*(6), e2002212. https://doi.org/10.1371/journal.pbio.2002212

Boer, D., & Fischer, R. (2013). How and when do personal values guide our attitudes and sociality? Explaining cross-cultural variability in attitude–value linkages. *Psychological Bulletin*, *139*(5), 1113–1147. https://doi.org/10.1037/a0031347

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

Brier, G. W. (1950). Verification of ~~Forecasts Expressed in Terms of Probability~~forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*(3), 306–307. https://doi.org/10.1207/s15327752jpa4803_13

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Almenberg, J., Altmejd, A., Chan, T., Heikensten, E., Holzmeister, F., Imai, T., Isaksson, S., Nave, G., Pfeiffer, T., Razen, M., & Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Camerer, C. F., & Johnson, E. J. (1991). The process-performance paradox in expert judgment: How can experts know so much and predict so badly? In K. A. Ericsson & J. Smith, *Toward a general theory of expertise: Prospects and limits* (pp. 195–217). Cambridge University Press.

Cesario, J. (2014). Priming, ~~Replication~~replication, and the ~~Hardest Science~~hardest science. *Perspectives on Psychological Science*, *9*(1), 40–48. https://doi.org/10.1177/1745691613513470

Champely, S. (2020). *Pwr: Basic Functions for Power Analysis. R package version 1.3-0*. https://CRAN.R-project.org/package=pwr

Crisp, R. J., & Birtel, M. D. (2014). Reducing ~~Prejudice Through Mental Imagery: Notes on Replication, Interpretation~~prejudice through mental imagery: notes on replication, interpretation, and ~~Generalization~~generalization. *Psychological Science*, *25*(3), 840–841. https://doi.org/10.1177/0956797613520169

Deffner, D., Rohrer, J. M., & McElreath, R. (2022). A ~~Causal Framework for Cross-Cultural Generalizability~~causal framework for cross-cultural generalizability. *Advances in Methods and Practices in Psychological Science*, *5*(3). https://doi.org/10.1177/2515245922110

DeJesus, J. M., Callanan, M. A., Solis, G., & Gelman, S. A. (2019). Generic language in scientific communication. *Proceedings of the National Academy of Sciences*, *116*(37), 18370–18377. https://doi.org/10.1073/pnas.1817706116

Delios, A., Clemente, E. G., Wu, T., Tan, H., Wang, Y., Gordon, M., Viganola, D., Chen, Z., Dreber, A., Johannesson, M., Pfeiffer, T., Generalizability Tests Forecasting Collaboration, & Uhlmann, E. L. (2022). Examining the generalizability of research findings from archival data. *Proceedings of the National Academy of Sciences*, *119*(30), e2120377119. https://doi.org/10.1073/pnas.2120377119

DellaVigna, S., & Pope, D. (2018). Predicting ~~Experimental Results: Who Knows What~~experimental results: who knows what? *Journal of Political Economy*, *126*(6). https://doi.org/10.1086/699976

Dijksterhuis, A. (2018). Reflection on the ~~Professor-Priming Replication Report~~professor-priming replication report. *Perspectives on Psychological Science*, *13*(2), 295–296. https://doi.org/10.1177/1745691618755705

Dorison, C. A., Lerner, J. S., Heller, B. H., Rothman, A. J., Kawachi, I. I., Wang, K., Rees, V. W., Gill, B. P., Gibbs, N., Ebersole, C. R., Vally, Z., Tajchman, Z., Zsido, A. N., Zrimsek, M., Chen, Z., Ziano, I., Gialitaki, Z., Ceary, C. D., Lin, Y., … Coles, N. A. (2022). In COVID-19 ~~Health Messaging, Loss Framing Increases Anxiety with Little-to-No Concomitant Benefits~~health messaging, loss framing increases anxiety with little-to-no concomitant benefits: Experimental ~~Evidence~~evidence from 84 ~~Countries~~countries. *Affective Science*, *3*(3), 577–602. https://doi.org/10.1007/s42761-022-00128-3

Dreber, A., Pfeiffer, T., Almenberg, J., Isaksson, S., Wilson, B., Chen, Y., Nosek, B. A., & Johannesson, M. (2015). Using prediction markets to estimate the reproducibility of scientific research. *Proceedings of the National Academy of Sciences*, *112*(50), 15343–15347. https://doi.org/10.1073/pnas.1516179112

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … Levitan, C. A. (2020). Many ~~Labs~~labs 5: Testing ~~Pre-Data-Collection Peer Review as an Intervention to Increase Replicability~~pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Ferguson, M. J., Carter, T. J., & Hassin, R. R. (2014). Commentary on the ~~Attempt to Replicate the Effect of the American Flag on Increased Republican Attitudes~~attempt to replicate the effect of the american flag on increased republican attitudes. *Social Psychology*, *45*(4), 299–311.

Forsell, E., Viganola, D., Pfeiffer, T., Almenberg, J., Wilson, B., Chen, Y., Nosek, B. A., Johannesson, M., & Dreber, A. (2019). Predicting replication outcomes in the Many Labs 2 study. *Journal of Economic Psychology*, *75*(A), 102117. https://doi.org/10.1016/j.joep.2018.10.009

Fraser, H., Bush, M., Wintle, B. C., Mody, F., Smith, E. T., Hanea, A. M., Gould, E., Hemming, V., Hamilton, D. G., Rumpff, L., Wilkinson, D. P., Pearson, R., Singleton Thorn, F., Ashton, R., Willcox, A., Gray, C. T., Head, A., Ross, M., Groenewegen, R., … Fidler, F. (2023). Predicting reliability through structured expert elicitation with the repliCATS (Collaborative Assessments for Trustworthy Science) process. *PLOS ONE*, *18*(1), e0274429. https://doi.org/10.1371/journal.pone.0274429

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad7243

Gordon, M., Viganola, D., Dreber, A., Johannesson, M., & Pfeiffer, T. (2021). Predicting replicability—Analysis of survey and prediction market data from large-scale forecasting projects. *PLOS ONE*, *16*(4), e0248780. https://doi.org/10.1371/journal.pone.0248780

Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, *7*(4), 493–498. https://doi.org/10.1111/2041-210X.12504

Haeffel, G. J., & Cobb, W. R. (2022). Tests of generalizability can diversify psychology and improve theories. *Nature Reviews Psychology*, *1*(4), 186–187. https://doi.org/10.1038/s44159-022-00039-x

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making*, *8*(3), 188–201. https://doi.org/10.1017/S1930297500005921

Henrich, J. (2015). Culture and social behavior. *Current Opinion in Behavioral Sciences*, *3*, 84–89. https://doi.org/10.1016/j.cobeha.2015.02.001

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2–3), 61–83. https://doi.org/10.1017/S0140525X0999152X

Hoekstra, R., & Vazire, S. (2021). Aspiring to greater intellectual humility in science. *Nature Human Behaviour*, *5*(12), 1602–1607. https://doi.org/10.1038/s41562-021-01203-8

Hoogeveen, S., Sarafoglou, A., & Wagenmakers, E.-J. (2020). Laypeople ~~Can Predict Which Social-Science Studies Will Be Replicated Successfully~~can predict which social-science studies will be replicated successfully. *Advances in Methods and Practices in Psychological Science*, *3*(3), 267–285. https://doi.org/10.1177/2515245920919667

Hruschka, D. J., Medin, D. L., Rogoff, B., & Henrich, J. (2018). Pressing questions in the study of psychological and behavioral diversity. *Proceedings of the National Academy of Sciences*, *115*(45), 11366–11368. https://doi.org/10.1073/pnas.1814733115

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., Ndukaihe, I. L. G., Bloxsom, N. G., Lewis, S. C., Foroni, F., Willis, M. L., Cubillas, C. P., Vadillo, M. A., Turiegano, E., Gilead, M., Simchon, A., Saribay, S. A., Owsley, N. C., Jang, C., … Coles, N. A. (2021). To which world regions does the valence–dominance model of social perception apply? *Nature Human Behaviour*, *5*, 159–169. https://doi.org/10.1038/s41562-020-01007-2

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Hilgard, J., Ahn, P. H., Brady, A. J., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J. T., Cromar, R., Gardiner, G., Gosnell, C. L., Grahe, J., Hall, C., Howard, I., … Ratliff, K. A. (2022). Many ~~Labs~~labs 4: Failure to ~~Replicate Mortality Salience Effect With and Without Original Author Involvement~~replicate mortality salience effect with and without original author involvement. *Collabra: Psychology, 8*(1), 35271. https://doi.org/10.1525/collabra.35271

Klein, R. A., Ratliff, K. A., Vianello, M., Adams ~~Jr.~~, R. B., Jr., Bahník, ~~S~~Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., ~~…~~. . . Nosek, B. A. (2014). Investigating ~~Variation in Replicability: A ““Many Labs”” Replication Project. *European Psychologist*, *14*~~variation in replicability: A “many labs” replication project. *Social Psychology, 45*(3), ~~260–262~~142–152. https://doi.org/10.1027/~~a000001~~1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many ~~Labs~~labs 2: ~~Investigating Variation in Replicability Across Samples and Settings~~investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Kline, M. A., Shamsudheen, R., & Broesch, T. (2018). Variation is the universal: Making cultural evolution work in developmental psychology. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *373*(1743), 20170059. https://doi.org/10.1098/rstb.2017.0059

Krizan, Z., & Windschitl, P. D. (2007). The influence of outcome desirability on optimism. *Psychological Bulletin*, *133*(1), 95–121. https://doi.org/10.1037/0033-2909.133.1.95

Krumrei-Mancuso, E. J., Haggard, M. C., LaBouff, J. P., & Rowatt, W. C. (2020). Links between intellectual humility and acquiring knowledge. *The Journal of Positive Psychology*, *15*(2), 155–170. https://doi.org/10.1080/17439760.2019.1579359

Krumrei-Mancuso, E. J., & Rouse, S. V. (2016). The ~~Development and Validation of the Comprehensive Intellectual Humility Scale~~development and validation of the comprehensive intellectual humility scale. *Journal of Personality Assessment*, *98*(2), 209–221. https://doi.org/10.1080/00223891.2015.1068174

Kukull, W. A., & Ganguli, M. (2012). Generalizability: The trees, the forest, and the low-hanging fruit. *Neurology*, *78*(23), 1886–1891. https://doi.org/10.1212/WNL.0b013e318258f812

Kunda, Z. (1990). The ~~Case for Motivated Reasoning~~case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. https://doi.org/10.1037/0033-2909.108.3.480

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

Leary, M. R., Diebels, K. J., Davisson, E. K., Jongman-Sereno, K. P., Isherwood, J. C., Raimi, K. T., Deffler, S. A., & Hoyle, R. H. (2017). Cognitive and ~~Interpersonal Features of Intellectual Humility~~interpersonal features of intellectual humility. *Personality and Social Psychology Bulletin*, *43*(6), 793–813. https://doi.org/10.1177/0146167217697695

~~Mandel, D. R., & Barnes, A. (2014). Accuracy of forecasts in strategic intelligence. *Proceedings of the National Academy of Sciences*, *111*(30), 10984–10989. https://doi.org/10.1073/pnas.1406138111~~¶

McBride, M. F., Fidler, F., & Burgman, M. A. (2012). Evaluating the accuracy and calibration of expert predictions under uncertainty: Predicting the outcomes of ecological research: Evaluating expert predictions under uncertainty. *Diversity and Distributions*, *18*(8), 782–794. https://doi.org/10.1111/j.1472-4642.2012.00884.x

McDiarmid, A. D., Tullett, A. M., Whitt, C. M., Vazire, S., Smaldino, P. E., & Stephens, J. E. (2021). Psychologists update their beliefs about effect sizes after replication studies. *Nature Human Behaviour*, *5*(12), 1663–1673. https://doi.org/10.1038/s41562-021-01220-7

Medin, D., Ojalehto, B., Marin, A., & Bang, M. (2017). Systems of (non-)diversity. *Nature Human Behaviour*, *1*(5), 0088. https://doi.org/10.1038/s41562-017-0088

Mellers, B., Stone, E., Atanasov, P., Rohrbaugh, N., Metz, S. E., Ungar, L., Bishop, M. M., Horowitz, M., Merkle, E., & Tetlock, P. (2015). The psychology of intelligence analysis: Drivers of prediction accuracy in world politics. *Journal of Experimental Psychology: Applied, 21*(1), 1–14. https://doi.org/10.1037/xap0000040

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., Chen, E., Baker, J., Hou, Y., Horowitz, M., Ungar, L., & Tetlock, P. (2015). Identifying and ~~Cultivating Superforecasters~~cultivating superforecasters as a ~~Method of Improving Probabilistic Predictions~~method of improving probabilistic predictions. *Perspectives on Psychological Science*, *10*(3), 267–281. https://doi.org/10.1177/1745691615577794

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., … Chartier, C. R. (2018). The ~~Psychological Science Accelerator: Advancing Psychology~~

~~Through a Distributed Collaborative Network~~psychological science accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515. https://doi.org/10.1177/25152459187976

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1). https://doi.org/10.1038/s41562-016-0021

Muthukrishna, M., Bell, A. V., Henrich, J., Curtin, C. M., Gedranovich, A., McInerney, J., & Thue, B. (2020). Beyond Western, Educated, Industrial, Rich, and Democratic (WEIRD) ~~Psychology~~psychology: Measuring and ~~Mapping Scales of Cultural and Psychological Distance~~mapping scales of cultural and psychological distance. *Psychological Science*, *31*(6), 678–701. https://doi.org/10.1177/0956797620916782

Nielsen, M., Haun, D., Kärtner, J., & Legare, C. H. (2017). The persistent sampling bias in developmental psychology: A call to action. *Journal of Experimental Child Psychology*, *162*, 31–38. https://doi.org/10.1016/j.jecp.2017.04.017

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration ~~Is Hard, And Worthwhile~~is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology*, *18*(3), e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology, 73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Olsson-Collentine, A., Wicherts, J. M., & van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*(10), 922–940. https://doi.org/10.1037/bul0000294

Peters, U., Krauss, A., & Braganza, O. (2022). Generalization ~~Bias in Science~~bias in science. *Cognitive Science*, *46*(9). https://doi.org/10.1111/cogs.13188

Plott, C. R., & Sunder, S. (1988). Rational ~~Expectations~~expectations and the ~~Aggregation of Diverse Information in Laboratory Security Markets~~aggregation of diverse information in laboratory security markets. *Econometrica*, *56*(5), 1085. https://doi.org/10.2307/1911360

Porter, T., Elnakouri, A., Meyers, E. A., Shibayama, T., Jayawickreme, E., & Grossmann, I. (2022). Predictors and consequences of intellectual humility. *Nature Reviews Psychology*, *1*(9), 524–536. https://doi.org/10.1038/s44159-022-00081-9

Psychological Science Accelerator Self-Determination Theory Collaboration. (2022). A global experiment on motivating social distancing during the COVID-19 pandemic. *Proceedings of the National Academy of Sciences*, *119*(22), e2111091119. https://doi.org/10.1073/pnas.2111091119

R Core Team. (2022). *R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.* https://www.R-project.org/

Rad, M. S., Martingano, A. J., & Ginges, J. (2018). Toward a psychology of *Homo sapiens*: Making psychological science more representative of the human population. *Proceedings of the National Academy of Sciences*, *115*(45), 11401–11405. https://doi.org/10.1073/pnas.1721165115

Schnall, S., Johnson, D. J., Cheung, F., & Brent Donnellan, M. (2014). Commentary and ~~Rejoinder~~rejoinder on Johnson, Cheung, and Donnellan (2014a). *Social Psychology*, *45*(4), 315–320. https://doi.org/10.1027/1864-9335/a000204

Schwarz, N., & Strack, F. (2014). Does ~~Merely Going Through the Same Moves Make~~merely going through the same moves make for a ““~~“Direct””” Replication~~direct””” replication? *Social Psychology*, *45*(4), 299–311.

Sears, D. O. (1986). College ~~Sophomores~~sophomores in the ~~Laboratory~~laboratory: Influences of a ~~Narrow Data Base on Social Psychology’s View of Human Nature~~narrow data base on social psychology’s view of human nature. *Journal of Personality and Social Psychology*, *51*(3), 515–530. https://doi.org/10.1037/0022-3514.51.3.515

Shih, M., Pittinsky, T. L., Moon, A., & Roeder, S. S. (2014). Commentary and ~~Rejoinder~~rejoinder on Gibson, Losee, and Vitiello (2014) and Moon and Roeder (2014). *Social Psychology*, *45*(4), 335–338. https://doi.org/10.1027/1864-9335/a000207

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012, October 14). *A 21 Word Solution*. http://dx.doi.org/10.2139/ssrn.2160588

Simons, D. J. (2014). The ~~Value of Direct Replication~~value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Simons, D. J., Shoda, Y., & Lindsay, D. S. (2017). Constraints on ~~Generality~~generality (COG): A ~~Proposed Addition to All Empirical Papers~~proposed addition to all empirical papers.

Perspectives on Psychological Science, *12*(6), 1123–1128. https://doi.org/10.1177/174569161770863

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning*, *13*(3), 225–247. https://doi.org/10.1080/13546780600780796

Syed, M., & Kathawalla, U.-K. (2022). Cultural psychology, diversity, and representation in open science. In K. C. McLean~~, *Cultural methods in psychology*~~ (Ed.), *Cultural Methods in Psychology: Describing and ~~transforming cultures~~Transforming Cultures* (pp. 427–454). Oxford University Press. https://doi.org/10.1093/oso/9780190095949.003.0015

Thalmayer, A. G., Toscanelli, C., & Arnett, J. J. (2021). The neglected 95% revisited: Is American psychology becoming less American? *American Psychologist*, *76*(1), 116–129. https://doi.org/10.1037/amp0000622

Tiokhin, L., Hackman, J., Munira, S., Jesmin, K., & Hruschka, D. (2019). Generalizability is not optional: Insights from a cross-cultural study of social discounting. *Royal Society Open Science*, *6*(2), 181386. https://doi.org/10.1098/rsos.181386

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454–6459. https://doi.org/10.1073/pnas.1521897113

Vazire, S. (2018). Implications of the ~~Credibility Revolution for Productivity, Creativity~~credibility revolution for productivity, creativity, and ~~Progress~~progress. *Perspectives on Psychological Science*, *13*(4), 411–417. https://doi.org/10.1177/~~174569161775188~~1745691617751884

Viechtbauer, W. (2010). Conducting ~~Meta-Analyses~~meta-analyses in *R* with the metafor ~~Package~~package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Viganola, D., Buckles, G., Chen, Y., Diego-Rosell, P., Johannesson, M., Nosek, B. A., Pfeiffer, T., Siegel, A., & Dreber, A. (2021). Using prediction markets to predict the outcomes in the Defense Advanced Research Projects Agency's next-generation social science programme. *Royal Society Open Science*, *8*(7), 181308. https://doi.org/10.1098/rsos.181308

Wang, K., Goldenberg, A., Dorison, C. A., Miller, J. K., Uusberg, A., Lerner, J. S., Gross, J. J., Agesin, B. B., Bernardo, M., Campos, O., Eudave, L., Grzech, K., Ozery, D. H., Jackson, E. A., Garcia, E. O. L., Drexler, S. M., Jurković, A. P., Rana, K., Wilson, J. P., … Moshontz, H. (2021). A multi-country test of brief reappraisal interventions on emotions during the COVID-19 pandemic. *Nature Human Behaviour*, *5*, 1089–1110. https://doi.org/10.1038/s41562-021-01173-x

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Zmigrod, L., Zmigrod, S., Rentfrow, P. J., & Robbins, T. W. (2019). The psychological roots of intellectual humility: The role of intelligence and cognitive flexibility. *Personality and Individual Differences*, *141*, 200–208. https://doi.org/10.1016/j.paid.2019.01.016

Table 1

Study Design Table

Question	**Hypothesis**	**Sampling Plan**	**Analysis Plan**	**Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis**	**Interpretation given different outcomes**
Can researchers accurately predict the generalizability of psychological phenomena?	We will observe a positive relationship between predicted and actual **outcomes** in aggregate.	A minimum of 120 observations (i.e., subsamples; four-projects x 15 ~~regions~~countries x two sample sources).	Correlation between the mean researcher predicted probabilities and the actual binary outcomes.	A power sensitivity analysis found that 120 observations would provide 90% power to detect moderate correlations (r = .290 with α = .05).	In aggregate, researchers [*are (at least somewhat)* * / *are not*] accurate in their predictions of when psychological effects will generalize across regional subsamples.
	We will observe a positive relationship between predicted and actual **outcomes** at the level of prediction.	100 participants per study (total n = 400) recruited via the Psychological Science Accelerator's newsletter, project updates, and social media.	Multilevel generalized linear model with researcher outcome probability predictions predicting actual binary outcome results including calculated researcher intercepts and random intercepts of study, sample source, and sample ~~region~~country.	A simulation-based power analysis found that 60 participants per study would provide more than 90% power to detect a very small effect of OR = 1.176 with α = .05.	Researchers [*are (at least somewhat)* * / *are not*] accurate in their predictions of when psychological effects will generalize across regional subsamples.

	We will observe a positive relationship between predicted and actual **effect sizes** in aggregate.	A minimum of 120 observations (i.e., subsamples; four projects x 15 ~~regions~~countries x two sample sources)	Correlation between the mean researcher predicted effect sizes and the actual effect sizes.	A power sensitivity analysis found that 120 observations would provide 90% power to detect moderate correlations (r = .290 with α = .05).	In aggregate, researchers [*are (at least somewhat)** / *are not*] accurate in their predictions of the sizes of psychological effects across regional subsamples.
	We will observe a positive relationship between predicted and actual **effect sizes** at the level of prediction.	100 participants per study (total n = 400) recruited via the Psychological Science Accelerator's newsletter, project updates, and social media.	Multilevel linear model with researcher effect size predictions predicting actual effect sizes including calculated researcher intercepts and random intercepts of study, sample source, and sample ~~region~~country.	A simulation-based power analysis found that 70 participants per study would provide more than 90% power to detect a very small effect of η^2 = .002 with α = .05.	Researchers [*are (at least somewhat)** / *are not*] accurate in their predictions of the sizes of psychological effects across regional subsamples.
What researcher characteristics predict the accuracy of their generalizability predictions?	Researcher characteristics† will predict **Brier scores** (i.e., the ~~mean~~ squared prediction error) for the outcome predictions.	100 participants per study (total n = 400) recruited via the Psychological Science Accelerator's newsletter, project updates, and social media.	Multilevel linear model with the researcher characteristic† predicting Brier scores with random intercepts of researcher, study, sample source, and sample ~~region~~country.	A simulation-based power analysis found that 70 participants per study would provide more than 90% power to detect a very small effect of η^2 = .003 with α = .05.	The researcher characteristic† is [*positively/negatively/ not*] related to the accuracy of outcome probability predictions across regional subsamples.

	Researcher characteristics[†] will predict **absolute effect size differences** between predictions and results.	100 participants per study (total *n* = 400) recruited via the Psychological Science Accelerator's newsletter, project updates, and social media.	Multilevel linear model with the researcher characteristic[†] predicting absolute effect size differences with random intercepts of researcher, study, sample source, and sample ~~region~~country.	A simulation-based power analysis found that 70 participants per study would provide more than 90% power to detect a very small effect of $\eta^2 = .003$ with $\alpha = .05$.	The researcher characteristic[†] is [*positively/negatively/ not*] related to the accuracy of effect size estimates across regional subsamples

Note. This table includes the primary research questions and their corresponding analyses. We excluded the theory column because outcomes won't be interpreted as evidence for or against a given theory.

* The degree of prediction accuracy will be inferred by the magnitude of the relationship.

[†] The researcher characteristics that will be tested in the primary analyses are as follows: prediction confidence, involvement in the project, highest degree, self-rated expertise in the project subfield, intellectual humility, and actively open-minded thinking. These measures will be tested individually in separate models. We will also include all six variables in the same models.