Eleven years of student replication projects provide evidence on the correlates of replicability in psychology

Veronica Boyce[1,*], Maya Mathur[1], Michael C. Frank[1]

[1]Stanford University

Abstract

Cumulative scientific progress requires empirical results that are robust enough to support theory construction and future extensions. Yet in psychology, some prominent findings have failed to replicate, and large-scale investigations suggest that failures to replicate are widespread. The identification of features that predict replication success is limited by data, and most analyses re-analyse the same set of ~170 replications. We introduce a new dataset of 176 replications from students in a graduate-level methods course. Replication results were close enough to original results to support student extensions in 49% of replications. 45% of replications with suitable numerical outcomes (N=136) had point estimates within the prediction interval of the original outcome. Larger original effect sizes and within-participants designs were predictive of replication. Consistent with prior reports, our results indicate that the robustness of the psychology literature is low enough to limit progress by student investigators.

[148 words / NHB max 150]

1 Introduction

Cumulative scientific progress requires empirical results that are robust enough to support theory construction and future extensions. Yet in psychology, some prominent individual findings have failed to replicate in multi-site replication attempts (ex. terror management theory, Klein et al. 2022, ego-depletion, Hagger et al. 2016), and a large-scale replication project pegged the replication rate for findings in top-tier psychology journals at around 40% (Open Science Consortium 2015). When scientists treat publication as a signal of truth and attempt to build on results without verifying them first, they can become frustrated pursuing extensions to a popular, but unreliable result. As each failed extension does not directly contradict the original result, many scientists may waste considerable time and resources pursuing results where there may be none.

To identify studies at risk for non-cumulative findings, we might want to know 1) what to measure in order to capture the waste and frustration due to non-cumulative failed extensions and 2) from a descriptive perspective, what features of studies are associated with being at risk for this non-replicability. Measuring replicability has been the focus of much prior work, but there is limited consensus on what exactly is being estimated.

1.1 Large-scale replication projects

A few large-scale investigations have measured replication rates in samples of psychology studies. Reproducibility Project: Psychology sampled roughly 100 studies from articles published in three top psychology journals in 2008 and distributed the studies across participating labs. They found an overall replication rate of around 40% (Open Science Consortium 2015). Many Labs investigated heterogeneity using short target studies that compared between two conditions each. These study designs were not representative of the psychology literature as a whole, and due to the heterogeneity goal, they had large overall samples across multiple sites. Across Many Labs 1-3, in these high-powered replications, only 29 of 51 target effects (57%) replicated (Klein et al. 2014, Ebersole et al. 2016, Klein et al. 2018). Camerer et al. (2018)

*Corresponding author. Email: vboyce@stanford.edu

included all 21 behavioral social science studies from Nature and Science from 2010-2015 that were feasible; they consulted original authors and had high power to detect effects smaller than those reported. In this well-resourced environment with expert input, the replication rate was around 60%.

While their sampling and methods varied, these previous approaches to replicability have focused on interpreting their results in terms of a potentially problematic estimand: the probability of a finding in the literature being somehow truly replicable. Critics have pointed out that "true" replicability may not be possible to estimate outside of a specific sample (Van Bavel et al. 2016) or even time period (Ramscar et al. n.d.). Further, the methods for estimating this quantity have been theoretically problematic. Sampling schemes did not reflect an entirely random sample from the literature; instead replication projects sampled from specific journals where results may be of more interest and adjusted the sample for feasibility concerns. These are reasonable sampling choices, but they undermine the claim that the estimand is the level of "truth" in the literature as a whole. Sampling truly at random from the literature may not even be desirable, as arguably a literature will succeed if useful discoveries come out of it, not if random findings are true (Wilson et al. 2020).

Rather than aim for some measure of "true" replicability, it makes more sense to contextualize replication efforts based on their methodologies and outcome measures. With this lens, we could say that RP:P looked at how well findings in well-known journals can be replicated in a typical psychology lab. Many Labs looked at which well-known two condition studies replicate with huge samples, and Camerer et al. (2018) looked at which prestigious journal findings replicate when conducted in a highly-resourced environment with expert involvement. All of these are potentially desirable estimands depending on what the specific circumstances and questions of interest are.

In practice, most scientific work is conducted by graduate students with limited time, limited budgets and limited access to experts. Therefore, an important question is how replicable findings in the literature are for graduate students operating under these less-than-ideal conditions. If students cannot replicate a finding under these circumstances, it is unlikely that they will be able to successfully build upon the study in these circumstances either.

Our goal in estimating replicability is to look for markers of when findings can (and cannot) support cumulative science, by looking for markers of findings that do (or do not) support replication by graduate students operating under typical real-word constraints. Our sample of studies is selected based on what studies students were interested in and wanting to replicate, with some filtering for feasibility; this sampling reflects how scientists choose what studies to build on: those that are interesting and relatively doable given methodological and budgetary constraints. Our primary estimand is whether a student researcher, on selecting a finding of interest from the literature, can successfully achieve a result close enough to the original that they could build on it in their own work. In addition to this subjective measure of replicability, we also compare the effect sizes of the original and replication on one key outcome per study using prediction interval and p-original (Mathur & VanderWeele 2020).

1.2 Predicting replication success

Despite variation in the methods and outcomes used by large-scale replication studies, they are often aggregated together in analyses looking at the predictablity of replication success. Prediction markets and elicitations have established that people can predict what studies will replicate above chance (Dreber et al. 2015, Camerer et al. 2018, Forsell et al. 2019, Hoogeveen et al. 2019), but have not identified concrete predictors that differentiate replications from non-replications. Machine learning approaches trained on the available replications are also above chance at predicting replication success (Yang et al. 2020, Youyou et al. 2023). In terms of specific attributes that predict replicability, Altmejd et al. (2019) examined statistical and demographic features of studies and identified larger sample sizes, larger effect sizes, and simple effects (as opposed to interaction terms) as predictive of replication, and Open Science Consortium (2015) looked at correlates of replicability in the RP:P sample and found that studies in cognitive psychology (as opposed to social psychology) and studies with larger effect sizes and smaller p-values were more likely to replicate. One potential set of correlates to replicability that has not been thoroughly examined are experimental design features, such as between or within subjects designs and repeated measures.

While most approaches to correlates of replicability have been correlational, experimental approaches can be used to test potential interventions. Protzko et al. (2020) showed, across 16 studies, that better methodological practices led to replication rates that matched theoretical expectations with replication

effect sizes comparable with the original. Many Labs 5 added expert (original author) advice to a replication process, and found that it did not substantially increase the replication rate for the 10 studies (Ebersole et al. 2020). These types of experiments are valuable for testing potential causes of non-replication, but they don't scale well due to expense, and not all influences on non-replication may be experimentally manipulable.

While they don't incur the expense of collecting new datasets, correlational approaches do depend on previously collected replication data, and their conclusions are limited by the available data. Large-scale replications are arduous and expensive to run, so only a few large-scale replication datasets exist, and most analyses analyses draw heavily on same small set of data points. In particular, the RP:P dataset is much discussed and reanalyzed (Anderson et al. 2016, Etz & Vandekerckhove 2016, Gilbert et al. 2016, Patil et al. 2016) to the point that much of what we think we know about replicability may be overfit to the 100 studies included in RP:P. More replication data is needed to identify markers of replicability and targets for potential future experimental approaches.

Our contribution is a new dataset of 176 replications of experimental studies from the social sciences, primarily psychology. These replications were conducted by students in a graduate-level experimental methods class between 2011 and 2022 as individual course projects. We investigate statistical and experimental-design predictors of replicability in this dataset and find that within-subjects designs and studies with large standardized effect sizes are positively correlated with replication success.

2 Results

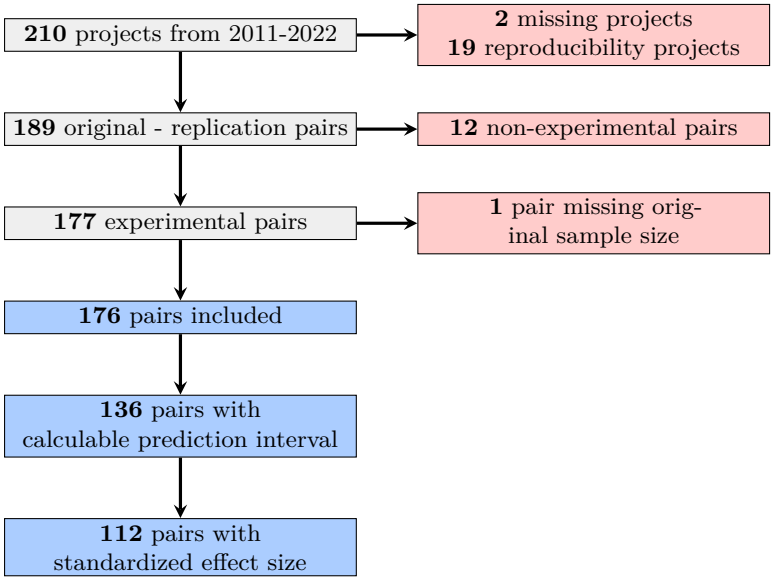

Figure 1: Of the 210 projects conducted for the class, 176 are included in our analysis, after excluding reproducibilty projects (with no new data collection), non-experimental replications, and missing projects. Of the 176, 136 report sufficient information to calculate prediction intervals, and 112 reported enough to calculated standardized effect sizes.

PSYCH 251 is Stanford Psychology's graduate-level experimental methods class taught by MCF. During the 10 week class, each student replicated a published finding. They individually re-implemented the study, wrote analysis code, pre-registered their study, collected data using an online platform, and wrote up a structured replication report. Students were free to choose studies related to their research interests, with the default recommendation being an article from a recent year of Psychological Science. The resultant sample of studies is not a random sample from the literature but is representative of studies that are of interest to and doable by first year graduate students.

The sample of replicated studies reflects the variability of the literature, including studies from different subfields, with different experimental methods and statistical outcomes. We leveraged the naturally occurring variability in this sample of replications to examine how different demographic, experimental design, and statistical properties predict replication success.

Of note, these replications were all conducted on short time scales, with constrained budgets, that in some cases limited the number of participants who could be recruited below what the original study had or what power analyses suggested. In nearly all cases, replications were conducted online, with recruitment from Amazon Mechanical Turk (default 2011-2020) or Prolific (default 2021-2022).

Many different measures can be used to define replication success of an individual statistical result (Simonsohn 2015, Gelman 2018/ed, Mathur & VanderWeele 2020). However, whether one feels confident in the results of a study given a replication is not always dependent on only one outcome measure (ex. interaction term) and particularly not dependent on only one statistical comparison between the two studies (ex. replication is p<.05 same direction as original). As our primary outcome, we use a subjective replication score which, unlike statistical measures, accommodated studies with multiple important outcome measures that together defined the pattern of interest and was applicable across the diverse range of statistical measures and reporting practices present in the sample. Importantly, a holistic measure of replication success had already been coded for most projects when they were turned in at the end of the class. For reliability, VB independently code the replication success from the students' written reports; discrepancies were resolved by discussion between MCF and VB (inter-rater reliability, Spearman's ρ: 0.88, disagreement in 26% of cases).

As a complement to our primary subjective outcome, we also used two statistical measures of replication on the subset of the data where they were computable for the key statistic of interest (136 cases, see Figure 1). We used p-original, the p-value on the null hypothesis that the original and replication statistics are from the same distribution, and prediction interval, a binary measure of whether the replication statistic fell within the prediction interval of the original statistic (Mathur & VanderWeele 2020). The prediction interval depends on the level of evidence of the original study; if the effect was marginal, the prediction interval could overlap zero; thus, a replication might fall within the predictive interval, and be consistent with the original outcome, but not provide compelling evidence for the claimed effect. Conversely, large original effects with precise point estimates may have prediction intervals that do not overlap a smaller replication effect size, and thus would be inconsistent with the original outcome, even though researcher intuition might classify it as a success. Thus, the statistical metrics quantify the similarity of a key statistic, but they will not always match researcher intuitions on whether a study replicated.

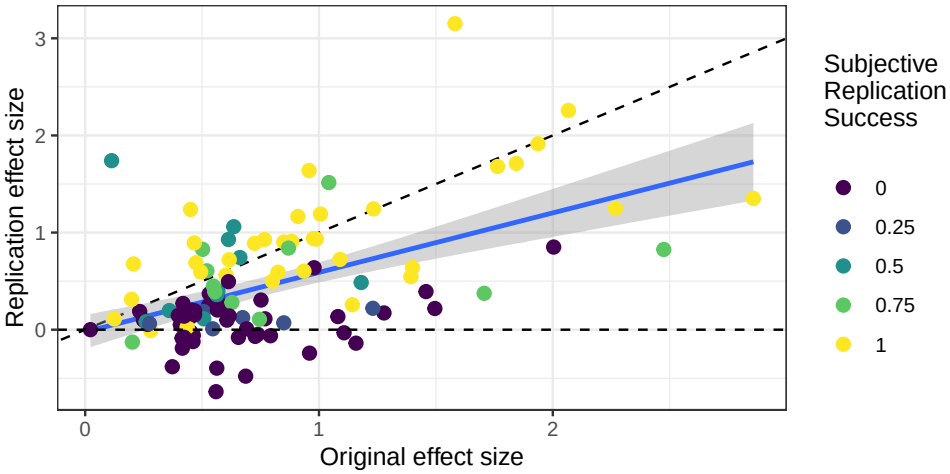

Figure 2: Relationship between effect size of the original study, effect size of the replication study, and subjective replication success rating, for those studies where effect size was applicable.

2.1 Overall replication rate

Across the 176 studies, the average subjective replication score was 49%, which we can interpret as an overall subjective replication rate. 46% (62/136) of replications has outcomes with point-estimates within the prediction interval of the original outcome. The median p_original value on the original and replication point-estimates coming from the same distribution was 0.03, representing the median probability that a replication study's estimate would be at least as extreme as was actually observed, if in fact the replication and original were statistically consistent. Figure 2 shows the relationship between original standardized effect size, replication effect size, and subjective replication score. Some studies replicated withed similar size effects to the original, and others failed to replicate, with replication effect sizes near zero. On average,

there was a diminution of effect sizes from original to replication. This pattern of results is consistent with the pattern of results found in RP:P (Open Science Consortium 2015).

Some multi-site replication projects have found heterogeneity in effect sizes across replication sites (Klein et al. 2018, Ebersole et al. 2020, Olsson-Collentine et al. 2020). If we assume that the level of heterogeneity in hypothetical multi-site replications of these effects would be the same as the average level heterogeneity found in Olsson-Collentine et al. (2020) (tau=.21 in SMD units), then 64.3% (72/112) of replication effect sizes are distributionally consistent with the original effect size. More work on understanding heterogeneity is needed to understand what levels of heterogeneity to expect across different implementations of the same experiment, and how considerations of heterogeneity should impact interpretations of both novel results and replication results.

[Is this the right place to talk about heterogeneity / what did we want to say?]

Table 1: The correlation between each individual predictor and the subjective replication score. For subfield, cognitive psychology is treated as the baseline condition. See Methods for how these variables were coded.

Predictors	r	p
Within subjects	0.333	0.000
Log n trials	0.182	0.015
Open data	0.150	0.047
Non psych	0.080	0.294
Other psych	0.075	0.322
Publication year	0.064	0.399
Open materials	0.002	0.979
Stanford	-0.027	0.725
Log rep/orig sample	-0.047	0.536
Log original sample size	-0.108	0.155
Switch to online	-0.158	0.037
Social	-0.246	0.001
Single vignette	-0.267	0.000

2.2 Single predictors

We investigated what features of the original study and replication were correlated with replication success, with the goal of being able to identify potential markers of replicability. We chose a set of predictor variables based on the correlational results of RP:P (Open Science Consortium 2015), our own intuitions of experimental design factors that might impact replication success, and some covariates related to how close the replication was. A full description of these features is given in Methods.

Many predictors individually correlated with subjective replication success (Table 1). Predictors of higher replicability included within-subjects designs, larger numbers of trials per subject, and the original study having open data. Predictors of lower replicability included single vignetted studies with only one induction or example per condition, social psychology studies, and original-replication pairs where the original study was in-person and the replication switched to online.

Distributions of study outcomes across some of these properties are shown in Figure 3. To quantify the effects of these single predictors, we ran logistic models predicting the subjective replication score (on 0-1) as a function of the single predictor, and exponentiated the coefficient to get an odds ratio. Both social and cognitive psychology studies were well represented in the replication sample, and the cognitive psychology studies replicated 2.45 times as often as social psychology studies. Within and between subjects designs were both common, and within-subjects designs replicated 3.35 times as much. Studies with multiple vignettes replicated 2.62 times as often as single vignetted studies. However, there were strong correlations among these experimental features and between these experimental features and subfield.

Studies with open data, which almost always also had open materials, tended to replicate more than studies without open data. Nearly all replications studies were conducted online, but original studies were split between using in-person and online recruitment. Replications that switched to online were less likely to replicate than those that had the same modality as the original (generally both online, in a few

cases both in-person). While online studies in general show comparable results to studies conducted in person (Crump et al. 2013), switching the modality does decrease the closeness of the replication, and some studies done in person may not have been well-adapted (ex. inductions may be weaker or attention checks inadequate to the new sample). These factors of open materials, open data, and online samples in original studies are more common in more recent studies, and so these effects may partially reflect temporal trends.

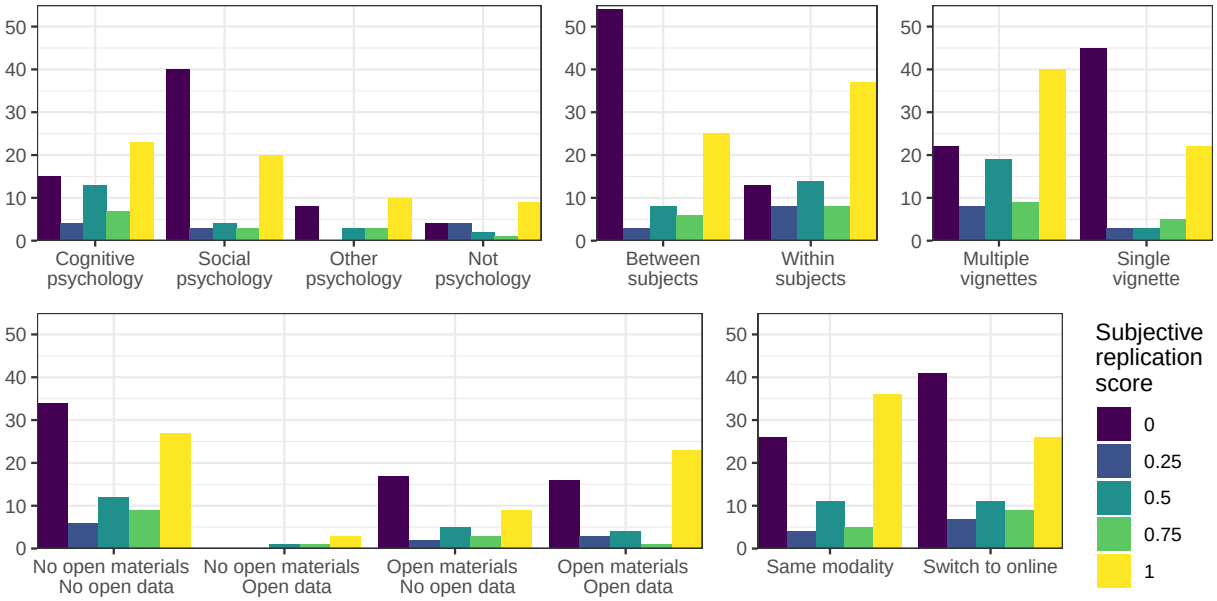

Figure 3: Distribution of subjective replication scores within categories. Bar heights are counts of studies.

2.3 Regression model

While a number of predictors show individual correlations with the subjective replication score, many of the predictors intercorrelate with one another. In order to determine which predictors were the strongest, we ran a series of pre-registered regression models regularized using a horseshoe shrinkage prior (Carvalho et al. 2009). We ran both 1) models using all the original-replication pairs, but without original effect size and original p value as predictors, as they were uncodable for some pairs, and 2) models including all predictors, but on only the subset of data where all predictors were available. The coefficient estimates from two models predicting the subjective replication scores are shown in Figure 4; see Supplement for full model results. Due to a large number of predictors coupled with a small and noisy dataset, there is much uncertainty around the coefficients even with strong regularization. The general directions of coefficients are consistent with the effects of the predictors in isolation.

Within-subjects designs stand out as the strongest indicator of replicability in the model with all the data (0.53, CrI= [0, 2]). In the model with all predictors, but less data, within-subjects designs remain predictive (0.66, CrI= [-0.04, 2.37]), and standardized effect size is also a strong predictor of subjective replication score (0.65, CrI= [0.27, 2.81]). Both effects are robust to a sensitivity analysis including only studies with close replications and matching statistical tests (within-subjects 0.87, CrI= [-0.01, 3.33]; effect size 1.05, CrI= [0.76, 4.88]).

We also ran models predicting our secondary outcome measures: whether the replication effect was within the prediction interval of the original effect and what the p-original was between the replication and original. Both these models had even more uncertain estimates. While the credible intervals were wide, the general patterns of predictor direction and relative strength were similar to the subjective replication models (See Supplement for plots of coefficients for these additional models). The strongest predictors for prediction interval were still within-subjects designs (0.62, CrI= [-0.18, 2.01]) and studies with larger effect sizes (0.3, CrI= [-0.29, 0.94]).

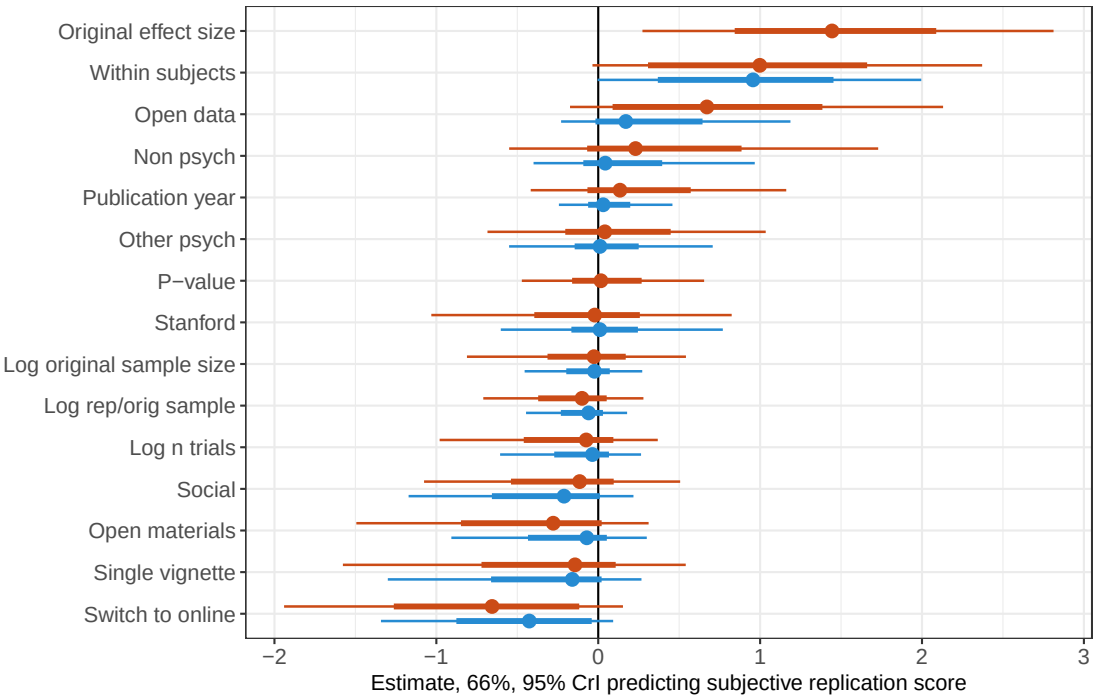

Figure 4: Coefficients and uncertaintly estimates from a model of all original-replication pairs (N=176, shown in blue) and a model of all pairs with full statistical information (N=112, shown in red) and predicting subjective replication scores as the dependent variable.

3 Discussion

Non-replications pose a problem for scientists who want to build on the empirical results in the literature, but the limited numbers of replications and limited research into specific predictors of replication failure mean that the reasons for non-replications are not well understood.

Here, we take a functional approach to assessing replicability, framing both our methods and interpretation around the idea of whether work can be repeated by an early-career scientist. We took advantage of 11 years of graduate student replication projects to look at correlational predictors of replication in a previously-unused dataset. In line with previous results, we found a 49% replication rate, with some studies showing effect sizes similar to the original and others much smaller. When we looked at individual correlates of replicability, within-subjects designs, work in the subfield of cognitive psychology, and the original and replication both using online samples stood out as the strongest correlates. As many of these predictors interrelate with one another, we ran regularized regressions with all the predictors at once. Due to our small sample, model estimates were uncertain, but within-subjects designs and large original effect sizes were the strongest predictors.

Replication results should be interpreted in light of their methods and estimand. Given our sample, we are able to estimate the correlates of studies replicating when the replications are conducted online, by graduate students with limited time and budget. We do not interpret our non-replications as indicating the original results were false positives (presumably some were and some weren't). There are many possible reasons for the non-replications in this sample. In some cases, the problem may be with the replication, such as too few participants, many exclusions for failed attention checks, or participants speeding through the study. When these issues are diagnosed, they suggest possible ways to "rescue" the replication by increasing the sample or changing the interface, without altering the underlying experiment; thus, while the replication did not succeed, after some troubleshooting, students may still be able to extend the work in the future. In other cases, there were a priori reasons to distrust the original study, such as exclusion criteria that seemed post-hoc or high-order interaction terms with a small sample. That said, not all scientists recognize the same factors as potential indications of low power or questionable research practices; students conducting these replications generally expected them to succeed. In many non-replication cases, it was unclear why the results failed to replicate.

We conducted an observational study of replication attempts, and our results are limited by the studies

we included, which are limited in number and may not be representative of the studies of interest to psychologists as a whole. Our predictor variables were not manipulated, so they cannot be interpreted as causing (non-)replication, but only as correlational markers. Some of the correlates are most easily interpreted as being about the original study, and others reflect the closeness of the replication to the original. For instance, while within-subjects designs are more likely to replicate than between-subjects designs, this could be related to power, or the types of experiments that tend to be run in each design. Given the predictive value, slightly more skepticism and critical reading of between-subjects designs may be warranted, but this correlation, by itself, does not mean scientists should prefer to run within-subjects designs.

Large scale replications are costly and arduous to run. The batch of replications presented here were pedagogical replications, done as part of a class. Trainee behavioral scientists need to learn experimental methods, and conducting replications as part of methods classes serve a dual purpose: they enables students to learn to do experiments in a scaffolded way, and they lead to more useful results than if students designed their own experiments from scratch on the same timescale (Frank & Saxe 2012, Wagge et al. 2019, Quintana 2021, Hawkins et al. n.d.). Pedagogy has an important role to play in open science more broadly – it's one thing to require or incentivize certain practices, but the tools and workflows of open science have to be learned.

4 Methods

Our pre-registration, code, and coded data are available at TODO OSF REPO.

4.1 Dataset

The dataset of replication projects comes from class projects conducted in PSYCH 251 (earlier called PSYCH 254) a graduate-level experimental methods class taught at Stanford by MCF from 2011 to 2022. This class is commonly taken by first year graduate students in psychology and related disciplines, and it has been a requirement of the Psychology PhD since around 2015. Each student chose a study to replicate, implemented the study, wrote analysis code, pre-registered their replication, ran the study, and turned in a structured final report including methods, analytic plan, changes from the original study, confirmatory and exploratory analyses, and discussion of outcomes. Students were encouraged to do experimental replications, but some students chose to replicate correlational outcomes or do computational reproducibility projects instead. We cannot include the full student reports for confidentiality reasons, but we include an example as well as the template given to students at TODO example and template.

Students were free to choose what study to replicate; the recommended path for students who did not have their own ideas was to pick an interesting study from a recent year of Psychological Science (this led to a high fraction of Psych Science articles in the replication sample, 80, 45% of studies).

We note that 4 (TODO check) of the replication projects were included in RP:P, and 10 of them were previously reported in Hawkins et al. (n.d.).

4.2 Coding procedure

We relied primarily on student reports to code the measured variables for the replications. We supplemented this with spreadsheets of information about projects from the time of the class and the original papers.

4.2.1 Measures of replication success

Our primary replication outcome was experimenter and instructor rated replication success. The subjective replication success was recorded by the teaching staff for the majority of class replications at the time they were conducted. Where the values were missing they were filled in by MCF on the basis of the reports. For all studies, replication success was independently coded by VB on the basis of the reports. Where VB's coding disagreed with the staff/MCF's code, the difference was resolved by discussion between VB and MCF (26% of studies). Subjective replication scores were coded on a [0, .25, .5, .75, 1] scale.

This subjective replication outcome was chosen because it already existed, could be applied to all projects (regardless of type and detail of statistical reporting), and did not rely solely on one statistical measure. As a complement, we also identified a "key" statistical test for each paper (see below for details), and if

possible, computed p-original and prediction interval at this statistic, following Mathur & VanderWeele (2020). p-original was a continuous measure of the p-value on the hypothesis that the original and replication samples come from the same distribution. Prediction interval was a binary measure of whether the replication outcome fell within the prediction interval of the original outcome measure.

4.2.2 Demographic properties

We coded the subfield of the original study as a 4 way factor: cognitive psychology, social psychology, other psychology, and non-psychology. For each paper, we coded its year of publication, whether it had open materials, whether it had open data, and whether it had been conducted using an online, crowd-sourced platform (i.e. MTurk or Prolific).

4.2.3 Experimental design properties

We coded experimental design on the basis of student reports, which often quoted from the original methods, and if that did not suffice, the original paper itself. To assess the role of repeated measures, we coded the number of trials seen per participant, including filler trials and trials in all conditions, but excluding training or practice trials.

We coded whether the manipulation in the study was instantiated in a single instance ("single vignette"). Studies with one induction or prime used per condition across participants were coded as having a single vignette. Studies with multiple instances of the manipulation (even if each participant only saw one) were coded as not being single vignette. While most studies with a single vignette only had one trial and vice versa, there were studies with a single induction and multiple test trials, and other studies with multiple scenarios instantiating the manipulation, but only one shown per participant.

We coded the number of subjects, post-exclusions. We coded whether a study had a between-subjects, within-subjects, or mixed design; for the analysis, mixed studies were counted as within-subjects designs. In the analysis, we used a log-scale for number of subjects and numbers of trials.

4.2.4 Properties of replication

We coded whether the replication was conducted on a crowd-sourced platform; this was the norm for the class projects, but a few were done in-person. For analysis, we coded this into a variable indicating if the recruitment platform changed between original and replication. This grouped the few in-person replications in with the studies that were originally online and stayed online in a "no change" condition, in contrast with the studies that were originally in-person with online replications.

We coded the replication sample size (after exclusions). This was transformed to the predictor variable log ratio of replication to original sample size.

As a control variable, we included whether the original authors were faculty at Stanford at the time of the replication. This was to account for potential non-independence of these replications (ex. if replicating their advisor's work, students may have access to extra information about methods).

We made note of studies to exclude for sensitivity analyses, due to not quite aligned statistics, extremely small or unbalanced sample sizes, or a student choosing a key statistical measure that was not of central importance to the original study.

4.2.5 Determination and coding of key statistical measure

For each study pair, we used one key measure of interest for which we calculated the predictor variables of p-value and effect size and the statistical outcome measures p_original and prediction interval. If the student specified a single key measure of interest and this was a measure that was reported in both the original paper and replication, we used that measure. If a student specified multiple, equally important, key measures, we used the first one. When students were not explicit about a key measure, we used other parts of their report (including introduction and power analysis) to determine what effect and therefore what result they considered key. In a few cases, we went back to the original paper to find what effect was considered crucial by the original authors. When the measures reported by the student did not cleanly match their explicit or implicitly stated key measure, we picked the most important (or first) of the measures that were reported in both the original and replication. These decisions could be somewhat subjective but importantly they were made without reference to replication outcomes.

Whenever possible, we used per-condition means and standard deviations, or the test statistic of the key measure and its corresponding degrees of freedom (ex. T test, F test). We took the original statistic from the replication report if it quoted the relevant analysis or from the original paper if not. We took the replication statistics from the replication report.

We then calculated p values, effect sizes, p-original, and prediction intervals. We choose to recalculate p values and effect sizes from the means or test statistic rather than use reported measures when possible because we thought this would be more reliable and transparent. The means and test statistics are more likely to have been outputted programmatically and copied directly into the text. In contrast, p-values are often reported as <.001 rather than as a point value, and effect size derivations may be error prone. By recording the raw statistics we used and using our available code to calculate other measures, we are transparent, as the test statistics can be searched for in the papers, and all processing is documented in code.

In some cases, p-values or effect sizes were not calculable either due to insufficient reporting (ex. reporting a p-value but no other statistics from a test) or key measures where p-values and effect sizes did not apply (ex. PCA as measure of interest). Where studies reported beta estimates and standard errors or proportions, standardized effects sizes are not an applicable measure, but we were still able to calculate p-original and prediction interval.

We separately coded whether the original and replication effects were in the same direction, based on raw means and graphs. This is more reliable than the statistics because F-tests don't include the direction of effect, and some students may have flipped the direction in coding for betas or t-tests. In the processed data, the direction of the effect of the replication was always coded consistently with the original study's coding, so a positive effect was in the same direction as the original and a negative effect in the opposite direction.

In regression analyses, we used standardized mean difference and log p-value as predictors.

4.3 Modelling

TODO citation around horseshoe prior

Due to the monotonic missingness of the data, we had more predictor variables and outcome variables for some original-replication pairs than for others. To take full advantage of the data, we ran a series of models, with some models having fewer predictors, but more data, and others having more predictors, but less data.

We ran a model predicting the subjective replication score on the basis of demographic and experimental predictors on the entire dataset. We ran two models predicting p_original and prediction interval from demographic and experimental predictors on the subset of data where we had p_original and prediction intervals. Then, on the smaller subset of the data where we had effect sizes and p-values, we re-ran these three models with those as additional predictor variables.

The subjective replication scores were coded on [0, .25, .5, .75, 1], and we ramapped these to 1-5 to run an ordinal regression predicting replication score. We ran logistic regressions predicting prediction interval and linear regressions predicting p_original.

We used a horseshoe shrinkage prior on the fixed effect coefficients because we had a lot of predictors compared to the amount of data (Carvalho et al. 2009). All models included random slopes for predictors nested within year the class occurred to control for variation between cohorts of students. We did not include any interaction terms in the models. All numeric predictor variables were z-scored after other transforms (e.g., logs) to ensure comparable regularization effects from the horseshoe prior. The priors we used were horseshoe(3) for betas, normal(0,.5) for standard deviation of random slopes, and lkj(1) for correlations between random slopes. Models were run in BRMS (Bürkner 2017).

As a secondary sensitivity analysis, we examined the subset of the data where the statistical tests had the same specification, the result was of primary importance in the original paper (i.e. not a manipulation check), and there were no big issues with the replication.

Results from these models not reported in the main paper are reported in the Supplement.

Acknowledgements

Acknowledge people here. `{-}` useful to not number this section.

References

Altmejd A, Dreber A, Forsell E, Huber J, Imai T, Johannesson M, Kirchler M, Nave G, Camerer C (2019) Predicting the replicability of social science lab experiments. *PLOS ONE* **14**:e0225826. doi:10.1371/journal.pone.0225826

Anderson CJ, Bahník Š, Barnett-Cowan M, Bosco FA, Chandler J, Chartier CR, Cheung F, Christopherson CD, Cordes A, Cremata EJ, Della Penna N, Estel V, Fedor A, Fitneva SA, Frank MC, Grange JA, Hartshorne JK, Hasselman F, Henninger F, Hulst M van der, Jonas KJ, Lai CK, Levitan CA, Miller JK, Moore KS, Meixner JM, Munafò MR, Neijenhuijs KI, Nilsonne G, Nosek BA, Plessow F, Prenoveau JM, Ricker AA, Schmidt K, Spies JR, Stieger S, Strohminger N, Sullivan GB, Aert RCM van, Assen MALM van, Vanpaemel W, Vianello M, Voracek M, Zuni K (2016) Response to Comment on "Estimating the reproducibility of psychological science." *Science* **351**:1037–1037. doi:10.1126/science.aad9163

Bürkner P-C (2017) brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* **80**:1–28. doi:10.18637/jss.v080.i01

Camerer CF, Dreber A, Holzmeister F, Ho T-H, Huber J, Johannesson M, Kirchler M, Nave G, Nosek BA, Pfeiffer T, Altmejd A, Buttrick N, Chan T, Chen Y, Forsell E, Gampa A, Heikensten E, Hummer L, Imai T, Isaksson S, Manfredi D, Rose J, Wagenmakers E-J, Wu H (2018) Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat Hum Behav* **2**:637–644. doi:10.1038/s41562-018-0399-z

Carvalho CM, Polson NG, Scott JG (2009) Handling sparsity via the horseshoe. In: Dyk D van, Welling M (eds) *Proceedings of the twelth international conference on artificial intelligence and statistics*. PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, p 73–80. Available from: https://proceedings.mlr.press/v5/carvalho09a.html

Crump MJC, McDonnell JV, Gureckis TM (2013) Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLOS ONE* **8**:e57410. doi:10.1371/journal.pone.0057410

Dreber A, Pfeiffer T, Almenberg J, Isaksson S, Wilson B, Chen Y, Nosek BA, Johannesson M (2015) Using prediction markets to estimate the reproducibility of scientific research. *Proc Natl Acad Sci* **112**:15343–15347. doi:10.1073/pnas.1516179112

Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, Banks JB, Baranski E, Bernstein MJ, Bonfiglio DBV, Boucher L, Brown ER, Budiman NI, Cairo AH, Capaldi CA, Chartier CR, Chung JM, Cicero DC, Coleman JA, Conway JG, Davis WE, Devos T, Fletcher MM, German K, Grahe JE, Hermann AD, Hicks JA, Honeycutt N, Humphrey B, Janus M, Johnson DJ, Joy-Gaba JA, Juzeler H, Keres A, Kinney D, Kirshenbaum J, Klein RA, Lucas RE, Lustgraaf CJN, Martin D, Menon M, Metzger M, Moloney JM, Morse PJ, Prislin R, Razza T, Re DE, Rule NO, Sacco DF, Sauerberger K, Shrider E, Shultz M, Siemsen C, Sobocko K, Weylin Sternglanz R, Summerville A, Tskhay KO, Allen Z van, Vaughn LA, Walker RJ, Weinberg A, Wilson JP, Wirth JH, Wortman J, Nosek BA (2016) Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology* **67**:68–82. doi:10.1016/j.jesp.2015.10.012

Ebersole CR, Mathur MB, Baranski E, Bart-Plange D-J, Buttrick NR, Chartier CR, Corker KS, Corley M, Hartshorne JK, IJzerman H, Lazarević LB, Rabagliati H, Ropovik I, Aczel B, Aeschbach LF, Andrighetto L, Arnal JD, Arrow H, Babincak P, Bakos BE, Baník G, Baskin E, Belopavlović R, Bernstein MH, Białek M, Bloxsom NG, Bodroža B, Bonfiglio DBV, Boucher L, Brühlmann F, Brumbaugh CC, Casini E, Chen Y, Chiorri C, Chopik WJ, Christ O, Ciunci AM, Claypool HM, Coary S, Čolić MV, Collins WM, Curran PG, Day CR, Dering B, Dreber A, Edlund JE, Falcão F, Fedor A, Feinberg L, Ferguson IR, Ford M, Frank MC, Fryberger E, Garinther A, Gawryluk K, Ashbaugh K, Giacomantonio M, Giessner SR, Grahe JE, Guadagno RE, Hałasa E, Hancock PJB, Hilliard RA, Hüffmeier J, Hughes S, Idzikowska K, Inzlicht M, Jern A, Jiménez-Leal W, Johannesson M, Joy-Gaba JA, Kauff M, Kellier DJ, Kessinger G, Kidwell MC, Kimbrough AM, King JPJ, Kolb VS, Kołodziej S, Kovacs M, Krasuska K, Kraus S, Krueger LE, Kuchno K, Lage CA, Langford EV, Levitan CA, Lima TJS de, Lin H, Lins S, Loy JE, Manfredi D, Markiewicz Ł, Menon M, Mercier B, Metzger M, Meyet V, Millen AE, Miller JK, Montealegre A, Moore DA, Muda R, Nave G, Nichols AL, Novak SA, Nunnally C, Orlić A, Palinkas A, Panno A, Parks KP, Pedović I, Pękala E, Penner MR, Pessers S, Petrović B, Pfeiffer T, Pieńkosz D, Preti E, Purić D, Ramos T, Ravid J, Razza TS, Rentzsch K, Richetin J, Rife SC, Rosa AD, Rudy KH, Salamon J, Saunders B, Sawicki P, Schmidt K, Schuepfer

K, Schultze T, Schulz-Hardt S, Schütz A, Shabazian AN, Shubella RL, Siegel A, Silva R, Sioma B, Skorb L, Souza LEC de, Steegen S, Stein LAR, Sternglanz RW, Stojilović D, Storage D, Sullivan GB, Szaszi B, Szecsi P, Szöke O, Szuts A, Thomae M, Tidwell ND, Tocco C, Torka A-K, Tuerlinckx F, Vanpaemel W, Vaughn LA, Vianello M, Viganola D, Vlachou M, Walker RJ, Weissgerber SC, Wichman AL, Wiggins BJ, Wolf D, Wood MJ, Zealley D, Žeželj I, Zrubka M, Nosek BA (2020) Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. *Adv Methods Pract Psychol Sci* **3**:309–331. doi:10.1177/2515245920958687

Etz A, Vandekerckhove J (2016) A Bayesian Perspective on the Reproducibility Project: Psychology. *PLOS ONE* **11**:e0149794. doi:10.1371/journal.pone.0149794

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

Frank MC, Saxe R (2012) Teaching Replication: *Perspect Psychol Sci*. doi:10.1177/1745691612460686

Gelman A (2018/ed) Don't characterize replications as successes or failures. *Behav Brain Sci* **41**:e128. doi:10.1017/S0140525X18000638

Gilbert DT, King G, Pettigrew S, Wilson TD (2016) Comment on "Estimating the reproducibility of psychological science." *Science* **351**:1037–1037. doi:10.1126/science.aad7243

Hagger MS, Chatzisarantis NLD, Alberts H, Anggono CO, Batailler C, Birt AR, Brand R, Brandt MJ, Brewer G, Bruyneel S, Calvillo DP, Campbell WK, Cannon PR, Carlucci M, Carruth NP, Cheung T, Crowell A, De Ridder DTD, Dewitte S, Elson M, Evans JR, Fay BA, Fennis BM, Finley A, Francis Z, Heise E, Hoemann H, Inzlicht M, Koole SL, Koppel L, Kroese F, Lange F, Lau K, Lynch BP, Martijn C, Merckelbach H, Mills NV, Michirev A, Miyake A, Mosser AE, Muise M, Muller D, Muzi M, Nalis D, Nurwanti R, Otgaar H, Philipp MC, Primoceri P, Rentzsch K, Ringos L, Schlinkert C, Schmeichel BJ, Schoch SF, Schrama M, Schütz A, Stamos A, Tinghög G, Ullrich J, vanDellen M, Wimbarti S, Wolff W, Yusainy C, Zerhouni O, Zwienenberg M (2016) A Multilab Preregistered Replication of the Ego-Depletion Effect. *Perspect Psychol Sci* **11**:546–573. doi:10.1177/1745691616652873

Hawkins RXD, Smith EN, Au C, Arias JM, Hermann E, Keil M, Lampinen A, Raposo S, Salehi S, Salloum J, Tan J, Frank MC Improving the Replicability of Psychological Science Through Pedagogy. :41

Hoogeveen S, Sarafoglou A, Wagenmakers E-J (2019) Laypeople Can Predict Which Social Science Studies Replicate. preprint. PsyArXiv. Available from: https://osf.io/egw9d [Last accessed 30 September 2019]. doi:10.31234/osf.io/egw9d

Klein RA, Ratliff KA, Vianello M, Adams RB, Bahník Š, Bernstein MJ, Bocian K, Brandt MJ, Brooks B, Brumbaugh CC, Cemalcilar Z, Chandler J, Cheong W, Davis WE, Devos T, Eisner M, Frankowska N, Furrow D, Galliani EM, Hasselman F, Hicks JA, Hovermale JF, Hunt SJ, Huntsinger JR, IJzerman H, John M-S, Joy-Gaba JA, Barry Kappes H, Krueger LE, Kurtz J, Levitan CA, Mallett RK, Morris WL, Nelson AJ, Nier JA, Packard G, Pilati R, Rutchick AM, Schmidt K, Skorinko JL, Smith R, Steiner TG, Storbeck J, Van Swol LM, Thompson D, Veer AE van 't, Ann Vaughn L, Vranka M, Wichman AL, Woodzicka JA, Nosek BA (2014) Investigating Variation in Replicability: A "Many Labs" Replication Project. *Social Psychology* **45**:142–152. doi:10.1027/1864-9335/a000178

Klein RA, Vianello M, Hasselman F, Adams BG, Adams RB, Alper S, Aveyard M, Axt JR, Babalola MT, Bahník Š, Batra R, Berkics M, Bernstein MJ, Berry DR, Bialobrzeska O, Binan ED, Bocian K, Brandt MJ, Busching R, Rédei AC, Cai H, Cambier F, Cantarero K, Carmichael CL, Ceric F, Chandler J, Chang J-H, Chatard A, Chen EE, Cheong W, Cicero DC, Coen S, Coleman JA, Collisson B, Conway MA, Corker KS, Curran PG, Cushman F, Dagona ZK, Dalgar I, Dalla Rosa A, Davis WE, Bruijn M de, De Schutter L, Devos T, Vries M de, Doğulu C, Dozo N, Dukes KN, Dunham Y, Durrheim K, Ebersole CR, Edlund JE, Eller A, English AS, Finck C, Frankowska N, Freyre M-Á, Friedman M, Galliani EM, Gandi JC, Ghoshal T, Giessner SR, Gill T, Gnambs T, Gómez Á, González R, Graham J, Grahe JE, Grahek I, Green EGT, Hai K, Haigh M, Haines EL, Hall MP, Heffernan ME, Hicks JA, Houdek P, Huntsinger JR, Huynh HP, IJzerman H, Inbar Y, Innes-Ker ÅH, Jiménez-Leal W, John M-S, Joy-Gaba JA, Kamiloğlu RG, Kappes HB, Karabati S, Karick H, Keller VN, Kende A, Kervyn N, Knežević G, Kovacs C, Krueger LE, Kurapov G, Kurtz J, Lakens D, Lazarević LB, Levitan CA, Lewis NA, Lins S, Lipsey NP, Losee JE, Maassen E, Maitner AT, Malingumu W, Mallett RK, Marotta SA, Međedović J, Mena-Pacheco F, Milfont TL, Morris WL, Murphy SC, Myachykov A, Neave N, Neijenhuijs K, Nelson AJ, Neto F, Lee Nichols A, Ocampo A, O'Donnell SL, Oikawa H, Oikawa M, Ong E, Orosz G, Osowiecka M, Packard G, Pérez-Sánchez R, Petrović B, Pilati R, Pinter B, Podesta L, Pogge G, Pollmann MMH, Rutchick AM, Saavedra P, Saeri AK, Salomon E, Schmidt K, Schönbrodt FD, Sekerdej MB, Sirlopú D, Skorinko JLM, Smith MA, Smith-Castro V, Smolders KCHJ, Sobkow A, Sowden W, Spachtholz P, Srivastava M, Steiner TG, Stouten J, Street CNH, Sundfelt OK, Szeto S,

Szumowska E, Tang ACW, Tanzer N, Tear MJ, Theriault J, Thomae M, Torres D, Traczyk J, Tybur JM, Ujhelyi A, Aert RCM van, Assen MALM van, Hulst M van der, Lange PAM van, Veer AE van 't, Vásquez- Echeverría A, Ann Vaughn L, Vázquez A, Vega LD, Verniers C, Verschoor M, Voermans IPJ, Vranka MA, Welch C, Wichman AL, Williams LA, Wood M, Woodzicka JA, Wronska MK, Young L, Zelenski JM, Zhijia Z, Nosek BA (2018) Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Adv Methods Pract Psychol Sci* **1**:443–490. doi:10.1177/2515245918810225

Klein RA, Cook CL, Ebersole CR, Vitiello C, Nosek BA, Hilgard J, Ahn PH, Brady AJ, Chartier CR, Christopherson CD, Clay S, Collisson B, Crawford JT, Cromar R, Gardiner G, Gosnell CL, Grahe J, Hall C, Howard I, Joy-Gaba JA, Kolb M, Legg AM, Levitan CA, Mancini AD, Manfredi D, Miller J, Nave G, Redford L, Schlitz I, Schmidt K, Skorinko JLM, Storage D, Swanson T, Van Swol LM, Vaughn LA, Vidamuerte D, Wiggins B, Ratliff KA (2022) Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement. *Collabra: Psychology* **8**:35271. doi:10.1525/collabra.35271

Mathur MB, VanderWeele TJ (2020) New statistical metrics for multisite replication projects. *J R Stat Soc Ser A Stat Soc* **183**:1145–1166. doi:10.1111/rssa.12572

Olsson-Collentine A, Assen MA van, Wicherts J (2020) Heterogeneity in direct replications in psychology and its association with effect size.

Open Science Consortium (2015) Estimating the reproducibility of psychological science. *Science*

Patil P, Peng RD, Leek JT (2016) What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspect Psychol Sci* **11**:539–544. doi:10.1177/1745691616646366

Protzko J, Krosnick J, Nelson LD, Nosek BA, Axt J, Berent M, Buttrick N, DeBell M, Ebersole CR, Lundmark S, MacInnis B, O'Donnell M, Perfecto H, Pustejovsky JE, Roeder SS, Walleczek J, Schooler J (2020) High Replicability of Newly-Discovered Social-behavioral Findings is Achievable. preprint. PsyArXiv. Available from: https://osf.io/n2a9x [Last accessed 5 April 2023]. doi:10.31234/osf.io/n2a9x

Quintana DS (2021) Replication studies for undergraduate theses to improve science and education. *Nat Hum Behav* **5**:1117–1118. doi:10.1038/s41562-021-01192-8

Ramscar M, Shaoul C, Baayen RH Why many priming results don't (and won't) replicate: A quantitative analysis.

Simonsohn U (2015) Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychol Sci* **26**:559–569. doi:10.1177/0956797614567341

Van Bavel JJ, Mende-Siedlecki P, Brady WJ, Reinero DA (2016) Contextual sensitivity in scientific reproducibility. *Proc Natl Acad Sci* **113**:6454–6459. doi:10.1073/pnas.1521897113

Wagge JR, Brandt MJ, Lazarevic LB, Legate N, Christopherson C, Wiggins B, Grahe JE (2019) Publishing Research With Undergraduate Students via Replication Work: The Collaborative Replications and Education Project. *Front Psychol* **10**

Wilson BM, Harris CR, Wixted JT (2020) Science is not a signal detection problem. *Proc Natl Acad Sci USA* **117**:5559–5567. doi:10.1073/pnas.1914237117

Yang Y, Youyou W, Uzzi B (2020) Estimating the deep replicability of scientific findings using human and artificial intelligence. *Proc Natl Acad Sci* **117**:10762–10768. doi:10.1073/pnas.1909046117

Youyou W, Yang Y, Uzzi B (2023) A discipline-wide investigation of the replicability of Psychology papers over the past two decades. *Proc Natl Acad Sci* **120**:e2208863120. doi:10.1073/pnas.2208863120