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 extension. Yet in psychology, some prominent findings have failed to replicate, and large-scale studies suggest replicability issues are widespread. The identification of predictors of replication success is limited by the difficulty of conducting large samples of independent replication experiments, however: most investigations 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 judged to be successful in 49% of replications; of the 136 where effect sizes could be numerically compared, 46% had point estimates within the prediction interval of the original outcome (versus the expected 95%). Larger original effect sizes and within-participants designs were especially related to replication success. Our results indicate that, consistent with prior reports, the robustness of the psychology literature is low enough to limit cumulative progress by student investigators.

1 Introduction

Cumulative scientific progress requires empirical results that are robust enough to support both future empirical extensions and the construction of synthetic theories. 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). One early large-scale replication project pegged the replication rate for findings in top-tier psychology journals at around 40% [RP:P; Open Science Consortium (2015)]. Low replicability has negative consequences for the field as a whole: when scientists attempt to build on published results, they stand a good chance of meeting with failure.

Addressing this issue requires a better understanding of the scope of the problem as well as the methodological and structural issues that might lead to replication failure (Simmons et al. 2011, e.g., Smaldino & McElreath 2016). Estimating replicability in the literature is a key starting point, but, as we review below, there is limited consensus on what quantity exactly should be estimated.

Our viewpoint here is that one important estimand is the probability that a graduate student can identify a finding in the literature and replicate it successfully enough that they can build on that finding in subsequent work. Taking this perspective, our contribution is a new dataset of 176 replications of experimental studies from the social sciences, primarily psychology. These replications were conducted as individual course projects by students in a graduate-level experimental methods class between 2011 and 2022. We use this dataset to investigate the rate of replicability for such projects as well as the correlates of replication success.

1.1 What are we estimating when we measure replicability?

A few large-scale investigations have measured replication rates in samples of psychology studies. The first of these, RP:P, sampled roughly 100 studies from articles published in three top psychology journals in 2008 and distributed the studies across participating labs, finding an overall replication rate of around 40% (Open Science Consortium 2015). The followup Many Labs studies investigated heterogeneity using short target studies that each compared two conditions. These study designs were not representative of the

*Corresponding author. Email: vboyce@stanford.edu

psychology literature as a whole, and due to the goal of measuring heterogeneity, they had large overall samples across multiple sites. Across Many Labs 1–3, only 29 of 51 target effects (57%) replicated (Klein et al. 2014, Ebersole et al. 2016, Klein et al. 2018). Camerer et al. (2018) included all 21 behavioral social science studies from Nature and Science from 2010-2015 that were feasible to replicate. They consulted original authors and had high power to detect even small effects; under these conditions and with this sample, the replication rate was around 60%.

While their sampling procedure and methods varied, these previous approaches to replicability have all focused on interpreting their results in terms of a potentially problematic estimand: the probability of a finding in the literature being 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. 2015).

Further, the methods used in these studies are not sufficient to yield an unbiased estimate of this quantity. In no case were studies randomly sampled from the literature; instead replication projects sampled from specific journals and adjusted the sample for feasibility concerns. These reasonable decisions further undermine the interpretation of the results from these studies as representing the proportion of true findings in the psychology literature as a whole.

Rather than aim for some measure of true replicability, perhaps we should contextualize the true estimand for replication efforts based on their methodologies and outcome measures. Through this lens, RP:P estimated the rate at which findings from relatively simple experiments published in a few well-known journals at a particular time could be replicated in a typical psychology lab. Many Labs estimated the rate at which well-known, two-condition findings replicate in very large samples. Camerer et al. (2018) estimated which prestigious journal findings replicate when conducted in a highly-resourced environment with expert involvement. All of these could be potentially desirable estimands.

But in practice, most scientific work is conducted by graduate students with limited time, limited budgets, and limited access to experts. If students cannot replicate a finding under these circumstances, they cannot build on it in their own empirical or theoretical work. How replicable are findings in the literature for graduate students operating under these less-than-ideal conditions? We address this question here.

Our current sample of replications is selected based on what experiments students were interested in and wanted to replicate, with some filtering for feasibility. This sampling is not at all random. It reflects how scientists generally choose which experiments to build on: those that are interesting and relatively feasible given methodological and budgetary constraints.

In the current study, we estimate the probability of successful replication in this sample, with the goal of also identifying markers of when findings can (and cannot) support cumulative science. Our hope is to extend previous work that has attempted to find key correlates of replication success.

1.2 When do replications succeed?

Despite variation in the methods and outcomes used by large-scale replication studies, they are often aggregated together in analyses looking at the predictability of replication success. Prediction markets and elicitations have established that people can predict above chance what studies will replicate (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), though again the precise features relating to success are unclear.

In search of such features, RP:P examined 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. Using these same data combined with a few other smaller samples, Altmejd et al. (2019) examined statistical and demographic features of replication studies and identified larger sample sizes, larger effect sizes, and simple effects (as opposed to interaction terms) as predictive of replication.[1]

These approaches are fundamentally limited by the available data. Large-scale replications are arduous

[1] While most approaches to correlates of replicability have been correlational, experimental approaches can be used to test potential interventions (Ebersole et al. 2020, Protzko et al. 2020). Experiments can be very valuable as tests of specific causes of non-replication, but they are expensive and time-consuming to conduct, and not all factors affecting replication success can be manipulated experimentally.

and expensive to run, so only a few large-scale replication datasets exist, and most 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 over-fit to the 100 studies included in RP:P.

Further, none of these studies focused on features of experimental design, such as within-participants designs or the use of repeated measures. There has been speculation that both of these factors should be linked to increased replicability due to their role in enabling increased statistical precision. Within-participants designs lead to more precise experimental estimates by allowing the estimation of correlated person-level variation across conditions; repeated measures allow for more precise estimation by averaging out measurement error. Both are often recommended by methodologists as part of good measurement practices, at least when they are feasible (Greenwald 1976, Rosenthal & Rosnow 2008, Frank et al. 2023).

In sum, our current study examines the overall rate of replicability as well as the statistical and design features that predict replicability in a new sample of student replications.

2 Results

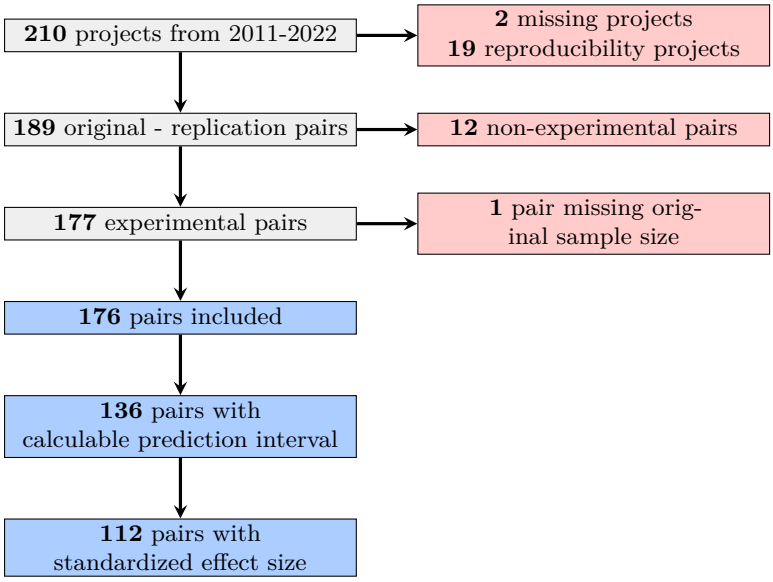

Figure 1: Of the 210 projects conducted for the class, 176 are included in our analysis, after excluding reproducibilty projects (with no new data collection), non-experimental replications, and missing projects. Of the 176, 136 report sufficient information to calculate prediction intervals, and 112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

PSYCH 251 is a graduate-level experimental methods class in experimental psychology taught at Stanford University. During the 10 week class, each student replicates a published finding. They individually re-implement the study, write analysis code, pre-register their study, collect data (typically using an online crowd-sourcing platform), and write a structured replication report. Students in the course are free to choose studies related to their research interests, with the default recommendation being an article from a recent year of Psychological Science.

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

These replications were all conducted on short time scales, within a constrained class budget. In some cases the budget limited the number of participants who could be recruited, occasionally below what the

[2]Replications had a median of 59 participants (interquartile range: 31 - 124), while original studies had a median of 101 participants (interquartile range: 40 - 180).

original study included or what power analyses suggested. The replications had a median post-exclusion sample size that was 86% of the original sample size (interquartile range 41%-105%)[2]. In nearly all cases, replications were conducted online, with recruitment from Amazon Mechanical Turk (the default from 2011 to 2020) or Prolific (the default from 2021 – 2022).

Many different measures can be used to define replication success of an individual statistical result (Gelman 2018/ed, Simonsohn 2015, Mathur & VanderWeele 2020). However, whether a replication should be considered successful is not always dependent on only one statistical comparison between the two studies. Often in original papers, multiple statistical tests are cited in support of the claim that a pattern of results matches a particular theoretical expectation.

As our primary outcome, we chose to use a subjective replication score (coded by two independent raters – one typically at the time of project completion – with discrepancies resolved by discussion). Unlike statistical measures, subjective replication success accommodates studies with multiple important outcome measures that together define the pattern of interest. Further, this measure was applicable across the diverse range of statistical measures and reporting practices present in the sample.

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, these two statistical metrics each quantify the similarity between a key statistic in the original study and the replication, but they will not always match researcher intuitions on whether a study replicated.

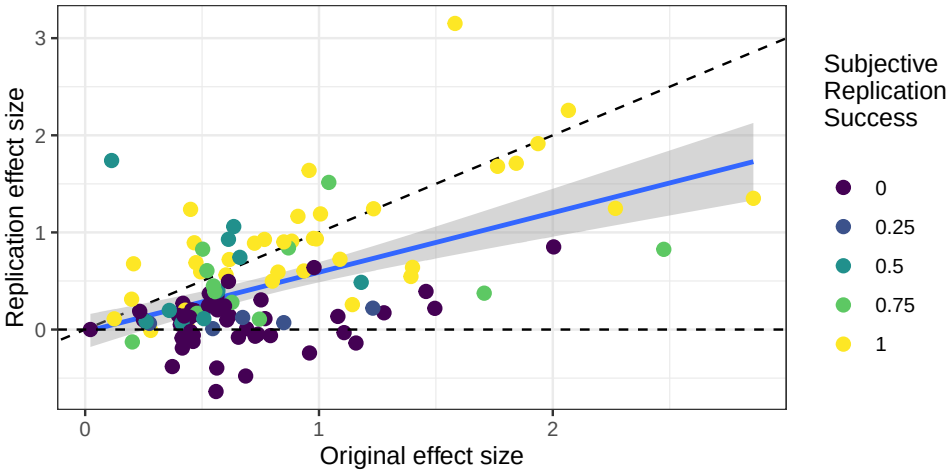

Figure 2: Relationship between effect size of the original study, effect size of the replication study, and subjective replication success rating, for those studies where effect size was applicable (N=112).

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 with 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 results of 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). As a test of sensitivity to heterogeneity, we assumed that the level of heterogeneity in hypothetical multi-site replications of our sampled articles was the same as the average level heterogeneity found by Olsson-Collentine et al. (2020) in prior multi-site replications in psychology ($\tau = .21$ in units of standardized mean difference). Under this assumption, 64% (72/112) of replication effect sizes are distributionally consistent with the original effect size. However, 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.

Table 1: The unadjusted Pearson correlations between each individual predictor and the subjective replication score. See Methods for how these variables were coded.

r	p	Predictors
0.333	0.000	Within participants design (versus between participants)
0.182	0.015	Log number of trials
0.150	0.047	Open data
0.080	0.294	Non psychology (versus cognitive psych)
0.075	0.322	Other psychology (versus cognitive psych)
0.064	0.399	Publication year
0.002	0.979	Open materials
-0.027	0.725	Stanford affiliation of original authors at time of replication
-0.047	0.536	Log ratio between replication and original sample sizes
-0.108	0.155	Log original sample size
-0.158	0.037	Switch to online for replication (versus same modality for original and replication)
-0.246	0.001	Social psychology (versus cognitive psych)
-0.267	0.000	Single vignette (versus multiple items/inductions per condition)

2.2 Bivariate correlates of replication success

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 the Methods.

Many predictors correlated with subjective replication success using unadjusted Pearson correlations (Table 1). Predictors of higher replicability included within-participants designs, larger numbers of trials per participant, and the original study having openly accessible data. Predictors of lower replicability included single-vignette studies (those with only one experimental stimulus per condition), classification as social psychology, and study 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. Both social and cognitive psychology studies were well represented in the replication sample, and the cognitive psychology studies replicated more often than social psychology studies (mean subjective replication scores: 0.58 versus 0.36). Within and between participants designs were both common, and within-participants designs replicated more often than between-participants designs (mean scores: 0.65 versus 0.36). Studies with multiple vignettes replicated more often than single vignetted studies (mean scores: 0.59 versus 0.36). However, there were strong correlations among these experimental features as well as between these experimental features and specific subfields.

Studies with open data, which almost always also had open materials as well, tended to replicate more than studies without open data. Nearly all replication 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 (e.g., inductions might have been weaker or instructions might have been insufficient). These factors – open materials, open data, and online samples – are more common in more recent studies, and so these effects may partially reflect temporal trends.

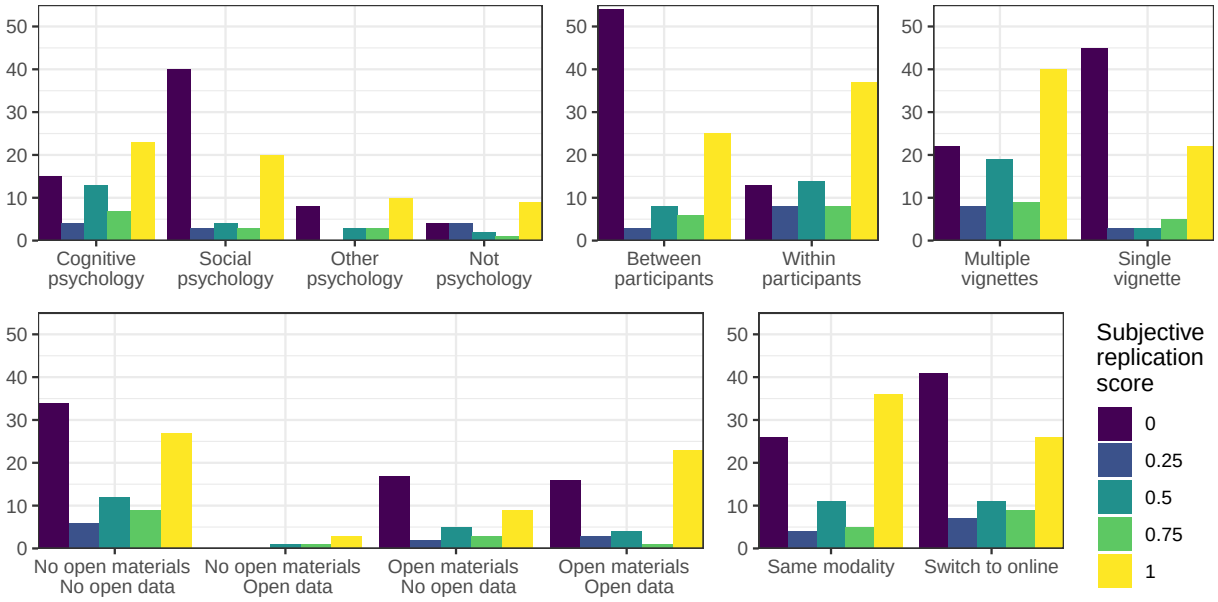

Figure 3: Distribution of subjective replication scores within categories. Bar heights are counts of studies.

2.3 Joint evaluation of the predictors of replicability

While a number of features show individual correlations with the subjective replication score, many correlate with one another. To determine which predictors were the strongest, we modeled subjective replication score as an ordinal outcome using ordinal Bayesian regression models with a logit link function, regularized using a horseshoe shrinkage prior (Carvalho et al. 2009). This model estimates odds ratios representing the association of the predictor with having a higher versus lower subjective replication score. We first ran models using all of the original-replication pairs (N=176), but without original effect size and original p value as predictors, as they were uncodable for some pairs. We next ran models including all predictors, but on only the subset of data where all predictors were available (N=112).

Coefficient estimates from the two ordinal models predicting the subjective replication scores are shown in Figure 4. 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-participants designs stand out as the strongest correlate of replicability in the model with all the data (OR = 2.56, 95% CrI = [1, 7.36]). In the model with all predictors, but less data, within-participants designs remain strong (OR = 2.84, 95% CrI = [0.95, 10.9]), and standardized effect size is also substantially related to subjective replication score (OR = 3.32, 95% CrI = [1.04, 11.99]). Both effects are robust to a sensitivity analysis including only studies with close replications and matching statistical tests (within-participants OR = 3.7, 95% CrI = [0.98, 21.2]; effect size OR = 7.34, 95% CrI = [1.25, 48.65]).

We also ran models predicting our secondary outcome measures: a logistic model predicting whether the replication effect was within the prediction interval of the original effect, and a linear model predicting what the p-original was between the replication and original. Both these models had 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 results from all models). The strongest predictors for prediction interval were still within-participants designs (OR = 1.89, 95% CrI

= [0.84, 8.98]) and studies with larger effect sizes (OR = 1.2, 95% CrI = [0.75, 2.81]).

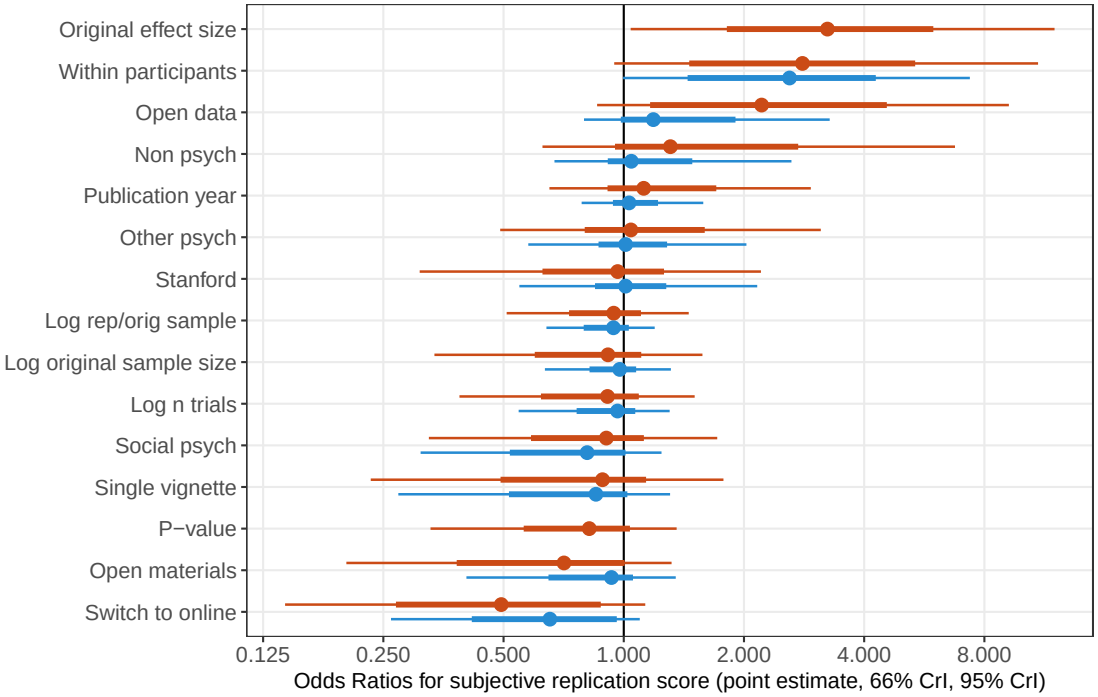

Figure 4: Odds ratios and corresponding 95% CrI on the likelihood of having a higher subjective replication score as a function of the independent variable. Estimates from a model of all original-replication pairs (N=176) are shown in blue, and from a model of all pairs with full statistical information (N=112) are shown in red. A value of 1 indicates no association, greater than one indicates an association with higher replication scores and less than 1 an association with lower replication scores.

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 study both the overall level of replication success and the 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. Within-participants designs, work in the subfield of cognitive psychology, and the original and replication both using online samples stood out as the strongest correlates of replication success. 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-participants designs and large original effect sizes were the strongest predictors.

We do not interpret our non-replications as indicating the original results were false positives: presumably some were and some were not. 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 can 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 – perhaps most – non-replications in our sample, it was unclear why the results failed to replicate.

Our results are limited by the sample of studies we included, which are limited in number and may not be representative of the studies of interest to psychologists as a whole. Further, 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-participants designs are more likely to replicate than between-participants designs, this predictor could also be related to the types of experiments that tend to be run in each design. Given the strong relationship of this factor to replicability, slightly more skepticism and critical reading of between-participants designs may be warranted, but this correlation, by itself, does not mean scientists should prefer within-participants designs.

Large scale replication studies are costly and arduous. The batch of replications presented here were pedagogical replications, done as part of a class. Trainees must learn experimental methods, and conducting replications as part of methods classes serves several purposes: it enables students to learn to do experiments in a supportive context, it often leads to more useful results than if students designed their own experiments from scratch, and it creates a resource for studying the literature (Frank & Saxe 2012, Hawkins et al. 2018, Wagge et al. 2019, Quintana 2021). We believe that this kind of pedagogy has an important role to play in improving methodological practices in psychology more broadly. The tools and workflows of rigorous, replicable science cannot simply be mandated: they have to be learned.

4 Methods

Our pre-registration, code, and coded data are available at https://osf.io/xwn9m/.

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 over 50 reports that we received permission to share and the template given to students at https://osf.io/xwn9m/.

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).

Four of the replication projects were included in RP:P, and 10 were previously reported in Hawkins et al. (2018) (which reported 11 student replications from the 2015-2016 class, one of those was excluded from the current sample for being non-experimental).

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

All original-replication pairs (except for one) had codable demographic and experimental features, while fewer pairs had codable effect sizes on some consistent scale, and fewer still had codable p-values and SMD effect sizes. Thus, we had more predictor variables and outcome variables for some original-replication pairs than for others, but which variables were codable was monotonic. 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 predicting whether the replication was in the 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

We are grateful to the students in PSYCH 254 and PSYCH 251 for conducting the replication projects included here. We thank Ben Prystawski and the FriSem audience for feedback on earlier versions of this work. MM was supported by R01LM013866.

Author Contributions

VB and MCF report no conflicts of interest. MM is the Associate Director of the Stanford Center for Open and Reproducible Science.

The authors made the following contributions: VB: Methodology, Data Curation, Formal Analysis, Writing - original draft; MM: Methodology, Writing - review & editing; MCF: Conceptualization, Methodology, Writing - review & editing.

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 án, 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,..., 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

Frank MC, Braginsky M, Cachia J, Coles N, Hardwicke T, Hawkins R, Mathur MB, Williams R (2023) Experimentology: An open science approach to experimental psychology methods.

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

Greenwald AG (1976) Within-subjects designs: To use or not to use? *Psychological Bulletin* **83**:314

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 RD, Smith EN, Au C, Arias JM, Catapano R, Hermann E, Keil M, Lampinen A, Raposo S, Reynolds J, Salehi S, Salloum J, Tan J, Frank MC (2018) Improving the replicability of psychological science through pedagogy. *Advances in Methods and Practices in Psychological Science* **1**:7–18

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 án, Bernstein MJ,..., 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,..., 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 (2015) Why many priming results don't (and won't) replicate: A quantitative analysis.
Rosenthal R, Rosnow RL (2008) Essentials of behavioral research: Methods and data analysis.
Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological science* **22**:1359–1366
Simonsohn U (2015) Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychol Sci* **26**:559–569. doi:10.1177/0956797614567341
Smaldino PE, McElreath R (2016) The natural selection of bad science. *Royal Society Open Science* **3**:160384. doi:10.1098/rsos.160384
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**
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