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

A cumulative science relies on knowing how much to trust published findings. Large-scale replication studies provide a measure of how reliable the published literature in a field is by determining what fraction of studies replicate. Because of the effort involved in conducting a large number of replications, few large-scale replications are available. We add 176 replication projects conducted by students of experimental studies primarily in psychology. Roughly half the studies were replicated successfully, and we report how features of the studies and the replications influenced the replication outcome.

introduction outline:

Things don't replicate that well and it feels like a problem for cumulative science * For a cumulative science, we want (many) of the load-bearing empirical results that form the bases of theories to hold up. * People were acting as if publication of results meant they were true/reliable, trying to build on things; caused crises when some prominent results failed to replicate. Called into question reliability of literature as a whole – lead to a few measurements. * Large scale replications of chunks of the literature, but aren't many (list and explain) * These rates feel lower than we want

Why? Many approaches, no definitive answers. * Why are rates what they are. Both want to know to fix but also to watch out / calibrate caution. Can be approached from theoretical, interventional, and correlational dimensions. * Theoretical – we know people do bad practices and can model how this leads to lower replicability. Requires things we already have ideas about * Interventional – try things they think will make it better (like not doing bad practices), see what happens. Not everything is intervention prone, need to have strong belief first. Expensive! * Correlational – markets/etc say there is some predictive value to be had, but don't say if it's transparent and explainable. RP:P. - Note that it's correlational, so can't do causality, but still useful for prediction, and as fodder for theory & intervention. (Interplay between them)

What's missing from the conversation and why our approach is good. * There's been a lot of discussion around replicabilty and it's role and yadda yadda (much overfocused on RP:P results because that's what's available). Growing sophistication and valid points, but we think it misses the mark about what practicing scientists are concerned about. * Don't want to build on something and fail a bunch trying to get your variant to work and then realize it just doesn't work. Replicability (as we mean it here) A useful framing is "ability to do cumulative science" – close to what people mean, may disagree with many things * How this approach differs and what it gets us * (built in feasibility under reasonable / small resource constraints) * interest based sampling * functional approach sidesteps issues of "truth" (may even be reliable in a different environment, but would be)

Teaser: * We do x,y,z, and find that larger effect sizes and within subjects designs are the two strongest correlates of replicability in this sample.

Questions: - mention primary outcome in intro? - switch out fig 4 for supp fig 2? - do we want to use the term p(cumulative)?

*Corresponding author. Email: vboyce@stanford.edu

TODO fill in cites flesh out theoretical approaches section

introduction outline:

Things don't replicate that well and it feels like a problem for cumulative science * For a cumulative science, we want (many) of the load-bearing empirical results that form the bases of theories to hold up. For a cumulative science, the empirical results that form the bases for theories and * People were acting as if publication of results meant they were true/reliable, trying to build on things; caused crises when some prominent results failed to replicate. Called into question reliability of literature as a whole – lead to a few measurements. A major concern in psychology, as in other fields, is that the published literature may not replicate as well as had previously been assumed. This is illustrated both by a few prominent findings that turn out to not replicate at all [ex. terror management theory; Klein et al. (2022)], and by large-scale replication projects such as RP:P which found a replication rate around 40% for a sample of the psychology literature (Consortium 2015). This low replication rates seems undesirable and seems to run counter to the idea of a cumulative science.

- Large scale replications of chunks of the literature, but aren't many (list and explain) We are aware of three large-scale replication efforts replicating experimental results in the existing psychology literature. RP:P sampled roughly 100 studies from articles published in three top psychology journals in 2008 and found a overall replication rate of around 40% (Consortium 2015). The ManyLabs series of replications investigating heterogeneity has sampled target studies that are short and compare only two conditions, which is not representative of the psychology literature as a whole. Across Many Labs 1-3, 29 of 51 target effects (57%) were found to replicate (Klein et al. 2014, Ebersole et al. 2016, Klein et al. 2018). Camerer et al. (2018) found a roughly 60% replication rate when they replicated all 21 behaviors studies published in Nature and Science from 2010-2015 that did not require special populations or special equipment. Those are all the large-scale replication projects of experimental psychology results available at this time. These roughly 170 results are the primary empirical results discussions of replication in pschology have to rely are.
- These rates feel lower than we want

Why? Many approaches, no definitive answers. What makes the replication rates in psychology so low? Knowing what predicts and drives non-replications would let us potentially increase the replication rate, and also calibrate our levels of caution and skepticism toward results. Theoretical, interventional and correlational approaches do not yet provide comprehensive answers.

- Theoretical – we know people do bad practices and can model how this leads to lower replicability. Requires things we already have ideas about

Theoretical approaches modelling how questionable research practices (which many psychologists admit to) and a bias towards positive results would lead to a high rate of (non-replicable) false positives in the literature (CITE). There's some disconnect between these theoretically causal factors and empirical results about what causes or correlates with replicability.

- Interventional – try things they think will make it better (like not doing bad practices), see what happens. Not everything is intervention prone, need to have strong belief first. Expensive! A couple studies have attempted to intervene on the low replication rate, but seeing if certain factors can affect whether studies replicate. For instance, across 16 studies, Protzko et al. (2020) showed that better methodological practices, such as transparency, large sample sizes, and confirmatory tests, led to replication rates that matched theoretical expectations based on the effect sizes and sample sizes, and replication effect sizes comparable with the original. Many Labs 5 examined how adding expert advice to a replication process might intervene to improve the replicate rate, and found that it mostly didn't, at least on the 10 studies they looked at (Ebersole et al. 2020). These types of experiments are valuable, but they aren't very scalable because they are expensive and some potential influences on non-replication are not experimentally manipulable.
- Correlational – markets/etc say there is some predictive value to be had, but don't say if it's transparent and explainable. RP:P. Much of the work looking at predictors of replicability has been correlational. 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 many concrete predictors that differentiate replications from non-replications. Machine learning approaches CITE similarly show predictive signal, but do not provide explainable predictor variables. In the RP:P sample, studies in cognitive psychology (as opposed to social psychology) and studies with larger effect sizes and smaller p-values were more likely to replicate CITE . All of these correlational approaches depend on data from replications, generally drawing heavily from the same small set of large-scale replication data points. In particular, the RP:P dataset itself is much discussed and reanalyzed (Anderson et al. 2016, Etz & Vandekerckhove 2016, Gilbert et al. 2016, Patil et al. 2016), so much of what we think we know about replicability may be overfit to the 100 studies included in in RP:P.

  - Note that it's correlational, so can't do causality, but still useful for prediction, and as fodder for theory & intervention. (Interplay between them)

What's missing from the conversation and why our approach is good. * There's been a lot of discussion around replicabilty and it's role and yadda yadda (much overfocused on RP:P results because that's what's available). Growing sophistication and valid points, but we think it misses the mark about what practicing scientists are concerned about.

What does replicability mean? Prior approaches to replicability have focused on interpreting 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 for prior work typically do not reflect an entirely random sample from the literature; instead they sample from specific journals where results may be of more interest and adjust 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 truely 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). How much we care about whether a study is "trustworthy" is not uniformly distributed across the literature.

In contrast, we have explicitly pursued a different estimand: the probability that a researcher, on selecting a finding of interest from the literature, can successfully achieve a result satisfactorily close enough to the original that they can build on it in their own work, with all the necessary compromises to the methods and sample of the original that may be required by the constraints of the situation.

This framing of replicability matches our methodology.

Rather than sampling at random from some parts of the literature; 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.

We use a subjective replication score as our primary metric. 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). This avoids bright line distinctions and accomodates the range of outcome measures in our diverse set of studies.

All our replications were conducted under short time scales and relatively small budgets, which mimics the constraints many scientists are under when starting new projects. From the point of view of cumulative science, researchers want to know if they can get paradigms to work under their real-world constraints. The same issue that may cause a study not to replicate under constrained circumstances (despite hypothetically replicating under more favorable circumstances such as expert administration or larger budgets) will also plague attempts to build off those studies in constrained circumstances. Thus, we believe it is relevant to estimate replicability as done by a graduate student with limited resources.

Overall, we take a functional approach to assessing replicability, by framing both our methods and interpretation around the idea of whether work can be repeated or built on by an early-career scientist.

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 graduate-level experimental methods

class between 2011 and 2022 as individual course projects. We investigated predictors of replicability in this dataset and found that within-subjects designs and studies with large standardized effect sizes were positively correlated with replication success.

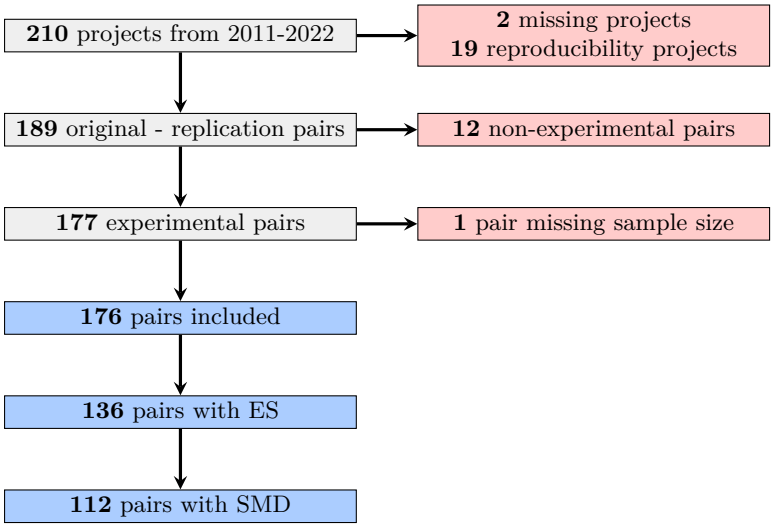

Figure 1: Which studies were excluded for what reasons, and how many original-replication pairs are left.

1 Results

PSYCH 251 is Stanford Psychology's graduate-level experimental methods class taught by MCF. During the 10 week class, students replicate a published finding. They individually re-implement the study, write analysis code, pre-register their study, collect data using an online platform, and write up a structured replication report. Students are free to choose studies related to their research interests, with the default recommendation being an article from a recent year of Psychological Science. While this choice results in a non-random sample from the literature, the sample 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, using different experimental methods and statistical outcomes. We leverage the naturally occurring variability in this sample of replications to examine how different demographic, experimental design, and statistical properties predict replication success.

Many different measures can be used to define replication success of an individual statistical result, and there is much discussion over what each measure represents and which make sense (Simonsohn 2015, Gelman 2018/ed, Mathur & VanderWeele 2020). Because we operationalized replicability as whether a study could be built upon, we used a subjective rating of replication success as our primary outcome measure. This measure has the benefit that it was applicable across the diverse range of statistical measures and reporting practices present in the sample. It, unlike statistical measures of replication, could easily accommodate studies where there were multiple important outcome measures that together defined the pattern of interest. A holistic measure of replication success had been coded for each project when it was turned in at the end of the class. For reliability, VB independently code the replication success from the replication reports; discrepancies were resolved by discussion between MCF and VB (26% of cases).

As a complement, we also used two statistical measure of replication on the subset of the data where they were computable (138 cases, see Figure 1). We measured p-original, the p-value on the null hypothesis that the original and replication statistics are from the same distribution, as a continuous variable, and we also determined whether the replication statistic fell within the prediction interval of the original statistic (Errington et al. 2021). These both measure how similar the estimates from the two sets of data are.

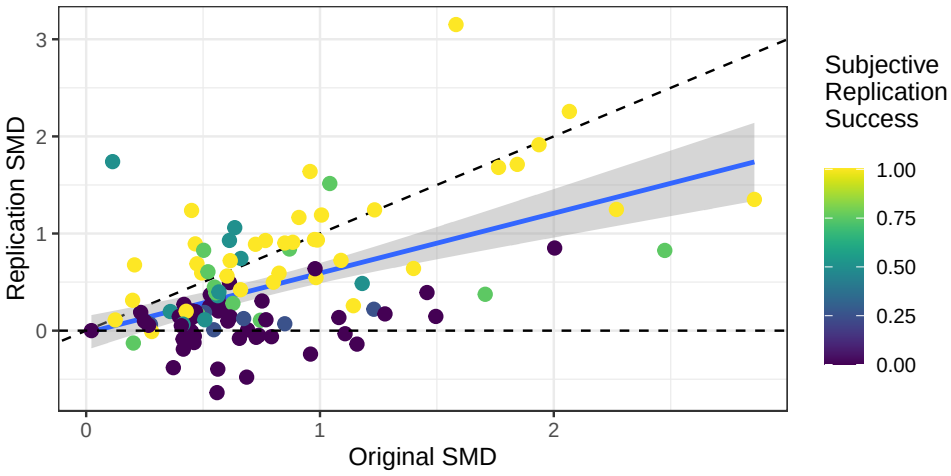

Figure 2: Relationship between SMD of the original study, SMD of the replication study, and subjective replication success rating, for those studies where SMD was applicable.

1.1 Overall replication rate

Across the 176 studies, the overall subjective replication rate was 49%. 45% (N=FOOBAR) of the studies had replication outcomes within the prediction interval of the original outcome. The median p_original value was 0.03. Figure 2 shows the relationship between original SMD, replication SMD, and subjective replication score. Roughly speaking, there's a cluster of studies that replicate with similar effect sizes to the original and another cluster that fail to replicate with replication effect sizes near zero. On average, there is a diminution of effect sizes from original to replication.

Table 1: The correlation of individual predictors with subjective replication outcomes. 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 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

1.2 Single predictors

Properties of both the original study and the replication can influence whether or not the replication is a success. We chose a set of predictor variables from the correlational results of RP:P and our own intuitions about experimental factors that might impact replication success as well as some covariates related to how close the replication would be. A full description of these features is given in methods.

Many of these predictors individually correlate with subjective replication success (Table 1). Predictors of higher replicability included within-subjects designs, higher numbers of trials, and open data. Predictors

of lower replicability included single vignetted studies, social psychology studies, and original-replication pairs where 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, and the cognitive psychology studies replicated at 2.45 times the rate of social psychology studies. Within and between subjects designs were both common, and within replicated 3.35 times as much. Similarly, studies with multiple vignettes replicated 2.62 times more than 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, although this may be linked to temporal trends.

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 (**crump2013?**)TODO MORE CITATIONS, 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).

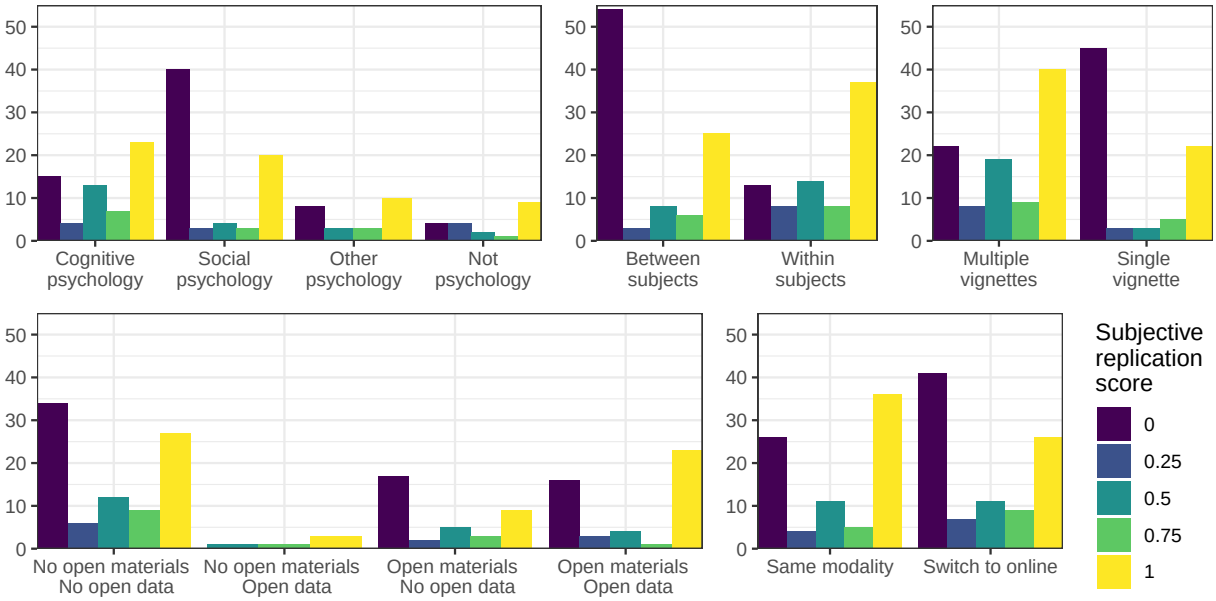

Figure 3: Distribution of subjective replication scores within categories. Bar heights are counts of studies.

1.3 Regression model

While a number of the predictors show individual correlations with the subjective replication score, many of the predictors are also correlated with one another. In order to determine which predictors were the strongest, we ran a series of pre-registered regularized regression models (see Methods for details; see Supplement for all estimates from all models). The coefficient estimates of the primary model, predicting the subjective replication scores based on all the data, are shown in Figure 4. Due to a large number of predictors coupled with a small and noisy dataset, even with strong regularization, there is much uncertainty around the coefficients. 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 without statistical predictors (0.55, CrI= [-0.01, 2]). When statistical predictors are added to the model, within-subjects designs remain predictive (0.64, CrI= [-0.03, 2.38]). Standardized effect size is another strong predictor of subjective replication score (0.59, CrI= [0.31, 2.58]). 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, 3.34]; effect size 0.97, CrI= [0.76, 4.59]).

We also ran models predicting whether the replication effect was within the prediction interval as 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 are wide, the general patterns of predictors are similar to the subjective replication models. The strongest predictors are still within-subjects designs (0.71, CrI= [-0.12, 2.38]) and studies with larger effect sizes (0.3, CrI= [-0.31, 0.89]).

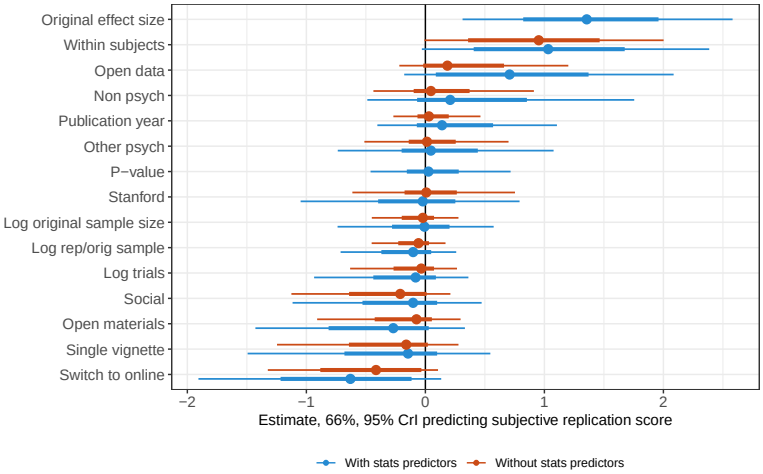

Figure 4: Coefficient estimates and uncertainty from a model predicting subjective replication scores from the full dataset.

2 Discussion

The replication rate in psychology is to the best estimates, somewhere around 50%, which is quite low. Many reasons for this low replication rate have been hypothesized, but studying it either experimentally or correlationally requires doing a number of replication experiments, which can be a lot of work. Here 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 replications, we found a FOOBAR replication rate, with some studies showing effect sizes similar to the original and others much smaller. Some individual correlates of replicability that stood out included within-subjects designs, work in the subfield of cognitive psychology, and the original and replication both using online samples. Many of these correlates interrelate with one another, and we are still limited in our sample, so models with multiple predictors have a lot of uncertainty around the effects of each predictor.

We are explicit in our goal of estimating how likely it is that a first-year graduate student, attempting to replicate a study of interest to them, gets replication results which are consistent enough with the original results that the student could build on the study in their own work. We can't estimate some platonic probability of truth for a study and neither can any replication study. Instead, we can estimate how likely studies are to get results of different levels of closeness in different circumstances. Much of the discussion around whether things should count as replication successes, or failures, or not replications at all is really about what the right thresholds are for closeness of results, and what the right circumstances are for a replication. Rather than try to find the one right answer to whether a direct replication was done in a certain right way, we think it's useful to explicitly label what the estimand really was: how likely it is to get the same results (under whatever metric) given circumstances like those used for the replication.

In our case, our estimand is about how well first-year graduate students will do at the replication, which takes into account the limited time, limited budget, and limited experience. We think this is an important estimand, as much of the work of psychology is done by trainees, in circumstances like these. Thus, if some studies have delicate results that require large samples and very exact methods to achieve, they

may not replicate under normal resources.

Other replication projects have other estimands: Camerer et al. (2018) seems to ask something closer to how likely studies are to replicate when preformed by an expert with a large budget, and Ebersole et al. (2020) asks how likely certain studies are to replicate when performed with extensive feedback from the original authors.

We do not interpret our results as saying that all non-replications were false positives (presumably some would be replicable under different implementations and budgets and others would not). There are many possible reasons for the non-replications in this sample. In some cases, it seemed that the problem may have been with the replication: for instance, if there were too few participants, or if there were high levels of wrong answers on attention checks, or participants speeding through without attention checks. For these cases, there was a clear next attempt that a student could make if they wanted to get the replication to work. In other cases, there might have been a priori reasons to distrust the original study results, such as exclusion criteria that seemed to be post-hoc, or a three-way interaction effect on a small sample (CITE THAT THIS IS SKETCH). In yet other cases, it's unclear why the results failed to replicate.

[somehow transition here] Pedagogy is important for open science. It's one thing to require or incentivize scientists to use open science practices and conduct replicable and reproducible research, but using the right tools and workflows to do open science is something that has to be learned. Teaching it in the classroom addresses the knowing how point at the beginning and shows students how to have open science practices integrated in to their science at the beginning, before other habits can ossify. Doing replications give students the motivation to care about open science, as they see how much easier it is to implement the study with open materials versus the study where they have to make guesses about the study instructions from the methods section. In presenting work with classmates, students see that there is variation in how well studies replicate, with some replicating very cleanly and others not at all. This sort of first hand experience teaches that not everything they read in the literature may just work if done again.

Our results are limited by the number and quality of the studies we included. These studies are not necessarily representative of the studies of interest to psychologists as a whole. The replication studies were not designed from the start with this analysis in mind, so analysis is limited by the choices and reporting used at the time of replications. TODO what are the limitations we want?

TODO quick wrap up

3 Methods

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

3.1 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 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 in the repo. TODO example and template

Students were free to choose what study they replicated; 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, 45.4545455% 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.).

3.2 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.

3.2.1 Measures of replication success

Our primary replication outcome is experimenter and instructor rated replication success (0-1). 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 (25.5681818% of studies). These 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 Errington et al. (2021). 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.

3.2.2 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).

3.2.3 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 analyses mixed studies were counted as within-subjects designs. In the analysis, we used a log-scale for number of subjects and numbers of trials.

3.2.4 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. As the predictor variable, we used whether the recruitment platform was changed between original and replication. This groups 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 is to account for potential non-independence of the replication (ex. if replicating their advisor's work, students may have access to extra information about methods).

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

3.2.5 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, ES, p_orig, and predInt. 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 and 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, SMD isn't 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, using 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 regressions, we used SMD and log p-value as predictors.

3.3 Modelling

Due to the monotonic missingness of the data, we had more predictor variables and outcome variables for some original-replication pairs than 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 more limited 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 SMD 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.

All models used a horseshoe prior in brms. 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.

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

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

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

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

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

Errington TM, Mathur M, Soderberg CK, Denis A, Perfito N, Iorns E, Nosek BA (2021) Investigating the replicability of preclinical cancer biology (R Pasqualini and E Franco, Eds.). *eLife* **10**:e71601. doi:10.7554/eLife.71601

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

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

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

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

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

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