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.

1 TODO list

- BETTER abstract
- many citations everywhere
- long list of things to gesture at in intro
- link to open stuffs
- ** verify code for ES ** with someone
- redo intro
- redo results
- redo discussion
- fill in todos

2 Text slush pile

[who knows where this paragraph should go] Replications are one way of approximating whether an effect in the target paper is true, and how likely results are to replicate in some sort of platonic ideal world. There is no one answer here; replication projects measure something else that could be treated as an approximation. However, what a replication is really measuring is how likely it to get that effect given some conditions. Which the right conditions are is a potential point of contention (see discussion around closeness and sample size, power etc). Here, we are explicit in what sample of replicators and replication conditions were are sampling, and thus what we can generalize to. We are estimating how likely replication is when done online by a graduate student, under constraints that are typical for graduate students (limited time, limited budget). As much of scientific process is in fact performed by graduate students, we think this in itself is an interesting question.

We do not interpret our results as saying that all non-replications were false positives (presumably some are replicable under other circumstances and others are not). Some of the factors we look at are more easily interpreted as being about the original study than others. We do not assign causal explanation to the predictors because there are multiple plausible interpretations: they could be correlates of QRPs in

*Corresponding author. Email: vboyce@stanford.edu

the original, they could be correlates of harder-to-detect or more fragile effects, they could be correlates of less-close replications, they could even be correlates of another stronger predictor.

[there is controversy here, but to even discuss it, we need empirics] Even those who argue that replication is not so essential still rely on data from replications to make their cases that replication rates are not as low as they seem and that these replication rates are acceptable. Lewandowsky & Oberauer (2020)

circle back to the idea that student replications (either as class or before extension into their own work) would be valuable

3 Citation slush pile

predict w/ ML Altmejd et al. (2019)

econ also has replications and issues Camerer et al. (2016)

moving goal posts around replication / interpreting results Nosek & Errington (2020)

some sort of modeling of replications Pawel & Held (2020)

Gilbert et al. (2016) Anderson et al. (2016) people argue a lot about how to interpret RPP

Schmidt (2009) early theory of replication as part of advancing knowledge

Simonsohn (2015) small telescopes

Gelman (2018/ed) Don't characterize replications as successes or failures.

Mathur & VanderWeele (2020) other metrics of replicability

pedagogy for replication in intro:

Frank & Saxe (2012) Hawkins et al. (n.d.)

O'Donnell et al. (2021) reps of some set of related things

Wagge et al. (2019) get UGs to do replications for

Quintana (2021) do reps as honors thesis Pownall et al. (2021) sips 2021 on teaching openness Jekel et al. (2020) teach open science w/ replication Jern (2018) in one study didn't show critical thinking benefits for students to make them do replications

4 introduction

Replicability is a desirable feature for any study that we build theory or future experiments off of. It's a pre-requisite for building theories that we have phenomena for these studies to explain, but results that cannot be observed consistently aren't results to build theories around. However, to be a load-bearing empirical result, a study needs to do much more than just replicate, it also needs to be interesting, have a believable theoretical interpretation, generalize appropriately, and have reliable constructs.

replication is important (but not everything) - good starting place * prereq of theory building that the results of interest "can be observed consistently" * but not everything (can still have other problems)

The replication rates in psychology are low, and replication rates are low in psychology (see studies)

big open question: why

experimental and correlational approaches to answering it

experimental is hard - protzko, ML5 etc. but they are expensive and you can't do them for many research questions

correlational has been pursued extensively, but is overfit to small data (prediction markets / ML models / regression of all types)

our contribution: provide more data! using pedagogical replications, which are good because they're p(build on) slightly different estimand but maybe a better one?

Psychology is in the middle of heated debate over its research practices and the subject of many reform efforts. Many of these issues concern how trustworth the published literature is (or isn't), how it got to be this way, and what measures can be taken to improve the quality of findings for the future. These issues center around issues of reproducibility and replicability. Concerns about replication rates have become a large point of discussion in psychology and other fields, with large scale replications spurring these discussions and providing the empirical data analyzed by all sides.

A major concern is that published literature may not replicate as much as previously thought or as desired. This has been illustrated by prominent findings/theories not replicating at all (ego depletion/ TMT?). To understand how replicable the literature as a whole is, we need empirical data about replication attempts across a wide range of studies. Most replications are targeted replications of individual results of interest to the researcher, but these are not sampled randomly from the literature, are performed in inconsistent ways, and may be selectively published depending on the results. Thus, they are not ideal for estimating the overall replicability of the field.

There have been a limited number of large-scale replication efforts, which sample studies from a discipline to estimate overall replicability. Due to the arduous process of replicating many studies, there have not been many large-scale replication efforts. Thus, all of the argumentation around predictors of replications and field-wide replication rates is fit to a small number of data points.

4.1 prior literature

We are aware of three large-scale replication efforts replicating experimental results in the existing psychology literature. RP:P sampled roughly 100 studies from top psychology journals in 2008 (Consortium 2015). They found an overall replication rate around 40%, which provided evidence to support growing concerns about the non-reliability of the literature. The RP:P dataset was itself much discussed and reanalyzed (Etz & Vandekerckhove 2016, Gilbert et al. 2016, Patil et al. 2016) [idk, maybe worth framing that this was a big deal study when it came out]. TODO some discussion of how many papers have reanalysed this cutting the data different ways.

The ManyLabs series of studies have also done large-scale replications of effects from psychology. Due to their primary goal of investigating different forms of hetereogeneity, their sampling has been non-representive, focusing on short studies with only two conditions. Across Many labs 1-3 foo bar replicated (Klein et al. 2014, Ebersole et al. 2016, Klein et al. 2018). Many labs 5 was a re-replication attempt on 10 of RP:P and rescued 2/10 (Ebersole et al. 2020). TODO look up details

Camerer et al. (2018) replicated the 21 behaviors studies published in Nature and Science from 2010-2015 that did not require special populations or special equipment. They found a roughly 60% replication rate.

4.2 predictors

In addition to determining estimated overall replicaiton rates for fields and journals, there's also merit in knowing what features of experiments (and replication attempts) are predictive of replication success. Blah blah stakeholders and resources. Scientists may want to build their work on studies that are likely to replicate, so that they can replicate and build on work without wasting resources. Similarly investing in interventions may want to start with things that are more likely to replicate. There's some signal here, as people are able to predict replication success at above chance CITATIONS (Dreber et al. 2015, Camerer et al. 2018, Forsell et al. 2019, Hoogeveen et al. 2019)

RP:P looked at how replication rates varied across subsamples of their studies (Consortium 2015). They found that cognitive psychology studies replicated at higher, but still low rates (50% v 25%) compared to social psychology. They also found that larger effect sizes and smaller p-values of original studies were predictive of replicating. TODO do we talk about any other correlates.

There are reasons to believe that experimental factors such as number of items or between and within subject designs may also be predictive (cite our old paper), and could pontentially be some of the reason for subfield differences. TODO maybe foreshadow that there are reasons to believe there are correlations!

While many replications of individual studies are conducted, these may have been selected for non-representative reasons, such as a high prior on replicating (e.g. as a demonstration), or a low or uncertain

prior (e.g. for a high-value study). When sampling is likely related to replicability, the studies are less good for estimating base rates and predictors of replicability in the literature as a whole.

4.3 current work

one strategy for the pedagogical bit at end of intro: stress "this is what people want to do, that is, pick up a project that they care about (not a random sample of the literature) and see if they can 'get it to work' " - and "subjective replication success is our primary measure for the same reason. . . ."

can even cite the "science is not a signal detection problem" paper on not randomly sampling from the literature.

Prior approaches to replicability have focused on 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 etc.) or even time period (ramscar).

Further, the methods for estimating this quantity have been theoretically problematic. Sampling schemes for prior work typically does not reflect a random sample from the literature – and even if they did, arguably a literature will succeed if useful discoveries come out of it, not if random findings are true ("Science is not signal detection" ref).

In contrast, we have 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 feel 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. SPELL OUT SAMPLING DIFFERENCE AND SUCCESS DIFFERENCE.

We estimate this quantity in the context of a particular pedagogical situation: that of a graduate student entering a doctoral program in psychology. . .

Prior literature, such as Frank & Saxe (2012) and Hawkins et al. (n.d.) have advocated using research methods classes to conduct replications, both for the education and scientific value.

Here we introduce 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. This approximately doubles the set of experiments in the large-scale replication literature. We investigate predictors of replicability in this new dataset.

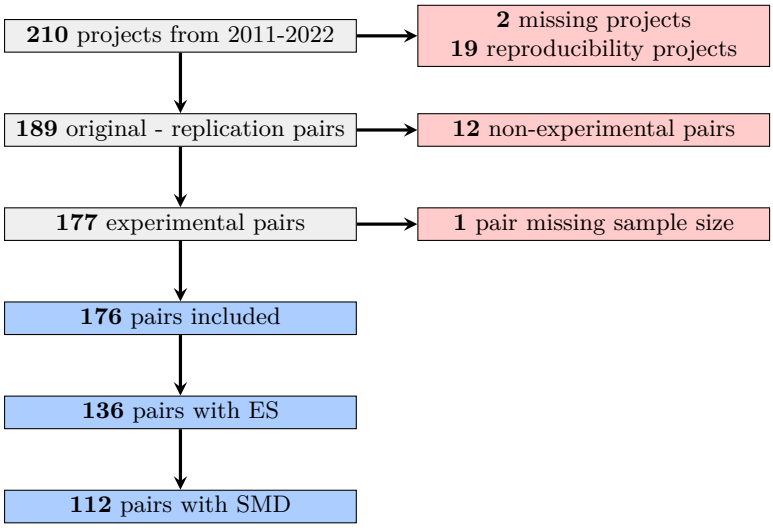

Figure 1: Which studies were excluded for what reasons, and how many original-replication pairs are left.

5 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.

TODO discuss more about why this one and what other measure we don't choose are

We used a subjective rating of replication success as our primary outcome measure. The instruction team had coded a holistic measure of replication success for each project when they were turned in at the end of the course. For reliability, VB independently code the replication success from the replication reports; discrepancies were resolved by discussion between MCF and VB (25.5681818 % of cases). This measure was applicable across the range of statistical measures and reporting practices and accommodated studies where there were multiple important outcome measures.

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).

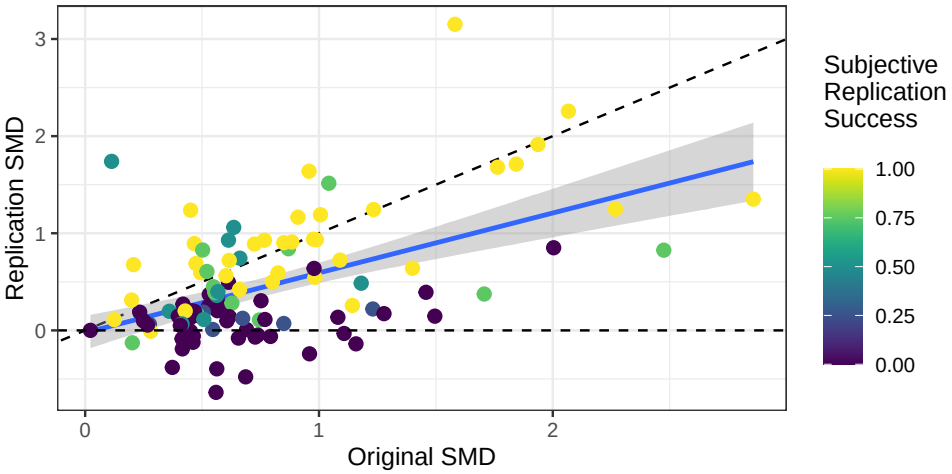

Figure 2: Relationship between SMD of the original study, SMD of the replication study, and subjective replication success rating, for those studies where SMD was applicable.

5.1 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

5.2 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 twice the rate of social psychology studies. Within and between subjects designs were both common, and within replicated four times as much. Similarly, studies with multiple vignettes replicated 1.5 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 TODO 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).

5.3 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 pre-registered regularized regression model (see Methods for details). The coefficient estimates 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 seem an especially strong indicator of replicability TODO include estimate and range in text!

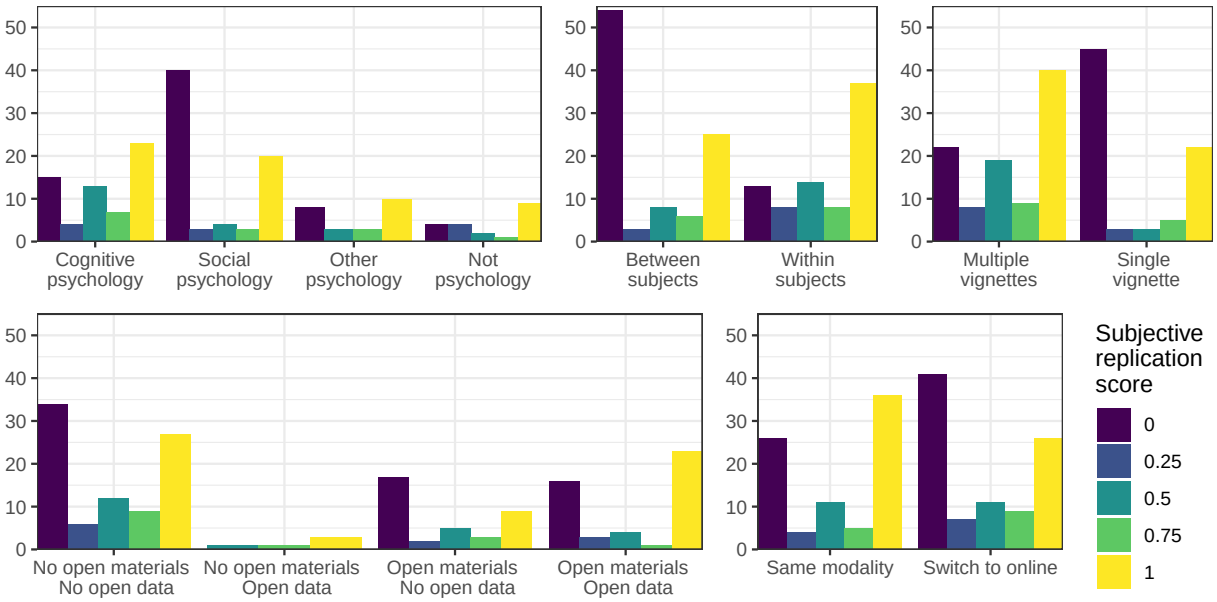

Figure 3: Distribution of subjective replication scores within categories. Bar heights are counts of studies.

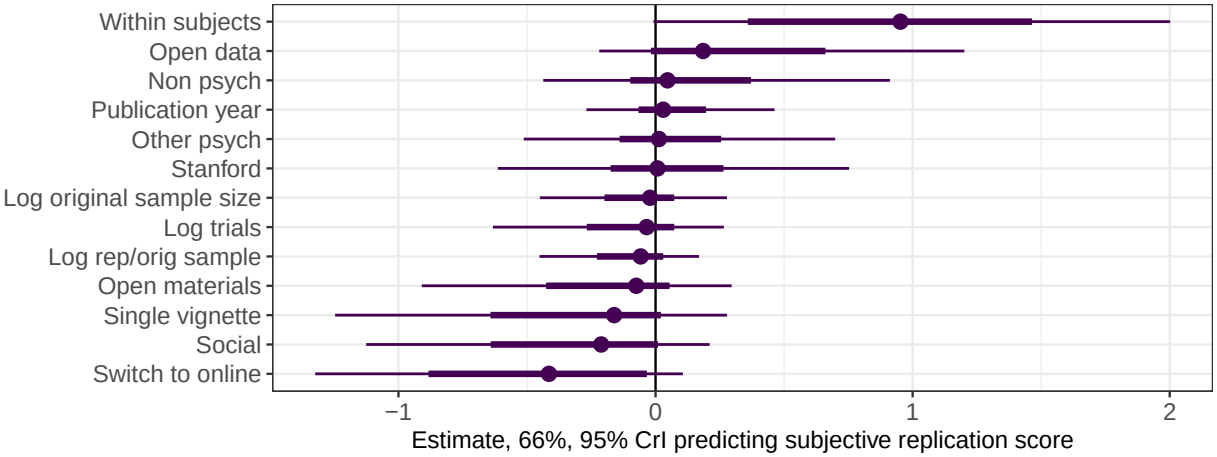

Figure 4: Coefficient estimates and uncertainty from a model predicting subjective replication scores from the full dataset.

Results from other models can be found in the supplement. TODO make this a true statement

6 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. TODO other possible things

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 (**ML5?**) TODOCITE 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 literature as a whole. TODO what are the limitations we want?

TODO quick wrap up

7 Methods

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

7.1 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 TODO FOOBAR %).

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

7.2 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.

7.2.1 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.

7.2.2 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).

7.2.3 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.

7.2.4 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.

7.2.5 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.

7.3 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

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

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

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

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

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

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

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

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

Jekel M, Fiedler S, Allstadt Torras R, Mischkowski D, Dorrough AR, Glöckner A (2020) How to Teach Open Science Principles in the Undergraduate Curriculum—The Hagen Cumulative Science Project. *Psychol Learn Teach* **19**:91–106. doi:10.1177/1475725719868149

Jern A (2018) A preliminary study of the educational benefits of conducting replications in the classroom. *Scholarsh Teach Learn Psychol* **4**:64–68. doi:10.1037/stl0000104

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

Lewandowsky S, Oberauer K (2020) Low replicability can support robust and efficient science. *Nat Commun* **11**:1–12. doi:10.1038/s41467-019-14203-0

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

Nosek BA, Errington TM (2020) The best time to argue about what a replication means? Before you do it. *Nature* **583**:518–520. doi:10.1038/d41586-020-02142-6

O'Donnell M, Dev AS, Antonoplis S, Baum SM, Benedetti AH, Brown ND, Carrillo B, Choi AL, Connor P, Donnelly K, Ellwood-Lowe ME, Foushee R, Jansen R, Jarvis SN, Lundell-Creagh R, Ocampo JM, Okafor GN, Azad ZR, Rosenblum M, Schatz D, Stein DH, Wang Y, Moore DA, Nelson LD (2021) Empirical audit and review and an assessment of evidentiary value in research on the psychological consequences of scarcity. *Proc Natl Acad Sci USA* **118**:e2103313118. doi:10.1073/pnas.2103313118

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

Pawel S, Held L (2020) Probabilistic forecasting of replication studies. *PLOS ONE* **15**:e0231416. doi:10.1371/journal.pone.0231416

Pownall M, Azevedo F, Aldoh A, Elsherif M, Vasilev M, Pennington CR, Robertson O, Tromp MV, Liu M, Makel MC, Tonge N, Moreau D, Horry R, Shaw J, Tzavella L, McGarrigle R, Talbot C,

Parsons S (2021) Embedding open and reproducible science into teaching: A bank of lesson plans and resources. *Scholarsh Teach Learn Psychol*:No Pagination Specified–No Pagination Specified. doi:10.1037/stl0000307

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

Schmidt S (2009) Shall we Really do it Again? The Powerful Concept of Replication is Neglected in the Social Sciences. *Rev Gen Psychol* **13**:90–100. doi:10.1037/a0015108

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

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**