PUT YOUR TITLE HERE

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

[1]Stanford University

Abstract

TODO abstract

Keywords
One keyword; Yet another keyword

1 TODO list

- title
- abstract
- need to do supplement with the stuff we preregistered (sensitivity analysis, all models)
- how to introduce coding methods (outcome measures)
- many citations everywhere
- long list of things to gesture at in intro
- link to open stuffs
- ** verify code for ES ** with someone

2 introduction

[replication: it's a hot issue and concerns around replication have done a lot, see crisis/revolution] Replication is a hot topic. Some argue that it is a cornerstone of a cumulative science, and findings of low replication rates are a problem. Concerns about replication rates have become a large point of discussion in psychology and other fields.

[there's . . . not a lot of empirics]To quantify and understand this problem, we need empirical data about replication attempts. However, due to the arduous nature of collecting samples of replications, there have not been many large-scale replication efforts, so all of the argumentation around predictors of replications and field-wide replication rates is fit to a small number of data points.

Many replications of individual effects have been performed, but these are less useful as pattern analysis because the varied sampling and reporting, and the fact they aren't all in one place. Also pub bias.

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. [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

[*] Corresponding author. Email: vboyce@stanford.edu

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.

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. There's some signal here, as people are able to predict replication success at above chance CITATIONS Hoogeveen et al. (2019). RP:P looked at how replication rates varied across subsamples of their studies. 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.

(could possibly frame with a bit of some=MCF have advocated teaching replication cite whatever, describe class, then thing? Frank & Saxe 2012, Hawkins et al. n.d.) Here we introduce a new dataset of replications in the behavioral sciences, primarily psychology. Over the years 2011-2022, students in a graduate-level experimental methods class have conducted online replications as individual course projects. From this, we have 176 experimental replications that were codeable. This approximately doubles the set of experiments in the large-scale replication literature. We investigate predictors of replicability in this new dataset.

2.1 quick methods

[some of this should actually just go in methods methods] Class, taught by last author, is an experimental methods class aimed at first-year graduate students, but taken by . . . Since around 2015, it has been required for incoming PhD students in the Psychology Department. Over the course of a quarter long class, students work through their individual replication projects, from choosing a study, to reimplementing it, to creating analysis code, piloting the study, pre-registering it, and finally running a full data collection and writing up their results. Students are free to choose studies related to their interests, although recent articles from Psychological Science are the recommended path under uncertainty (leading to a high proportion of Psych science articles in this replication sample). While the sampling procedure is non-random, it is generally representative of studies that are of interest to and doable by first year grad students.

Studies vary along a number of dimensions, including statistical properties, subfield, and experimental properties. We leverage this naturally occurring variation to see if these properties predict replication success.

3 Results

[again unclear the intro/ methods/results split] Over the years, there have been a number of student projects, of these we are able to include 176 (see Figure 1 for exclusions). Because of variance in the experimental designs and reporting practices, we have variable statistical information. To address this, we conducted a series of models, balancing between including as many data points as possible, and including potentially relevant outcome and predictor variables. Here we focus on one set of models, the prediction of subjective replication outcome on the basis of the whole dataset without the statistical predictors. Other model results, including those from a sensitivity analysis, are available in the supplement.

As our primary outcome measure, we use a subjective measure of replication success. At the end of the class, the instruction team coded each project on a 0-1 scale indicating whether or not it replicated. This allows for nuance around marginal effects, covers studies with a broad range of statistical outcomes (including those that do not lend themselves to NHST), and includes studies with multiple important outcome measures. For robustness, first author independently coded all projects on the same scale off of the final write-ups. Disagreements (which occured foobar % of the time) were resolved by discussion between first and third authors.

As secondary outcomes, we also calculated prediction intervals following Errington et al. (2021) on the key measure of interest from each study. Because of variable statistics and reporting, this was calculable on FOOBAR studies.

3.1 Overall replication rate

Across the 176 studies, the overall replication rate was 49% TODO derive in code. FOOBAR % of the studies had replication outcomes within the prediction interval of the original outcome. The average p_original value (median?) was TODO.

For the FOOBAR studies with SMD, we see cite FIGURE.

3.2 Single predictors

We drew 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 co-variates related to how close the replication would be. Many of these predictors individually correlate with subjective replication success (Figure 1. In particular, within-subjects designs, higher numbers of trials, and open data were predictive of replication. Single vignetted studies, social psychology studies, and original-replication pairs where the replication switched to online were less likely to replicate. Distributions of study outcomes across some of these properties are shown in Figure TODO.

Both social and cognitive psychology studies were well represented, and the cognitive psychology studies replicated at a FOOBAR higher rate. Within and between subjects designs were both common, and within replicated at YY more. Similarly, studies with multiple vignettes replicated FOOBAR more. However, there are strong correlations between these experimental features and between experimental features and subfield.

TODO say somthing about openness

The study sample is split between studies that kept the same modality as the original (generally meaning that the original study used an online crowd-sourced sample) and those that switched. While TODO cite online studies are respectable, this 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).

3.3 Regression model

We were interested in determining what the strongest predictors were, which ones are most meaningful, so we threw it all into a big ol regularlized regression model. Despite this, wide uncertainty due to parameterization and stuff. The coefficients see Figure 3 . . . TODO

4 discussion

[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. It contrasts with questions like how likely

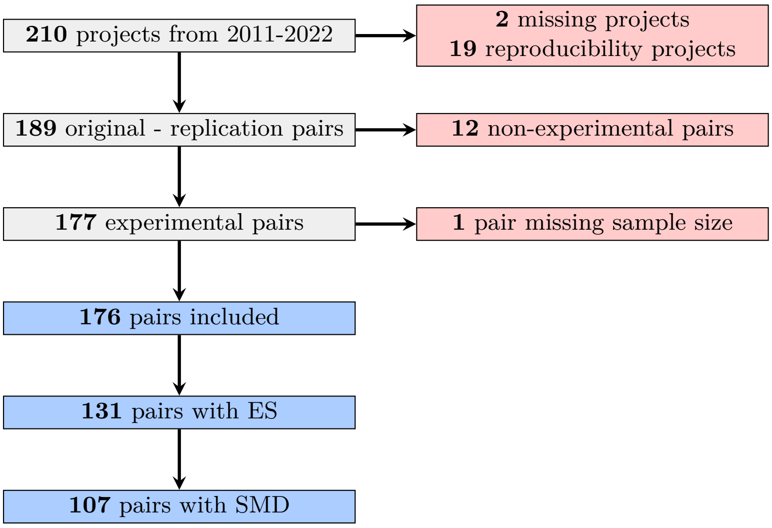
210 projects from 2011-2022
2 missing projects
19 reproducibility projects
189 original - replication pairs
12 non-experimental pairs
177 experimental pairs
1 pair missing sample size
176 pairs included
131 pairs with ES
107 pairs with SMD

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

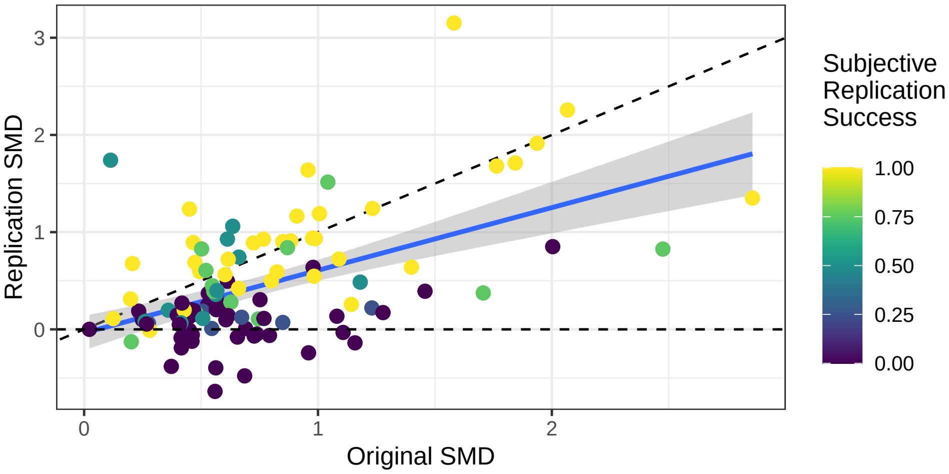

Figure 2: Relationship between SMD of the original study, SMD of the replication study, and subjective replication success rating, for those studies where SMD was applicable.

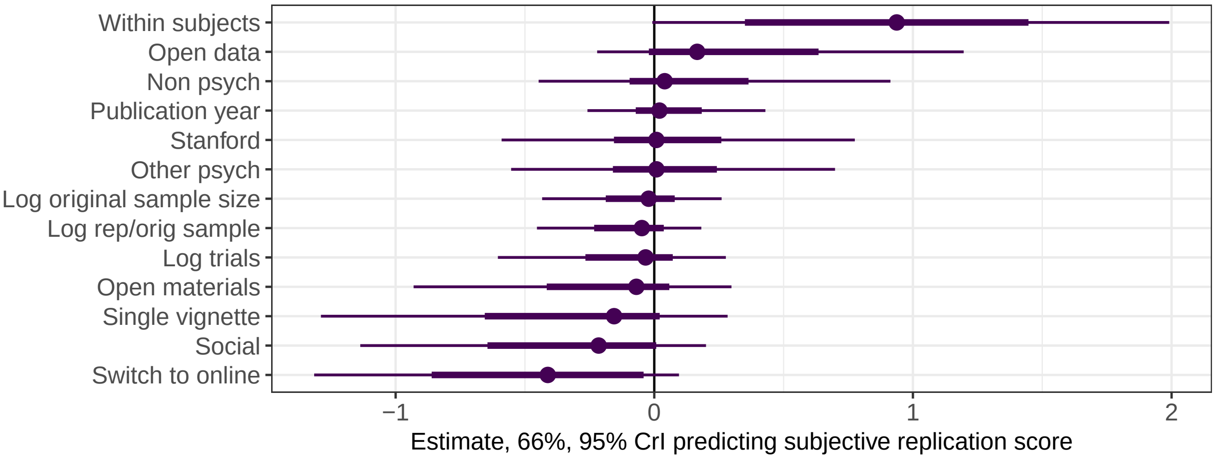

Figure 3: Coefficient estimates and uncertainty from a model predicting subjective replication scores from the full dataset.

Table 1: The correlation of individual predictors with subjective replication outcomes. For subfield, cognitive psychology is treated as the baseline condition.

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

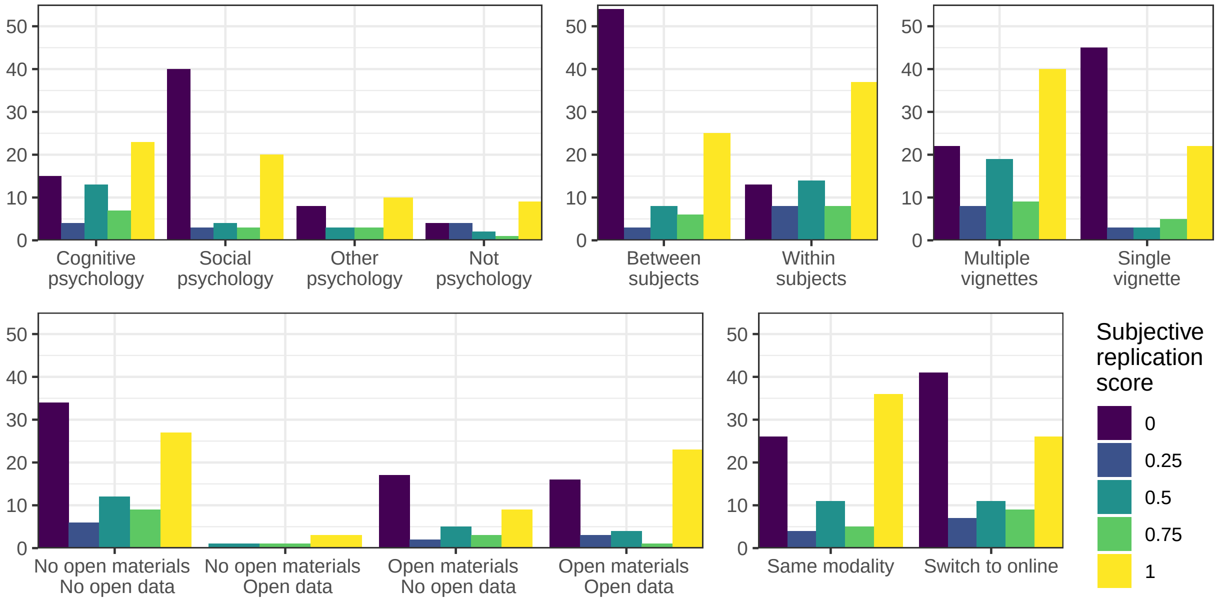

Figure 4: Distribution of subjective replication scores within categories. Bar heights are counts of studies.

something is to replicate when performed by an expert with a large budget (perhaps this is the right framing on Camerer) or when performed with extensive feedback from the original authors, etc.

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 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)

5 Methods

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

5.1 Dataset

The dataset of replication projects comes from class projects conducted in PSYCH 251 (earlier called PSYCH 24) a graduate-level experimental methods class taught by MCF from 2011 to 2022. 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

5.2 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.

5.2.1 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. 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.

5.2.2 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).

5.2.3 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.

5.2.4 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.

5.2.5 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.

5.3 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 will include random slopes for predictors nested within years of the class (year) 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 of more models in supplement. TODO

Acknowledgements

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

References

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*

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

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

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

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

Appendix A

Some appendix text.

Appendix B

More appendix text.