Estimating the replicability of psychology experiments after an initial failure to replicate

Veronica Boyce[1,*], Ben Prystawski[1], Adani Abutto[1], Emily Chen[1], Ziwen Chen[2], Howard Chiu[3], Irmak Ergin[1], Anmol Gupta[1], Chuqi Hu[4], Bendix Kemmann[5], Nastasia Klevak[1], Verity Y. Q. Lua[1], Mateus M. Mazzaferro[3], Khaing Mon[4], Dan Ogunbamowo[1], Alexander Pereira[5], Jordan Troutman[6], Sarah Tung[1], Raphael Uricher[1], Michael C. Frank[1]

[1]Department of Psychology, Stanford University
[2]Graduate School of Business, Stanford University
[3]Graduate School of Education, Stanford University
[4]Symbolic Systems Program, Stanford University
[5]Department of Philosophy, Stanford University
[6]Department of Computer Science, Stanford University

Abstract

When a replication fails, scientists have to decide whether to make a second attempt or move on. Psychology researchers who attempt to replicate studies before building on them often face this decision, given the relatively low rates of replication success in psychology. Here, we report on 17 re-replications of experiments with failed replications reported in Boyce et al. (2023). In 5/17 of these "rescue" projects (29%), the re-replication mostly or fully replicated the original results, albeit with smaller effect sizes. While a few replications were rescued with larger sample sizes and minor methodological changes, most re-replications failed again. In many situations, it may be most efficient to stop pursuing an effect after a single failed replication.

TODO FIX CITATION / BIB INFO (re OSC for RP:P but only after we unlink from zotero)

TODO compile everything into OSF repo & link TODO link individual pre-reg's in repo TODO link overall pre-reg TODO include in repo & link the reasons for replication failure that were offered to students TODO redo/double check closeness scores for first rep and rescues TODO add something about Verity's to case-study? TODO author contributions, acknowledgements

1 Introduction

Imagine the scenario where you are a new graduate student, and you run a replication of a study in the literature that you are interested in building upon in your research. The replication fails: perhaps the interaction you hoped for is directionally correct, but the point estimate is small and the confidence interval definitely overlaps 0, with a p-value of .3. Or perhaps the interaction is numerically in the wrong direction, and also the main effects look different. Whatever the details of the replication failure, you are left with a question: Should you *try again* and run a re-replication, or should you *give up* and pick a different study to replicate and build on?

In psychology, large scale replication projects have found that around half the studies successfully replicate. Across 100 studies with positive results, RP:P replicated 36 - 47% depending on the metric for replication success (Consortium, n.d.). With large multi-site samples, Many Labs 1 replicated 11/13 effects (84.6%), Many Labs 2 replicated 14/28 effects (50%), and Many Labs 3 replicated 3/10 effects (30%) (Ebersole et al., 2016; Klein et al., 2014; Klein et al., 2018). Camerer et al. (2018) included 21 behavioral social science studies and replicated 12-14 of them (57%-67%) depending on the metric used. Boyce et al. (2023) reported an average replication score of 49% for 176 replications primary in psychology. Psychology is not

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

the only discipline where many studies do not replicate. Large scale replication projects in other disciplines have found 39/97 of studies with positive effects (40%) replicating in cancer biology (Errington et al., 2021), 11/18 studies (61%) replicating in economics (Camerer et al., 2016), and 31/40 studies (78%) replicating in experimental philosophy (Cova et al., 2021). Across these disciplines, and especially in psychology, large-scale replications indicate a substantial chance of scientists encountering failed replications if they try to replicate published findings.

Replications can fail for many reasons. In principle, any difference between the original study and replication could be a root cause of the difference in their outcomes. In practice, the relationship is not symmetric, and many causes of replication failure can be attributed more to the original study or more to the replication. One potential reason for replication failure is that the original result is fragile in some way: it could be a statistically unlikely result that was achieved through chance or p-hacking, or it could be sensitive to the exact conditions and time it was run under ("hidden moderators"). Another possibility is that the reported effect size might be inflated due to some combination of heterogeneity, p-hacking, low power, and publication bias. One consequence of inflated effect sizes is that replication studies might power for the reported effect size and then have too small a sample to reliably detect the (true) smaller effect. Replication studies might also end up underpowered due to unexpected attrition or noise in the replication sample.

Replications could also get different results from an original study due to differences in how the question of interest was instantiated. These differences in design could be intentional adaptations, unavoidable changes from lack of access to the original instructions or materials, or changes that were not expected to be critical. A difference in experiment platforms (ex. online versus in-person) might mean that an implementation no longer is appropriate, or the data quality controls and attention checks are inadequate. If there are differences in time, place, or subject population between the two studies, corresponding changes to the materials, instructions, or procedure may be needed to provide a good *translation* – too many changes could cause differing results, but so could not enough changes.

When replications fail, we generally do not know why they failed, although we may speculate. For the metascientific goals of most large-scale replication projects, it doesn't matter why any particular study failed to replicate, because the aim is to estimate the true proportion of effects in a literature.

In contrast, when individual scientists attempt to replicate an effect with the goal of building on it, they do care about why an individual replication failed. In an uncertain literature, scientists might not want to commit resources to building on an effect without first checking that they, with their resources and participant sample, can get the core effect. In this case, it's very important whether the replication failure is due to something fixable about the replication or something not-fixable about the original result.

While one can theorize after the fact about potential causes of replication failure, in most circumstances it is hard to diagnose for certain what made for the difference in the results. An exception might be if a specific error in one study or the other is found, such that the results were not what was thought (analytic error) or the data was not what was thought (experimental error). Aside from these rare cases, we usually can't know whether the replication failure is fixable just by looking at the replication failure itself.

Instead, re-replication could serve as a way of triangulating on reasons for re-replication failure, as well as another chance for a scientist to get a working protocol that picks up the effect of interest. Thus, scientists may want to know what the chances are that a re-replication succeeds at detecting an effect that indicates future experiments with this protocol are worth pursuing.

Re-replication has not been studied much, so it is uncertain how often re-replications recover the results of an original study. That is, it is not well known how definitive a first replication failure is.

One estimate of re-replication success rates comes from Ebersole et al. (2020) which re-replicated 10 psychology studies that had failed replications reported in RP:P (Consortium, n.d.). After Consortium (n.d.) published their results with a replication rate around 40%, Gilbert et al. (2016) raised concerns that methodological differences were the reason for the low replication rate. In 11 of the RP:P studies, the original authors had raised concerns about protocol fidelity prior to data-collection. Of these 11 with concerns, 10 failed to obtain significant results in the RP:P replication (Consortium, n.d.). In a follow-up, Ebersole et al. (2020) re-replicated 10 of these 11 in larger samples, using both the RP:P protocol and a new protocol revised under advice of the original authors or other experts, thus testing whether larger sample sizes and/or "better" methodologies would "rescue" the failed RP:P replications.

The result was that 0 of the RP:P protocols found significant results, and 2 of the 10 revised protocols did (but not the 1 that had a significant result in the original RP:P sample). Even when a significant effect was recovered, the effect size was much diminished (Ebersole et al., 2020). We could frame Ebersole et al. (2020) as an attempt to recover the original effects of the 9 studies with significant original effects and non-significant RP:P effects by a combination of high power and design tweaks. This suggests a successful re-replication rate of 2/9 (22%) for studies with non-significant first replications, under fairly favorable, high resource circumstances.

While it is tempting to speculate on the reasons for failed replications, the Gilbert et al. (2016) and Ebersole et al. (2020) example shows these speculations on causes of replication failure do not always pan out in re-replication outcomes.

While Ebersole et al. (2020) looked at re-replication success rate in a high resource setting with access to large, multi-site samples and expert input, most early career scientists do not have those kinds of resources to devote to re-replication attempts. On the other hand, early career researchers may also be prone to a different distribution of causes of initial replication failure. Thus, it is an open question what the re-replication rate is for failed replications of early career researchers, and estimating this rate could inform decisions on how to act after failed replications.

Here, we investigate re-replication in the context of graduate student replication class projects. Students in a graduate methods class on experimental methods were learning about best practices for experimental research and had the option to "rescue" an experiment that a student in a prior year had previously failed to replicate as their class project (first replications reported in Boyce et al. (2023)). Thus, students looked at both the original paper and the failed replication through the lens of best practices to see what potential issues in the replication they could fix. In the present paper, we report the results of 17 re-replication "rescue" projects which each re-replicated a study which had a failed or only partially-successful replication in Boyce et al. (2023). We find that a minority of projects (5/17, 29%) successful recovered the original effect.

2 Methods

PSYCH 251 is a graduate-level experimental methods class taught by MCF. In previous years, students have conducted replication projects, as reported in Boyce et al. (2023). In Fall 2023, students in PSYCH 251 were offered the option to do a "rescue" project where they re-replicated one of the unsuccessful replications from a previous year. Students could also opt to do a normal replication instead. We report on the result of 17 rescue projects that opted to be part of the paper and completed data collection.

A spreadsheet of projects, individual project write-ups (both first replications and rescues), links to individual project data and analyses for rescue projects, and the analytic code for this paper are all available at TODO OSF LINK GOES HERE.

Our analysis plan was pre-registered after students had selected projects, but before final data collection on the projects. Each project was also individually pre-registered by the student conducting it. The overall analysis is at LINK, individual pre-registrations are linked from LINK. We note one deviation from our pre-registration here: we pre-registered visual comparisons between original, first replication, and rescue projects using prediction intervals. Prediction intervals depend on both the original effect size and variance and the variance of the comparison (replication or rescue) study (Patil et al., 2016). Thus we cannot show *the* prediction interval for the original study, but would have to show a prediction interval between each pair of studies, which we thought would not offer clarity.

2.1 Sample

The experiments that were re-replicated were chosen from studies that failed to replicate in Boyce et al. (2023). We created an initial list of 49 rescue-eligible studies that had received a subjective replication success score of 0, .25 or .5 (on a 0-1 scale) in Boyce et al. (2023), where the replication had a Github repository available (github repositories were used starting in academic year 2015-2016), and where the original experiment had 200 or fewer participants (to ensure we could afford to increase the sample size). We then contacted the replication project authors for permission to share their report and repository with a new student and include it as a supplement on a resulting paper. This left 27 options for the students to choose from. 20 students chose to do rescue projects, and 3 students took an incomplete or did not indicate interest in being part of this paper, leaving a final sample of 17 rescue projects.

2.2 Procedure

Students conducted their rescue projects over the course of the 10-week class. Once they had chosen a project we gave them access to the original replicators' write-up and repository, which often included the data, experiment code, and analytic code. In many cases, students were also given the contact information of the original replicator (a few original replicators opted not to be contacted by students).

Students were required to think of reasons the first replication might not have worked, and address them if they could. A list of possible reasons and solutions TODO LINK was given to students. In general, we encouraged students to add manipulation checks as appropriate, and better adapt materials to online studies. For instance, the rescue of Paxton et al. (2012) switched from using the CRT which many online participants are overfamiliar with for the newer CRT-2 which is less overused but gets at the same construct (Thomson & Oppenheimer, 2016). The rescue of Jara-Ettinger et al. (2022) discovered that the replication had accidentally used the drawn version of the stimuli rather than the photographic stimuli used in the original and reverted to the photographic stimuli. Tarampi et al. (2016) had participants do a timed navigation task where they wrote L or R to indicate which way to turn at each intersection on a paper map. The replication ported this experiment online and had participants click a drop down to select left or write for each turn. The rescue switched for the unwieldy dropdown menus to having participants press keys on the keyboard to indicate the direction of a turn, which seemed like a more natural interface. Once students experimental designs and analytic plans were approved by TAs (VB and BP), students pre-registered and ran their samples.

With one exception, samples were collected on Prolific (the rescue of Yeshurun & Levy (2003) ran in-person on the Stanford student subject pool). We tried to power studies adequately (with a target of 2.5x original following Simonsohn (2015)), but due to cost constraints, not all studies were powered at this level. The rescues had on average 1.48 times the original sample post-exclusion (median: 1.07, IQR: 0.94 - 2.4, minimum: 0.48, maximum: 2.96, see Table 3 for all sample sizes). Across the 16 Prolific studies, we spent $5471, for an average of $342 per project.

2.3 Coding of results

We followed Boyce et al. (2023) in the properties of the studies we measure and how we quantify replication success.

Each project was rated on the basis of subjective replication success both by MCF and by one of VB and BP. Disagreements were resolved through discussion. As a complement to the subjective rating of overall success, we statistically compared one key measure of interest for each study, following Boyce et al. (2023). In order to statistically compare the key measures, we needed effect sizes reported in the same way for each original study, first replication, and rescue. When effects were not reported in consistent ways across original and replications, we recalculated effects from raw data when necessary to obtain comparable values.

We recorded the same set of potential correlates that were used in Boyce et al. (2023) for original, first replication, and rescue (these were already rated for original and first replication). These potential correlates included the subfield of the study (cognitive, social, or other psychology), its publication year, experimental design features including whether it was a within- or between- subject design, whether each condition was instantiated with one vignette or multiple, how many items each participant saw, and whether there were open materials and open data.

For the original study and each replication, we recorded the number of participants post-exclusions. For studies where some extra conditions were dropped in some replications, we counted only the participants in the key conditions all replications had for comparability. For instance, if an original study compared between two critical conditions but also had a baseline control, we did not count the participants in the baseline condition if a replication did not include this condition. We also recorded whether each study was conducted online using crowd-sourced platform or not.

3 Results

Our primary question of interest was how many of the 17 rescue projects succeeded at replicating the results in the original study. When a replication fails to obtain the same results, one may have intuitions about what may have gone wrong – these rescue projects test how often addressing these potential issues in fact works.

3.1 Overall replication rate

All rescue projects were rated holistically for how well they replicated the original results. We thought about replication in terms of how confident one would be to build on the line of work given the replication results, rather than focusing on any singular numeric result or significance cut-off. All projects were rated both by the instructor (MCF) and by one of the TAs (VB or BP); the interrater reliability was 0.9.

Across the 17 rescue projects, 5 mostly or fully replicated the original results according to the subjective replication ratings. 12 had a rating of 0, 2 got a rating of .75, and 3 got a rating of 1. Thus, a first pass answer to the question “how often can a failed replication be salvaged?” is 29% (bootstrapped 95% CI: 12% - 53%) of the time.

In the original replication sample from Boyce et al. (2023), 76 out of 176 replications (43%) mostly or fully replicated (i.e. received a subjective replication score of .75 or 1). Note that Boyce et al. (2023) report the average replication score as a percent success (49%), but given that we considered studies with a subjective score of .5 as eligible to be rescued, we recomputed the success rate where scores of 0, .25, and .5 are considered failures and .75 and 1 successes. If the re-replication rate in our sample is representative of the re-replication rate for the initially non-replicating studies, then the combined chance of mostly or fully replicating in a first replication or one follow-up replication is 60%.

3.2 Effect sizes

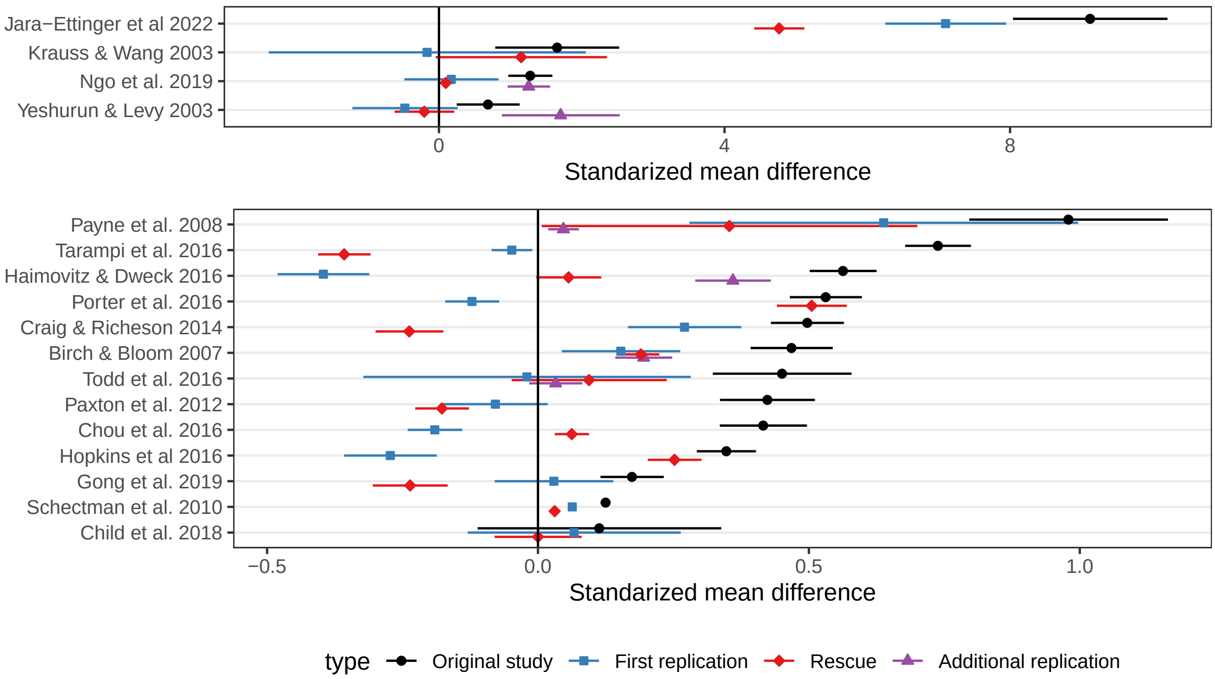

Figure 1: Standardized effect sizes of original studies, first replications, rescues, and additional replications if available. Due to the large effect size of a couple studies, large effect studies are shown in a separate panel. In a few cases, the first replication’s key effect was non-zero in the same direction as the original; however, in these cases the larger pattern of results was not fully consistent between original and first replication.

As a complement to the holistic, qualitative replication ratings used above, we also statistically compared the effect sizes of the rescue, first replication, and original study on one key measure per study. We followed Boyce et al. (2023) in determining a key measure for each study. When we were aware of additional direct replications (either from other class projects, or external replications in the literature), we also consider the effect sizes obtained in these additional replications.

We standardized all effect sizes into standardized mean difference (SMD) units. One potential issue with

Table 1: P-original values between different sets of experiments. The primary analysis is between the original result and the meta-analytic aggregation of all replications. All p-originals assume an imputed heterogeneity value of tau=.21.

	P-original comparing between			
	Original and			Rescue and
Paper	All reps	Rescue	Non-rescue	Other reps
Birch & Bloom 2007	0.192	0.194	0.191	0.989
Child et al. 2018	0.669	0.641	0.858	0.778
Chou et al. 2016	0.054	0.100	0.005	0.233
Craig & Richeson 2014	0.145	0.001	0.302	0.020
Gong et al. 2019	0.262	0.057	0.511	0.228
Haimovitz & Dweck 2016	0.069	0.018	0.180	0.864
Hopkins et al 2016	0.290	0.654	0.004	0.015
Jara-Ettinger et al 2022	0.014	0.000	0.006	0.000
Krauss & Wang 2003	0.271	0.519	0.139	0.311
Ngo et al. 2019	0.106	0.000	0.385	0.257
Paxton et al. 2012	0.011	0.005	0.023	0.649
Payne et al. 2008	0.021	0.031	0.075	0.923
Porter et al. 2016	0.370	0.905	0.002	0.003
Schectman et al. 2010	0.712	0.654	0.770	0.877
Tarampi et al. 2016	0.000	0.000	0.000	0.145
Todd et al. 2016	0.062	0.124	0.058	0.778
Yeshurun & Levy 2003	0.614	0.017	0.943	0.474

comparisons using SMD is that noisier measures will have smaller standardized effect sizes even if the effect on the original scale is the same. In general, the replication and rescue effect sizes were smaller than the original effect sizes, and in a few cases the effects were in the opposite direction (Figure 1).

Scientists' intuitions about whether a replication is successful and whether an effect provides support for a hypothesis (including heuristic cutoffs like p<.05) do not always align with measures of statistical consistency (Patil et al., 2016). For instance, two studies may both find that condition 1 results in a significantly higher outcome measure than condition 2, but the effect magnitudes may be sufficiently different that they are statistically unlikely to have come from the same population. On the other hand, one study may find a statistically significant, but imprecisely estimated effect in a small sample, and second study may find a near-zero (null) effect, but the effect estimates from the two studies may be statistically compatible, despite one supporting a hypothesized difference and the other not.

Here, we measure statistical consistency using *p*-original, the *p*-value on the null hypothesis that one study's population effect comes from the same distribution of population effects as another study or studies (Mathur & VanderWeele, 2020) If we assume no heterogeneity between studies, then this is the same as asking how consistent one study's estimate of the population-level effect is with another study's estimate of the population-level effect. However, meta-analyses of closely related studies, such as those from different sites of multi-site replications, indicate that there is often heterogeneity between studies. We cannot accurately estimate the level of heterogeneity for an effect meta-analytically as we have only a small number of replications per study. Instead, we impute a heterogeneity value of $\tau = .21$ SMD, which is the average level of heterogeneity found by Olsson-Collentine et al. (2020) in prior multi-site replications in psychology.

As our primary comparison, we compare the effect size of the original study to the meta-analytic effect of the totality of the replications (first replication, rescue, and additional if found). The value of *p*-original thus represents how likely an effect size equal to or more extreme than the original effect size is to occur if it comes from a distribution of population effect sizes defined by the mean and standard error of the replications (combined meta-analytically) with a between-population standard deviation of $\tau = .21$. That is, it *p*-original is a measure of the statistical consistency of the original result with the totality of the replications, assuming a given level of heterogeneity of effects between studies. Smaller *p*-original values indicate more inconsistency.

Table 2: Correlations between an individual predictor and the subjective replication score of the rescue project. The first set of predictors were pre-registered based on the correlates used in Boyce et al. (2023). The last three predictors were added post-hoc.

Predictors	r	p
Social	0.133	0.612
Other psych	-0.295	0.250
Within subjects	0.210	0.418
Single vignette	-0.030	0.910
Switch to online	0.175	0.501
Open data	0.231	0.373
Open materials	0.471	0.057
Stanford	-0.233	0.369
Log trials	0.005	0.985
Log original sample size	0.027	0.919
Log rescue/original sample size	0.024	0.927
Log replication sample size	-0.258	0.317
Log replication/original sample size	-0.487	0.048
Log rescue/replication sample size	0.490	0.046

The median value of p-original between an original study and its replications was 0.15 [IQR: 0.05 - 0.29]. 24% of the p-original values were less than .05, indicating by conventional thresholds a rejection of the null hypothesis that the original comes from the same distribution of population effects as the replications. Individual p-original values for each study are shown in Table 1.

As secondary measures, we calculated the p-original values between a) the original and the rescue, b) the original and non-rescue replications, and c) the rescue and other replications. For a) the original versus the rescue, median value of p-original was 0.06 [IQR: 0.01 - 0.52], and 47% of the p-original values were less than .05. For b) the original versus all the non-rescue replications, median value of p-original was 0.14 [IQR: 0.01 - 0.38], and 35% of the p-original values were less than .05. For c) the rescue versus the other replications, median value of p-original was 0.31 [IQR: 0.14 - 0.78], and 24% of the p-original values were less than .05.

Overall, allowing for a heterogeneity level of τ=.21 SMD, a number of original effects are not statistically consistent with rescues and replication effects. The pattern of inconsistency does not align with which studies were rated as having replicated. In all cases, the point estimate of the re-replication is smaller than, or in the opposite direction of, the original effect on the key measure of interest.

3.3 Correlates of rescue success

Are there signals in the features of a study or it's replication that predict whether a re-replication will succeed? To address this question, we ran correlations between the set of predictor variables used in Boyce et al. (2023) and the subjective replication scores of the rescues. We also added some (non-preregistered) predictors related to the sample size of the first replication, after seeing the successful re-replications of Ngo et al. (2019) and Krauss & Wang (2003), both of which had small replication samples.

The number of rescues is small, and many of these predictors are correlated, so we caution against over interpretation. All of the correlations are presented in Figure 2. The strongest correlates of rescue success were open materials, a small sample size on the first replication, a small sample size on the first replication relative to the original sample size, and a large rescue sample size relative to the first replication. None of the pre-registered correlates meet the conventional significance threshold; the two correlates based on ratios that reflect a relatively smaller replication sample size are marginally significant.

Small replication samples relative to original and rescue could be due to both a) powering a replication according to a reported large effect size or b) difficulties with recruitment or high exclusion rates leading to a smaller than intended sample. Since relative sizes of the studies may play a role in replication success and how probitive replications are, we show the sample sizes in Table 3.

An additional factor that influences the interpretation of a replication is how close the replication's

Table 3: Comparison of sample size for original, replication, and rescue samples and measures of closeness for replication and rescue samples.

Paper	Score	N			closeness	
		Original	Replication	Rescue	Replication	Rescue
Krauss & Wang 2003	1.00	101	19	75	close	very close
Ngo et al. 2019	1.00	31	12	77	very close	very close
Todd et al. 2016	1.00	63	26	55	very close	very close
Jara-Ettinger et al 2022	0.75	144	147	426	exact	exact
Porter et al. 2016	0.75	145	168	136	close	very close
Birch & Bloom 2007	0.00	103	73	247	very close	very close
Child et al. 2018	0.00	35	40	98	very close	very close
Chou et al. 2016	0.00	100	158	252	close	very close
Craig & Richeson 2014	0.00	121	76	127	exact	exact
Gong et al. 2019	0.00	155	90	137	far	far
Haimovitz & Dweck 2016	0.00	132	97	141	exact	exact
Hopkins et al 2016	0.00	147	93	161	very close	very close
Paxton et al. 2012	0.00	92	82	160	close	close
Payne et al. 2008	0.00	48	23	23	far	very close
Schectman et al. 2010	0.00	22	20	21	close	close
Tarampi et al. 2016	0.00	139	212	166	close	close
Yeshurun & Levy 2003	0.00	18	10	18	close	very close

methods were to the original. In the rescue projects, we aimed to have methods be as close as was feasible or appropriate. However, rescue projects varied in how close the re-replications actually were, often due to limitations in the availability of original stimuli and original instructions, in addition to the use of primarily online subject pools. Table 3 shows the closeness of each first replication and rescue according to the classification scheme from LeBel et al. (2018). TODO CLOSENESS NUMBERS SHOULD BE DOUBLE CHECKED AND FIRST REPS REVISED AS NEEDED

Overall, we do not have a clear picture of why certain studies replicated in the rescue sample and others did not, other than a few cases where fixing sample size issues may have helped.

3.4 Case studies

Given the mix of successful and unsuccessful rescues, we discuss a few projects where we have speculations about why they turned out the way they did.

One of the rescues that went from a replication with score of 0 to a rescue with score of 1 was the rescue of Krauss & Wang (2003). This study looked at the influence of a guided thinking on whether or not people gave correct justifications (drawn or written) for their answer on the Monty Hall problem. The original paper reported correct justification from 2/67 (3%) in the control condition and 13/34 (38%) in the guided thinking condition. The first replication struggled to recruit participants who were naive to the problem (an exclusion criterion), and many participants give very short text responses in the provided text box (only textual responses were allowed). The replication found 0/8 correct justifications in the control and 0/11 in the guided thinking condition. While we can't know for sure what caused the non-replication, there were clear problems observable from the small final sample and low-quality responses. The rescue targeted these issues by adding a pre-screen for naivete to the Monty Hall problem, switched the name of the problem (to reduce googling for answers), and had participants upload drawings for their justifications. Collectively, these changes brought the rescue closer to the intent of the original. The rescue had 1/40 (2%) correct justifications in the control group and 6/35 (17%) in the guided thinking group. The rescue effect is smaller, but the overall pattern of results replicated, and the online adaptation in the rescue feels like it could be built on.

Another successful rescue was that of Ngo et al. (2019). Here, the original study found a large effect. The first replication, powering for 80% power on the reported effect, recruited a small sample of 12 people. It failed to find the effect. The rescue, powered using 2.5x the original sample (as recommended by Simonsohn (2015)), recovered a clear effect (albeit a much smaller one). There are reasons to think that some effect sizes in the literature may be inflated Camerer et al. (2018), and it is also possible that changes to experiments or switches to online could result in noisier samples and thus smaller effect sizes. Therefore, replications with smaller samples than the original (even if powered to the original effect size), may not be very diagnostic, and could potentially benefit from a re-replication.

Not all rescues of small replications succeeded, however. Payne et al. (2008) was a study of the effects of sleep versus wake on memory consolidation that showed participants a number of images and then hours later (after either sleep or no sleep) measured their recall for parts of the images. The first replication struggled to recruit participants and only got 23 (the original had 48). The rescue attempted to recruit a larger sample (target 88), but due to difficulties getting participants to complete the second part of the experiment 12 hours after the first, the rescue only managed to recruit 23 people. The lesson here may be that sleep research is difficult to conduct online. However, an online replication by Denis et al. (2022) reports qualitatively similar (but quantitatively smaller) results to Payne et al. (2008) on related but not identical measures.

(TODO other successes or failures we want to discuss?) (Ben: Emily's project was interesting: the rescue succeeded but with big stimulus-level random effects. Verity's project was also interesting: the previous replication, as well as another in the literature, showed similar but smaller findings and it really looked like she could get the effect with a bigger sample, but she ended up finding totally different results.)

4 Discussion

We presented the results of 17 new replications that attempted to "rescue" previous failed replications reported in Boyce et al. (2023) by identifying and ameliorating possible causes of non-replication. 5 of these rescue projects (29%) mostly or fully replicated the original results.

We don't have qualitative or quantitative explanations for why the specific studies did or did not replicate. In some cases, increasing sample size and fixing internal validity issues in the replication seems to have led to a successful rescue (although we can't establish causality even in these cases). However, there were other cases where the first replications had issues with a small sample or deviations in the implementation, and the rescue addressed these issues but still failed to replicate the original results. We can't predict what replication failures are likely to resolve given another, more thoughtful try, beyond that suggestion that glaring problems and low samples may sometimes be resolvable.

The rescues all showed smaller effect sizes than their original studies, regardless of whether the pattern of effects replicated. A large minority of replications had effect sizes that were statistically inconsistent with the original effect, even accounting for expected levels of heterogeneity between studies. These diminished and inconsistent effects suggest that even if a re-replication "works", it may be difficult to build upon as follow-up studies will need large samples to detect small effects.

The reported rescue projects are a small sample of replications. They are also chosen non-randomly, as they have been selected for twice by student interest. However, this selection bias is likely to correlate with how graduate students choose what topics to work on and what studies to build on. Nonetheless, with a small sample all our estimates are highly uncertain.

The authors of the rescue projects put substantial effort into trying to set up rescues that had a good chance of success, but projects were constrained by budget limitations, a short timeline, and primarily running online studies. These limitations are representative of the sort of resource limitations often faced by early-career researchers. That said, it is possible that different results might be obtained in better-resourced settings, by scientists with more expertise and more time. Thus, where re-replications failed, we do not make any statements about whether the original results were false positives, we merely claim the original results do not support cumulative research by early-career researchers under the constrained conditions we tested.

We opened this paper with a question about what an early-career psychology researcher should do given a failed replication: should they *try again* or *give up* and move on? From our sample of testing the *try again* approach, it seems that the odds of a re-replication working are low (consistent with Ebersole et al. (2020)). Especially if there is not a clear, identifiable reason for the first replication's failure, another try is unlikely to recover the original result, and it may be more efficient to *give up* and build on a different result instead.

Acknowledgements

TODO

Author Contributions

TODO

References

Boyce, V., Mathur, M., & Frank, M. C. (2023). *Eleven years of student replication projects provide evidence on the correlates of replicability in psychology.* https://doi.org/10.1098/rsos.231240

Camerer, C. F., 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*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., … Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Consortium, O. S. (n.d.). Estimating the reproducibility of psychological science. *Science*. Retrieved March 17, 2023, from https://www.science.org/doi/full/10.1126/science.aac4716?casa_token=IJ35TwwlcjsAAAAA%3AqiP68QbVAHleIg9zD3WugKWuV6Oa5rswS0VQnDsCq5I14ME4WIQabNGVD_T6SBSuAt6voVHNnWc0sw

Cova, F., Strickland, B., Abatista, A., Allard, A., Andow, J., Attie, M., Beebe, J., Berniūnas, R., Boudesseul, J., Colombo, M., Cushman, F., Diaz, R., N'Djaye Nikolai Van Dongen, N., Dranseika, V., Earp, B. D., Torres, A. G., Hannikainen, I., Hernández-Conde, J. V., Hu, W., … Zhou, X. (2021). Estimating the Reproducibility of Experimental Philosophy. *Review of Philosophy and Psychology*, *12*(1), 9–44. https://doi.org/10.1007/s13164-018-0400-9

Denis, D., Sanders, K. E. G., Kensinger, E. A., & Payne, J. D. (2022). Sleep preferentially consolidates negative aspects of human memory: Well-powered evidence from two large online experiments. *Proceedings of the National Academy of Sciences*, *119*(44), e2202657119. https://doi.org/10.1073/pnas.2202657119

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B. V., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., Lazarević, L. B., Rabagliati, H., Ropovik, I., Aczel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., … Nosek, B. A. (2020). Many Labs 5: Testing Pre-Data-Collection Peer Review as an Intervention to Increase Replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331. https://doi.org/10.1177/2515245920958687

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *eLife*, *10*, e71601. https://doi.org/10.7554/eLife.71601

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, *351*(6277), 1037–1037. https://doi.org/10.1126/science.aad7243

Jara-Ettinger, J., Levy, R., Sakel, J., Huanca, T., & Gibson, E. (2022). The origins of the shape bias: Evidence from the Tsimane'. *Journal of Experimental Psychology: General*, *151*(10), 2437–2447. https://doi.org/10.1037/xge0001195

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating Variation in Replicability: A "Many Labs" Replication Project. *Social Psychology*, *45*(3), 142–152.

https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Krauss, S., & Wang, X. T. (2003). The psychology of the Monty Hall problem: Discovering psychological mechanisms for solving a tenacious brain teaser. *Journal of Experimental Psychology: General*, *132*(1), 3–22. https://doi.org/10.1037/0096-3445.132.1.3

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). *A Unified Framework to Quantify the Credibility of Scientific Findings.*

Mathur, M. B., & VanderWeele, T. J. (2020). New statistical metrics for multisite replication projects. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *183*(3), 1145–1166. https://doi.org/10.1111/rssa.12572

Ngo, C. T., Horner, A. J., Newcombe, N. S., & Olson, I. R. (2019). Development of Holistic Episodic Recollection. *Psychological Science*, *30*(12), 1696–1706. https://doi.org/10.1177/0956797619879441

Olsson-Collentine, A., Wicherts, J. M., & Van Assen, M. A. L. M. (2020). Heterogeneity in direct replications in psychology and its association with effect size. *Psychological Bulletin*, *146*(10), 922–940. https://doi.org/10.1037/bul0000294

Patil, P., Peng, R. D., & Leek, J. T. (2016). What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*(4), 539–544. https://doi.org/10.1177/1745691616646366

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and Reasoning in Moral Judgment. *Cognitive Science*, *36*(1), 163–177. https://doi.org/10.1111/j.1551-6709.2011.01210.x

Payne, J. D., Stickgold, R., Swanberg, K., & Kensinger, E. A. (2008). Sleep Preferentially Enhances Memory for Emotional Components of Scenes. *Psychological Science*, *19*(8), 781–788. https://doi.org/10.1111/j.1467-9280.2008.02157.x

Simonsohn, U. (2015). Small Telescopes: Detectability and the Evaluation of Replication Results. *Psychological Science*, *26*(5), 559–569. https://doi.org/10.1177/0956797614567341

Tarampi, M., Heydari, N., & Hegarty, M. (2016). *A Tale of Two Types of Perspective Taking: Sex Differences in Spatial Ability.* https://journals.sagepub.com/doi/full/10.1177/0956797616667459

Thomson, K. S., & Oppenheimer, D. M. (2016). Investigating an alternate form of the cognitive reflection test. *Judgment and Decision Making*, *11*(1), 99–113. https://doi.org/10.1017/S1930297500007622

Yeshurun, Y., & Levy, L. (2003). Transient Spatial Attention Degrades Temporal Resolution. *Psychological Science*, *14*(3), 225–231. https://doi.org/10.1111/1467-9280.02436