TODO (re-replication) (student) (cumulative science) (try again) (replication failure)

Veronica Boyce[1,*], TODO[1], Michael C. Frank[1]

[1]Stanford University

Abstract

Title ideas:

Can failed replications be rescued? Evidence from N re-replication experiments suggests rarely.

Try again? Attempts to rescue failed replications were generally unsuccessful

A low rate of successfully replicating experiments with a prior unsuccessful replication

Once is probably enough. Low replication rates when re-replicating N studies with failed replications.

- may want to include "student" or "class"

Abstract:

In order to do cumulative science, results need to be replicable. However, an individual direct replication could get a different result than an original ("replication failure") for a number of reasons, which vary in whether they would apply to a potential second replication. Given one failed replication, how likely is a second attempt at replication to succeed? As a pragmatic matter, when a replication fails, scientists have to decide: do they make a second attempt at replication, or move on?

We re-replicated 17 experiments that had previously had a failed student replication, reported in Boyce et al. (2023). In N/17 of these rescue projects (TODO %), the results mostly or fully replicated. We discuss the relative effect sizes and levels of statistical consistency between original effect sizes and replication effect sizes.

While a few studies with failed replications were rescued with larger sample sizes and experimental tweaks, our overall results indicate that most failed replications stayed failed. If the goal is to do cumulative science, one may wish to stop after one failed replication.

TODO MAKE THIS ABSTRACT BETTER!

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. Should you try again, running another replication, perhaps with a larger sample or other tweaks to possibly fix the replication? Or should you give up on replicating that specific study, and find a different study to replicate and build upon?

There has been much discussion in the literature about how we should interpret replication results in terms of the existence of a true effect or effect size. TODO CITATIONS

We take a different angle on replications, and instead focus scientists decisions on what studies to run. Individual studies (and replications) are rarely definitive when it comes to scientific evidence, but they often influence what studies a scientist conducts. If a study works, scientists are likely to continue to purse that result, but if a study doesn't work, scientists have to choose whether to try again, or switch to a different study.

How definitive is a replication failure? That is, if a study fails to replicate, what are the chances that an additional replication succeeds. After CITE OSF found an overall replication rate of WHATEVER, concerns were raised by (**gilbert?**) that method differences were the problem. In 11 of the OSF studies, there had been concerns about protocol fidelity raised by original authors prior to data-collection. 10 out

*Corresponding author. Email: vboyce@stanford.edu

of these 11 failed to obtain significant results in the replication. In a follow-up, Ebersole et al. (2020) re-replicated 10 of these 11 in large samples, using both the RP:P protocol and a new protocol revised under advise of the original authors or other experts. None of the RP:P protocols found significant results, 2 of the 10 revised did (but not the 1 that was significant in the original RP:P sample). The replication effect size was much smaller CITE NUMBER. We could frame Ebersole et al. (2020) as an attempt to rescue 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.

1.1 Why replication failure?

The idea of replication failure implies a certain time direction: first a "successful" study and (later) a "failed" replication. (CITE TIME INVARIANCE) We can also consider this as two very-related studies that got different results, and thus the question is what difference between these two studies could account for the differing results.

[not sure if it would be better to structure this differently]

A replication can have different results from the original study for a number of different potential reasons (see also Ebersole et al. (2020)). The original result may have been a fluke, through chance or p-hacking. The original result might be of much less generalizability than is assumed, and very sensitive to the exact conditions and time it was run in ("hidden moderators"). The original could have an inflated effect size, due to some combination of heterogeneity, p-hacking, luck, and publication bias.

The replication could have been underpowered due to any of inflated original effect size, attrition in sample, or the original also being underpowered. A manipulation might have failed in the replication (this is a proximate cause that also raises the question of why the manipulation check failed). There might have been a difference in materials or instructions between the two, that turns out to matter (more likely if materials are not available). The experiment might not port well to online due to some details. A change in sample (ie. to online) may decrease data quality without sufficient attention checks. The replication might have made changes (for convenience or a new sample) that turn out to matter.

[maybe try to make the point here that if there are changes in time/place/population, you want a good *translation* of the experiment not an identical copy]

While one can theorize after the fact about potential causes, 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).

1.2 Re-replication

What would these potential causes of replication failure suggest would happen in another replication? For our purposes, we're considering another replication with roughly the same environment, knowledge, and resources of the first replication, but with the benefit of hindsight on the first replication.

Heterogeneity and lack-of-generalizability reasons mostly suggest another failed replication, as the second replication is likely to be closer to the first replication than to the original in these random factors. CITE THAT PAPER about types of heterogeneity and which way they would fall

If the original result was p-hacked or a luck false positive, then another replication is also likely to fail.

Manipulation check failure may not be fixable, but it might at least be diagnosable.

Underpoweredness and inflated effect size may be recoverable depending on the extent – if the effect size was inflated and the first replication powered for the effect, a better powered (but not huge) replication may be able to recover a smaller, but still convincing effect.

Changes between original and replication may or may not be fixable – if the original materials are still unavailable, a second replication is no better off. If a second replication can get closer to the original, either by using closer materials, or better adapting the study to a new environment, the rescue might succeed at better approximating the results of the original.

Sensitivity to environment or not working well online are unlikely to be fixable if the replications remain constrained to using online methods. Similarly, if the result is specific to a special population, then a new

replication is also likely to fail, unless it has gained access to this population.

Overall, it seems that there are both potential causes of replication failure that may be resolvable even with a similar level of resources and potential causes that are not resolvable, either at all, or given resource constraints.

1.3 Current study

[Probably needs a better transition in] Here 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, Mathur, & Frank (2023). This is a non-random sample, and given the small number of projects, we provide descriptive statistics and discuss qualitatively some potential sources of differential re-replication outcomes.

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 replications and rescues), links to individual project data and analyses for rescue projects, and the analytic code for this paper are all available at OSF LINK GOES HERE.

2.1 Sample

The experiments that were re-replicated were a non-random sample of studies from 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 (for feasibility reasons if we needed to increase power). We then contacted the replication project authors for permission to share their report and github repository with a new student and include it as a supplement on a resulting paper. This left 27 options that were offered to the students. 20 students chose to do rescue projects; 3 students took an incomplete or did not indicate interest in being part of the rescue 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 original replication might not have worked, and address them if they could. A list of possible reasons and solutions LINK was given to students. 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. (See table TODO for post-exclusion sample sizes of original, replication, and rescues). Across the 16 studies, we spent $5471, for an average of $342 per project.

2.3 Pre-registration

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.

2.4 Coding of results

We followed Boyce et al. (2023) in what study properties we measured and what measures of replication success we used.

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 compliment to the overall subjective rating of success, we followed Boyce et al. (2023) in also doing a statistical comparison on one key measure of interest for each study.

We also recorded the same set of potential correlates that were used in Boyce et al. (2023) for original, replication, and rescue (these were already rated for original and replication). These potential correlates included features of the original study including the subfield of the study (cognitive v social v other psychology), it's 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, and 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, we count 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 would 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 with a crowdsourced platform or not.

3 Results

Our primary question of interest is how many of these 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 whether 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 this in terms of how confident one would be to build on this 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 0.903. Across the 17 replications, 5 mostly or fully replicated the original results according to the subjective replication ratings. 11 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% of the time. TODO DO WE WANT TO BOOTSTRAP A CI ON THIS?

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.

Table 1: Correlations between an individual predictor and the subjective replication score of the rescue project.

Predictors	r	p
Social	0.110	0.686
Other psych	-0.320	0.228
Within subjects	0.299	0.261
Single vignette	-0.065	0.810
Switch to online	0.140	0.606
Open data	0.213	0.427
Open materials	0.449	0.081
Stanford	-0.251	0.347
Log trials	0.095	0.727
Log original sample size	-0.060	0.825
Log rescue/original sample size	0.005	0.986
Log replication sample size	-0.374	0.153
Log replication/original sample size	-0.515	0.041
Log rescue/replication sample size	0.495	0.051

Another successful replication was that of Ngo, Horner, Newcombe, & Olson (2019). Here, the original study found a large effect, and so the first replication, powering for 80% power on the reported effect, recruited a small sample of 12 people, and then failed to find the effect. The rescue, powered using 2.5x the original sample (as recommended by (**simonsohn2015a?**)), recovered a clear effect (albeit a much smaller one). There are reasons to think that some effect sizes in the literature may be inflated CITATIONS, and separately potential reasons that slight changes to experiments, or switches to online, could result in noisier samples (and thus smaller effect sizes). Thus, replications with smaller samples than the original (even if powered to the original effect size), may not be that diagnostic, and could potentially benefit from a re-replication.

Not all rescues of small replications succeeded, however. Payne, Stickgold, Swanberg, & Kensinger (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.

(TODO other successes or failures we want to discuss?)

(could discuss tarampi as an example where there were potential issues, we fixed them and it still didn't work?)

3.2 Correlates of rescue success

We ran correlations between the set of predictor variables used in Boyce et al. (2023) and the subjective replication scores of the rescues (TABLE WHATEVER). As the number of rescues is small, and many of these predictors are correlated, these correlations should not be overinterpreted.

Given that a couple of the successful rescues seemed to succeed in part because of larger sample sizes, we also added predictors related to the first replication sample size and the relative sizes of the samples (post-hoc, not pre-registered). While not significant, 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.

Small replication samples relative to original and rescue could be due both to a) aiming for a small replication sample due to aiming for power for a reported large effect size or b) difficulties with recruitment or high exclusion rates leading to a smaller than intended sample.

Two measures associated with how satisfying a replication attempt was are its sample size (relative to the

Table 2: 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	NA	18	10	NA	close	NA

original) and how close the replication was to the original. We show these measures (using the classification scheme of LeBel, McCarthy, Earp, Elson, & Vanpaemel (n.d.) for closeness) in TABLE TODO.

Aside from the suggestion that fixing sample size issues may have helped, it's unclear why some projects replicated this time and most did not.

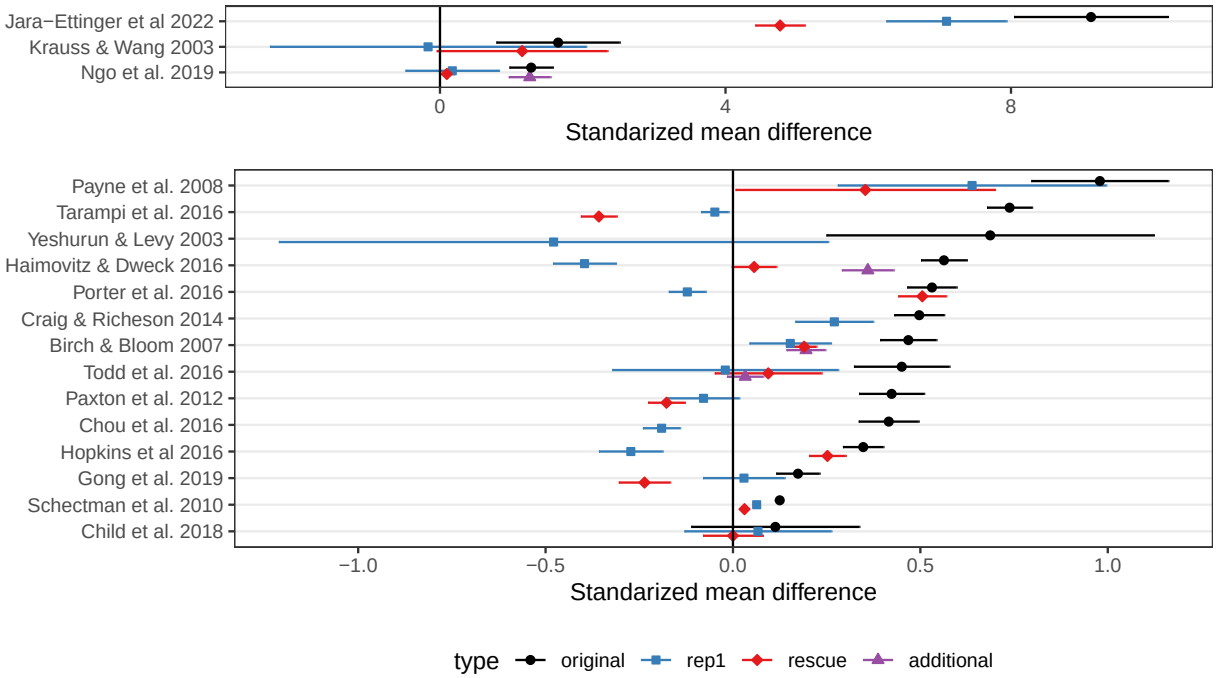


Figure 1: Standardized effect sizes of original studies, first replications, rescues, and additional replications if available. Due to the large effect size of a couple studies, large effect studies are shown in a separate panel.

Table 3: 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	NA	NA	0.005	NA
Craig & Richeson 2014	NA	NA	0.302	NA
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	NA	NA
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	0.071	0.031	0.245	0.388
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	NA	NA	0.016	NA

3.3 Effect sizes

In addition to looking at replication success overall, we also statistically compared the rescue, original replication, and original study on a key effect (we follow Boyce et al. (2023) in the determination of a key effect). When we were aware of additional direct replications (either from other class projects, or external replications in the literature), we also include these.

We first compare the original and replications effect sizes on the key effect of interest. We report this comparison in standardized mean difference (SMD) units. One potential issue with comparisons using SMD is that noisier measures will have smaller standardized effect sizes even if the effect on the original scale is the same. The effect sizes are shown in FIGURE TODO. In general, the replication and rescue effect sizes were smaller than the original effect sizes, and in a couple cases the effects were in the opposite direction.

3.4 Consistency of effects

We statistically compared the original effects with the replication effect sizes to determine their level of statistical consistency, that is whether these effects were likely to be drawn from the same distribution, given a certain level of heterogeneity.

We note that statistical consistency does not always align with our intuitions of replication or with whether an effect provides support for the hypothesis (Patil, Peng, & Leek, 2016). A replication may be statistically consistent with the original effect but not provide any evidence for the claimed result.

We use p-original to evaluate how consistent the original effect size is with the totality of replications (first replication, rescue, and additional if found). Because of the small number of replications, we impute the heterogeneity value of $\tau = .21$ SMD, which is the average level of heterogeneity found by (**olsson2020?**) in prior multi-site replications in psychology. P-original measures the p-value on the null hypothesis that the original effect and the replications come from the same distribution.

P-original values are shown in TABLE TODO. The median value of p_original was 0.15 [IQR: 0.06 - 0.29]. 21% of the p_original values were less than .05, indicating by conventional thresholds a rejection of the null hypothesis that the original and the replications came from the same distribution for the given

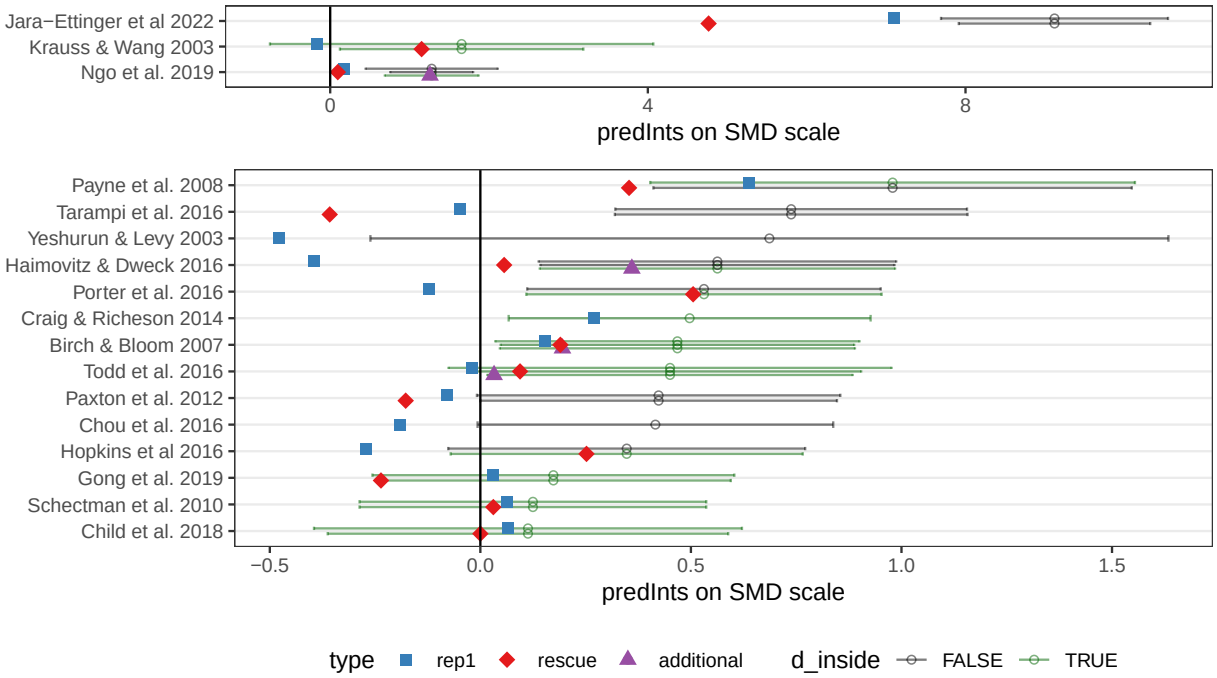
imputed level of heterogeneity.

As secondary measures, we also calculated the p-original values for a) the original and the rescue, b) the original and non-rescue replications, and c) between the rescue and other replications. For a) the rescue versus the original, median value of p_original was 0.09 [IQR: 0.01 - 0.61], and 43% of the p_original values were less than .05. For b) the non-rescue replications versus the original, median value of p_original was 0.12 [IQR: 0.01 - 0.32], and 44% of the p_original values were less than .05. For c) the rescue versus the other replications, median value of p_original was 0.39 [IQR: 0.14 - 0.78], and 23% of the p_original values were less than .05.

NOTE THERE ARE A LOT OF MISSING DATA HERE!!!

Visualizations of the statistical consistency between the original and each replication are shown in TODO FIGURE. The range around the open circle is the 95% predictive interval, representing the interval where, assuming that both are drawn from the same distribution, the replication will fall 95% of the time, given the precision (standard deviation) of the replication (Patil et al., 2016). We again impute a heterogeneity value of $\tau = .21$ SMD for the construction of the prediction intervals.

Overall, a number of these rescues and replications are not statistically consistent with the original effects allowing for this level of heterogeneity.



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 possible causes of non-replication and ameliorating them. 5 of these replications (29%) mostly or fully replicated.

We don't have qualitative or quantitative explanations for why some replicated and some didn't. In a couple 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 studies that had small replication samples, or implementational deviations in the replication, and rescues were still unsuccessful. We can't predict what replication failures are likely to resolve given another try (or a more thoughtful try), beyond that suggestion that glaring problems and low samples may sometimes (but not always) be resolvable.

Another pattern we observed was that the effect sizes of even the successful replications tended to be substantially smaller than the original effect.

[SAY something about statistical consistency]

4.1 Limitations

The reported rescue projects are a small sample of replications, and the effects explored are non-random, as they have been doubly selected by student interest. However, this non-random may be a useful selection bias as it is correlated with how students choose what to work on.

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. Thus, we do not make statements about whether these studies are "true" or "false-positives", merely that they do not support cumulative research by early-career researchers under these conditions.

We opened this paper with a question about what an early-career researcher should do given a failed replication. Should one try again or 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)), so especially if there isn't a super clear failure mode to point to, another try probably won't fix it.

Acknowledgements

Author Contributions

References

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

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., … 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. http://doi.org/10.1177/2515245920958687

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. http://doi.org/10.1037/0096-3445.132.1.3

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

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

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. http://doi.org/10.1177/1745691616646366

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. http://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. http://doi.org/10.1177/0956797614567341

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