Historical summary of “High replicability of newly discovered social-behavioural findings is achievable” paper

John Protzko, Brian Nosek, Sebastian Lundmark, James Pustejovsky, Nick Buttrick, and Jordan Axt

Protzko et al. (2023) provided evidence of high replicability across multiple indicators from a prospective replication project. We hypothesized that high replicability occurred because of a suite of rigor-enhancing practices. We provided evidence against an important alternative explanation: that the selected findings were trivially true or obviously predictable, compared to other research in the social-behavioral sciences that have shown weaker replication success. This proof-of-concept project demonstrated that high replicability is achievable in a prospective context. We called for more research to clarify the causal contributors to replicable findings.

Bak-Coleman and Devezer (2024) identified an important erroneous claim in our paper. We said,

> "*All confirmatory tests, replications and analyses were preregistered both in the individual studies (Supplementary Information section 3 and Supplementary Table 2) and for this meta-project (https://osf.io/6t9vm)*."

The 80 individual studies were preregistered but only a subset of the analyses for the meta-report were preregistered. Specifically, the statistics under the heading 'Replication Rates' in the Results section were not preregistered, and the comparison of self-confirmations and independent replications depicted in Figure 2 was not preregistered. Furthermore, there were some discrepancies between the models reported under the results section and what was preregistered, though the results are very similar with or without those discrepancies (see the draft correction in the Appendix below). We are embarrassed by our erroneous statement about what was preregistered.

How it happened

This collaboration began at a conference in 2012, organized by Jonathan Schooler and funded by the Fetzer Franklin Fund. The meeting included researchers from many fields who had observed evidence of declining and disappearing effects. Some presenters, like Brian Nosek, Leif Nelson, and Jon Krosnick, offered conventional explanations for observing weaker evidence in replications compared with original studies such as selective reporting, *p*-hacking, underpowered research, and poor methodological transparency. Other presenters offered unconventional explanations, such as the act of observation making effects decline over time, possibilities that Schooler believed worthy of empirical investigation (Schooler, 2011) and others have dismissed as inconsistent with the present understanding of physics.

Schooler, Nosek, Nelson, and Krosnick discussed conducting a prospective replication project that would try to eliminate decline in replicability by including presumed solutions for many

conventional explanations. This way, if conventional explanations were sufficient to understand decline, then high replicability might be observed, demonstrating that decline is not inevitable. And, if the solutions for the conventional explanations were not sufficient, then decline might still be observed.

While the unconventional explanations were not considered plausible (or even possible) to most of the team, they agreed on an approach that included tests of those possibilities within a so-called “best practices” proof-of-concept study pursuing high replicability. As Schooler (2014) noted in a commentary in *Nature* about the project published prior to data collection: “If the studies replicate flawlessly, we will have established a gold standard for reproducible studies. If they do not, then our approach will present an opportunity to rigorously assess the reasons.”

Protzko and Schooler (2017, p. 100) introduced the project and its purpose in a published article as the following:

> “Research teams at UC Berkeley, Stanford, and the University of Virginia have joined with our lab (at UC Santa Barbara) to examine the replicability of new findings that are uncovered while engaging in hypothesized ‘best practices’ for maximizing the reliability of findings. This project (supported by the Fetzer Franklin Fund) is carefully documenting all aspects of newly developed scientific studies, using highly powered research designs, and then repeating the studies at the various universities. Such prospective replication experiments may illuminate the factors that govern the replicability of scientific findings, including: researchers’ investment in the hypothesis, the number of times a protocol is repeated, and the manner in which methodologies and outcomes are communicated.”

During project planning (2013-2014), we decided to assess whether effects decline with several popular descriptive indicators and with a design that would enable an inferential test. We planned to report descriptive outcomes mostly parallel to the *Reproducibility Project: Psychology* (Open Science Collaboration, 2015) that was being conducted and reported around the same time (i.e., whether we observe declines in statistical significance rates, average effect size, effects within confidence intervals, and significance of all evidence combined meta-analytically; We did not plan to use a fifth indicator from the Reproducibility Project – subjective assessments of replication success). In the planning documentation, we described the design and how inferential tests could be conducted with it, but we failed to summarize the descriptive indicators.

In 2018, we recruited a statistician to prepare the meta-analytic modeling for the inferential tests. We did not ask them to develop an analysis plan for the descriptive outcomes. While some of us have grown to appreciate the value of analysis plans for all analyses, at that time, the benefits of preregistration were considered most acute for inferential testing. This project, like the *Reproducibility Project: Psychology* (Open Science Collaboration, 2015), adopted an approach of preregistering the underlying replications, and summarizing them in a meta-project with several unregistered descriptive indicators to highlight robustness or heterogeneity across methods of conceptualizing replicability.

In our original submission of the paper in 2020, we wrote (bold added):

> "*Each of the 16 discoveries was obtained through pilot and exploratory research conducted independently in each laboratory. Labs, using their own criteria, decided which discoveries to submit for replication. Labs introduced 4 provisional discoveries each, resulting in 16 self confirmatory tests and 64 replications (3 independent and 1 self-replication for each), testing replicability and decline.* ***All discoveries, replications, and the analyses presented here were preregistered.***"

This referred to the 80 individual studies comprising the meta-project, but it could also be taken to refer to all analyses for the meta-project. Unfortunately, following a reviewer's request to provide a link to the meta-project preregistration, we revised the sentence to be unambiguous *and* incorrect:

> "***All confirmatory tests, replications and analyses were preregistered both in the individual studies (Supplementary Information section 3 and Supplementary Table 2) and for this meta-project (https://osf.io/6t9vm).***"

The statements about preregistration of the individual studies and meta-project should have been made separately. Only ***some*** of the analyses for the meta-project were preregistered (many preregistered analyses are summarized in a subsection labeled "Confirmatory Analyses" in the methods section; preregistrations of follow-up studies to assess predictability of the discoveries are not reported there). We failed to identify and correct this error through subsequent revisions. This is not the only error in the paper (see our submitted correction in the Appendix), but it is an important one.

Decision to Retract

We are embarrassed by our elementary errors with reporting on the preregistration, and agree that they need to be addressed. We drafted a correction notice in December 2023. In July 2024, the journal responded that they perceived the problems to be sufficiently serious to merit retraction and resubmission following revision.

We discussed this with the journal and decided that we could defer to the journal's judgment and accept retraction based on our preregistration errors. However, the journal perceived additional problems—most especially they believed that we had not planned the project as we described above, instead decided to make the project about replicability and report the descriptive indicators after observing the outcomes (see Bak-Coleman & Devezer, 2024). Because these claims are inaccurate, we could not endorse those reasons for retraction.

We are grateful to *Nature Human Behaviour* for inviting us to submit a new version of the paper. We will revise the paper to address the inaccuracies and consider other critiques about the substance of our claims during revision. Also, our findings spurred interest in what occurred in

the pilot phase of the project. We are likewise quite interested to learn more about the pilot phase. Our current paper examined replicability after employing design features that we believed eliminated conventional reasons for decline. For that purpose, the evidence began with a confirmatory test of a hypothesized discovery and ignored the pilot data where there are conventional reasons to expect decline. Investigation of the pilot phase may reveal additional insights about what happened before that test that could have promoted or inhibited replicability. An plan for investigating the pilot phase was preregistered in March 2019 to be conducted as a separate project (https://osf.io/7yn2z). As of September 2024, that project is still ongoing by a subset of the team.

Finally, beyond the events leading to the journal's determination that our paper should be removed from the scholarly record that are addressed here, we welcome the substantive critiques of the project's evidence and claims. We hope that a revised paper will further facilitate that scholarly discussion.

References

Bak-Coleman, J. & Devezer, B. (2024). Claims about scientific rigour require rigour. *Nature Human Behaviour*.

Protzko, J. & Schooler, J.W. (2017). Decline Effects: Types, Mechanisms, and Personal Reflections. In S.O. Lilienfeld & I. D. Waldman. (Eds.) *Psychological Science Under Scrutiny*, 85-107. https://doi.org/10.1002/9781119095910.ch6

Protzko, J., Krosnick, J., Nelson, L., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S., Walleczek, J., & Schooler, J. W. (2023). High replicability of newly discovered social-behavioral findings is achievable. *Nature Human Behaviour, 8*, 311-319. https://doi.org/10.1038/s41562-023-01749-9

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, 470(7335), 437–437. https://doi.org/10.1038/470437a

Schooler, J. W. (2014). Metascience could rescue the 'replication crisis'. *Nature*, 515(7525), 9. https://doi.org/10.1038/515009a

Appendix: Draft correction from December 2023

The correction below was drafted by our team in December 2023 as a summary of the errors that we recognized at that time. It is offered for historical purposes, and to highlight additional errors in Protzko et al. (2023) that we did not discuss above.

A careful reader discovered deviations between our preregistrations and what we reported in the paper.

In the original article, we indicated that "All confirmatory tests, replications and analyses were preregistered both in the individual studies (Supplementary Information section 3 and Supplementary Table 2) and for this meta-project (https://osf.io/6t9vm)." We did not, however, preregister that we would report the descriptive statistics of the replication rate as the

percentage of replications that achieved p < .05 in the hypothesized direction and the descriptive statistics summarizing effect sizes. We also did not preregister the reported power analyses and some sensitivity and heterogeneity analyses.

Preregistered analyses reported in the article were: tests of decline in effect sizes over sequential replications, tests of differences between original confirmatory findings and self-replications, tests of declining effect sizes between the first 750 and second 750 participants, and tests of blinding effects and observer effects. The preregistered analysis plan is here: https://osf.io/ba8p7, and the preregistered description of methods and aims of the research is here: https://osf.io/vdr97, alongside historical versions of those plans: https://osf.io/ahdy7/files/osfstorage. Many other documents are part of that preregistration, but they are less relevant for assessing these reporting discrepancies. Also, the 80 confirmatory and replication studies were each preregistered and used statistical significance in the hypothesized direction as a criterion for observing evidence for the hypothesis. Analysis of the survey data examining the predictability of reported effects among laypeople was preregistered; the preregistered analysis plan is here: https://osf.io/mep5w. Finally, the design and a simple analysis plan of the survey data examining the predictability of reported effects among experts were preregistered here: https://osf.io/3c5nx.

There were also three occasions in which specific analyses that were preregistered were reported differently in the paper, and a fourth analysis that was not preregistered from one of the follow-up surveys.

First, we preregistered an exploratory analysis of potential variation in replicability across labs, but in the reporting of that analysis in the supplementary information we described the analysis as confirmatory (pages 15-16, and 39). Because we did not have an a priori hypothesis, the analysis should have been described as a preregistered exploratory analysis.

Second, we preregistered an analysis comparing confirmatory tests and self-replications which is reported in the supplementary information (pages 11-12). However, in the main text, we reported a simpler, basic random effects model that did not include a predictor for the number of independent replications between the initial confirmation study and self-replication. We omitted the term for number of replications in the main text because it would only be relevant in the event of a decline effect being observed. Given that no significant declines were observed, presenting the simpler model was warranted because it was more conventional, provided greater robustness to the estimates, and better aligned with the data depicted in Figure 2 of the main text. In addition, a simpler model was deemed more appropriate in the main text to facilitate generalizability of future meta-scientific replication projects not investigating reasons for decline effects. Finally, the primary finding from the two analyses was the same (Main text analysis: -0.0022, SE = 0.0147; Preregistered analysis in the supplement: -0.0023, SE = 0.0149).

Third, we preregistered an analysis examining the change in effect sizes across successive replications. The analysis reported in the main text deviates from the model described in the

natural language preregistration at https://osf.io/ba8p7. However, the change to the analysis strategy was made in the code prior to observing the real data. The code for the preregistered analysis, developed using simulated data, implemented the model slightly differently than proposed in the preregistration. Rather than using random study-specific slopes as originally proposed, the model implemented used an AR(1) autoregressive structure over subsequent waves. Compared to a random slopes model, the approach implemented is a more parsimonious approach to capturing dependence across waves. The primary finding from testing ESs sequentially over time based on the preregistered model was highly consistent with the reported model: no evidence for a decline in ESs from the self-confirmatory test through the final replication was observed (Main text analysis: b = −0.002, t(73) = −0.38, P = 0.71 95% CI, −0.02 to 0.01; Preregistered analysis: b = −0.001, t(73) = −0.15, P = 0.88, 95% CI, −0.01 to 0.01) and the findings did not differ when we removed the fixed effect for each lab. These results were the same for 'blind' and 'not blind' studies (Main text analysis: b = 0.02; t(73) = 1.75; P = 0.10; 95% CI, −0.01 to 0.05; Preregistered analysis b = 0.02, t(73) = 1.68; p = 0.115; 95% CI, -0.01 to 0.04).

Fourth, we preregistered a descriptive study of expert predictions of replication outcomes. We reported the preregistered analysis in the main text, but in the supplementary information Supplementary Figure 4, we reported "The Spearman correlation between prediction accuracy and observed effect size was .104, p = .712, two-tailed)", which was not preregistered. The reported Spearman correlation was not used for conclusions made in the main text or supplementary information.