UCSB

UNIVERSITY OF CALIFORNIA
SANTA BARBARA

April 2, 2023

Dear Mary Elizabeth;

Thank you for the interest in our manuscript “High Replicability of Newly-Discovered Social-behavioral Findings is Achievable” and encouragement to submit a new version with additional data addressing Reviewer’s 1 critique (point #2). Preparation of this revision took much longer than we had anticipated, but we were able to complete the follow-up study and believe that the evidence is compelling to address the core concern that we discussed by Zoom following the action letter. In this response letter, we first address this primary concern and then summarize responses and revisions to the other issues raised by the reviewers. We are happy to provide additional detail as needed.

Regarding the primary concern, we now provide evidence that the findings generated during this research are not *a priori* more predictable than other social-behavioral findings that have shown relatively low replication rates in other large-scale investigations. The additional evidence increases the plausibility of our conclusion that the high replicability rate is due to improved research practices that help identify replicable findings rather than a biased selection of findings that are inherently more obvious than other findings in social-behavioral research.

To address this concern empirically, we asked an independent group to revise and generate brief descriptions of the studies included in this paper (“high replicability”) and for studies included in a prior large-scale replication study known as Many Labs 2 (Klein et al., 2018) in which only half of the studies replicated successfully (“low replicability”). We then asked a sample of 1,180 participants to predict the outcomes of studies from both samples. The outcomes from our “high replicability” sample were no more predictable than the outcomes from the “low replicability” sample, and we included a variety of checks to ensure that the two samples of studies were highly comparable in other ways. In summary, our findings were replicable but no more predictable in advance than findings that were much less replicable from the same area of research. This evidence is discussed briefly in the main text, and more extensively in the SOM. For the remainder of this cover letter, we address other concerns raised by the reviewers.

Reviewer 1’s third point wondered whether our evidence that the discoveries were non-obvious to researchers was due to our brief, conceptual descriptions of the findings. The follow-up study addressed this concern directly by having the independent group who revised and created the descriptions do so with more operational descriptions. Also, we used an independent group to generate descriptions of both our “high replicability” studies and the previous “low replicability” studies so that we would not inadvertently bias the descriptions for the “high replicability” group to be more vague than the descriptions for the “low replicability” group. We provide empirical evidence of their equivalence in the SOM and all of the materials are openly accessible for the reader to assess this directly.

Reviewer 1 also raised a philosophical concern: “The central claim is obviously true. The paper's central claim—i.e., that a high rate of replication failures is not inevitable if one uses optimal practices—is obviously true, and requires no empirical support. The only way it could be false is if there were no reliable effects at all in social science—which is on its face an absurd proposition.” Reviewer 1 then expands on this point in section (1) of his excellent review.

His concern is that it is obvious we would observe high replicability when following rigorous methods, otherwise social-behavioral research would violate basic realist assumptions about science. While many of us agree with his core philosophical commitments, these are not universally shared or understood. A similar argument could have been made (and was) that the low replicability observed in the Reproducibility Project: Psychology (RP:P; https://science.sciencemag.org/content/349/6251/aac4716) was obvious and inevitable based on existing evidence of low rigor. Nevertheless, that paper was heralded as a breakthrough paper for raising awareness of low replicability in psychology and motivation to address it. This occurred despite the fact that many methodologists found the findings obvious because of prevailing research practices and their realist assumptions. The broader research community needed empirical evidence to assess, understand, and accept the same point-of-view.

The present manuscript is the complement to RP:P. Some methodologists, including some of the present authors, perceive finding high replicability with high rigor to be obvious given our philosophical commitments. But, this is far from understood and accepted in the social-behavioral sciences (and beyond).

For example, a number of critics of RP:P and other replication efforts explicitly embrace commitments expecting low replicability regardless of rigor because of presumptions about high complexity of social-behavioral phenomena among other reasons (e.g. Iso-Ahola, 2017, 2020; Stroebe & Strack, 2014; Strack & Stroebe, 2018). Likewise, there is resistance to the methods used in this research (such as preregistration) over concerns that they will not, in fact, improve rigor and reproducibility, and might actually disrupt research progress (Goldin-Meadow, 2016a, 2016b; Scott, 2014; Shiffrin et al., 2018; Szollosi et al., 2020). Finally, although the growth of adopting these rigor enhancing practices does suggest that many share our philosophical commitments that they will improve replicability, the behaviors are still far from universal and the rigor of their implementation is not uniformly high.

As such, we believe that there is substantial value in an empirical demonstration of high replicability following high rigor. We believe that this paper will have an impact much like RP:P did in advancing the behavioral reform of the social-behavioral science research community toward higher rigor and replicability.

Reviewer 1’s fourth point raised a number of issues that were confusing in our description of the study and analyses, including the distinction between a confirmation study and a self-replication, the purpose and use of splitting samples of 1500 into two subsamples of 750, the blinding procedures, and the references to the decline effect. We revised the main text and SOM to address these concerns and improve clarity. The short answer to the purpose of many of these features was to design the study *a priori* to address exotic possibilities for the decline effect that are at the fringes of scientific discourse. For example, some have posited that effects decline, not necessarily because of questionable research practices in the discovery of new findings, but because the act of observation influences the future occurrence of such findings. While most of the present authors do not believe that such an influence is plausible, or even possible, we all agreed that it was prudent to design the study to address both conventional and exotic explanations. In the event that decline was observed this design would have helped to rule out the exotic options as explanations for the observed findings. The fact that decline was not observed made moot all of these design factors for guiding interpretation. As such we revised the manuscript so as not to dwell on the decline effect hypotheses and kept that discussion and motivation for the design elements to the SOM.

Reviewer 1's fifth point was that we overemphasized the role of the discovery process in producing high replicability in some of the paper's interpretations, when we did not actually present evidence about the discovery process itself. This is an accurate critique, and we have revised the text accordingly to more specifically call out that the evidence for replicability enhancing behaviors provided here begins with the confirmatory study that followed apparent discoveries.

Reviewer 1 raised three minor concerns that were all useful in improving the text and providing all relevant links.

Reviewer 2 raised concerns about the clarity of the goal of the project and how that gets positioned in the introduction of the paper. There is some connection in the reviewer's points to Reviewer 1's philosophical concern that high replicability should be anticipated with rigorous practices, and Reviewer 1's concerns about confusing discussion of the decline effect. The comments were very useful in improving the revision to be clearer about the goals and approach. The study was designed with both motivations: (1) to assess whether declining effect sizes is inevitable and, if so, rule out some explanations for it; and (2) to assess whether rigorous research practices are associated with high replicability. Functionally, these are the same purpose from our perspective as we believe that the posited decline effect is due to the lack of rigorous research practices.

Following Reviewer 2's second suggestion critiquing "the decline effect" as a named phenomenon, we revised the main text to highlight two points of view why smaller effect sizes might be observed in replications of original findings – sub-optimal research practices such as selective reporting and p-hacking versus inherent qualities of social-behavioral findings such as high context sensitivity. Then, instead of emphasizing "the decline effect" as a phenomenon, we discuss how our adoption of so-called optimal research practices should eliminate sub-optimal practices as a reason for observing smaller effects in replication studies. So, the only reasons left to observe declining effect sizes would be due to inherent qualities of social-behavioral findings. We retain the term "decline" to refer to the focus on evaluating whether replication effect sizes are smaller than original findings as it provides information about direction compared to using terms like “change” or “difference”.

Reviewer 2’s other major concern was about our claims that the studies were highly powered. He accurately pointed out that this is not formally correct given the observed evidence for a couple of the findings that elicited very small effect sizes. We revised the manuscript to emphasize large sample sizes and relatively precise estimation rather than strong claims about the studies being highly powered. When we do discuss power, we are more circumspect about observed power and couch the interpretations accordingly.

Reviewer 2 raised a few more minor concerns that were all helpful in revisions to improve clarity and to ensure that the appropriate links were provided.

Reviewer 3 raised some points that are addressed in the comments above for Reviewer 1 and 2, and then several specific points that were addressed in line-item revisions in the manuscript. They were quite helpful in improving clarity of the manuscript. The full text of all reviewer comments is below the signature.

Thanks again for your interest and consideration of this manuscript. Please let us know if you need any additional information from us to consider this revision for publication.

Regards,

John Protzko, Brian Nosek, and Jonathan Schooler, on behalf of the author team

Dear Dr Protzko,

Your manuscript entitled "High Replicability of Newly-Discovered Social-behavioral Findings is Achievable" has now been seen by 3 referees, whose comments are attached below. While both they and we find your work of potential interest, they have raised serious concerns. In our opinion, these reservations are sufficiently important that we cannot, at this stage, invite a revision.

One of the main concerns that has been raised is with the conceptual rationale for the study. Our reading of the study falls along the lines of R1's suggestion that the paper is about identifying the conditions under which a certain degree of replicability is achievable, and the impact certain procedures make (on replicability). However, if this is the case, then there are quite a few methodological concerns (which I won't detail here as they are spelled out very clearly in the reports). These technical concerns could, in principle, be addressed by the addition of experimental data. Given the study design, we understand that this might not be possible.

If you would consider undertaking the additional experiments that would allow you to provide strong evidence answering the conditions under which replicability is achievable/the specific procedures that contribute to replicability, we would be happy to reconsider our decision.

I'm sorry that, given these reports, we cannot offer to publish your study at this point. We are very interested in publishing such studies, and so hope that you find our referees' comments helpful.

Please let me know if you would like to discuss our decision further.

Best wishes,

Mary Elizabeth

Mary Elizabeth Sutherland,
Senior Editor

Referees' comments:

Referee #1 (Remarks to the Author):

Protzko and colleagues report a series of large, pre-registered replications of 16 different online social-behavioral studies. They arguing that their findings demonstrate that, when rigor-enhancing procedures are used in replication studies, observed replication rates are substantially higher than those observed in previous replication efforts, suggesting that, as the title puts it, "high replicability of newly-discovered social-behavioral findings is achievable."

Although the manuscript reports on the results of a very large effort that spans multiple sites, and focuses on a topic currently of considerable interest in psychology and other social sciences, I think the current manuscript suffers from a number of serious conceptual and methodological problems that render it unsuitable for answering the question it claims to address (or, as I discuss below, more charitable interpretations of that question). I describe my concerns below in roughly descending order of importance.

(1) The central claim is obviously true. The paper's central claim—i.e., that a high rate of replication failures is not inevitable if one uses optimal practices—is obviously true, and requires no empirical support. The only way it could be false is if there were no reliable effects at all in social science—which is on its face an absurd proposition. The authors write that they "report the results of a prospective replication examining whether low replicability and declining effects are inevitable when using current optimal practices." But how could low replicability and declining effects possibly be "inevitable", either with or without optimal practices? Do the authors doubt that there are robust findings throughout the social sciences, including many that researchers have successfully replicated many times, even without preregistration, large sample sizes, etc.?

It seems to me that the following two statements are undeniable: (1) some effects in the social sciences are robust and others are not, and (2) best practices like preregistration and using high-powered samples increase (on average) the probability that a robust effect will be successfully replicated in any given study. Neither of these premises can plausibly be denied. Denying the former statement would amount to saying that there are no reliable generalizations in the social sciences, which is clearly absurd (e.g., would anyone doubt that people are on average more likely to marry people they like than people they despise?). And the latter claim is mathematically or logically entailed for a great many practices—e.g., it is necessarily true that a high-powered study has a higher probability of replicating a real effect than a low-powered one, and it is similarly true by definition that reducing the number of unaccounted-for researcher degrees of freedom means the reported
p-values provide a better estimate of the true false positive rate (conditional on the assumed model), other things being equal. So, clearly, the goal here cannot really be to address the binary question posed in the paper's title. It is a foregone conclusion that replicability of newly-discovered social-behavioral findings *is* achievable; it could not be otherwise.

A charitable reading of the authors' central claim might be that what they are really trying to do is quantitatively estimate the impact of using best practices on the replicability of previous findings. That is, the question is not really whether or not replicability is *achievable* (of course it is), but *under what conditions a certain degree of replicability is achievable*, and *how much of an impact certain procedures seem to make*. If the paper were explicitly framed this way, I think it would be posing an important question. A serious effort to quantify the impact of implementing specific rigorous methodologies on replicability would be a valuable service to the field. However, framing the paper this way would also make it clear that, at least in its current presentation, the study has a number of major design limitations that in my view preclude it from providing an informative answer to the question. I discuss some of these in the next few points.

(2) The generalization target is unclear. It is never made clear in the manuscript what population of effects the authors take their conclusions to apply to. Looking at Table 1, it seems clear that all or nearly all of the effects studied here involve self-report judgments about narrowly operationalized social behaviors. In this sense, there is probably little or no basis for generalization to "social-behavioral findings" in general, let alone to other areas of psychology or science. The authors acknowledge this in the Discussion, where they write: "A question might be the constraints on the generalizability of these findings to other research. Our 16 novel findings in social-behavioral sciences each involved two between-subject conditions that could be administered online." This seems appropriate, but I think the paper's title and introduction, which do not clearly state the scope of the modeled effects, should probably also be narrowed correspondingly to make it clear what the
modeled population of studies is.

More problematically, the authors provide essentially no explanation of how the 16 effects they studied were selected for inclusion. They write: "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." But the failure to disclose the criteria used to select each effect falls short of the authors' claim of "complete methodological transparency", and the interpretation of the authors' results—and specifically, our understanding of what population of studies the results generalize to—depends critically on knowing how the studies were selected.

Suppose, for example, that each of the 16 effects had been selected based on a p-hacked p < .05 result from a single small, exploratory study. Then the results reported here would presumably be generalizable to a broader population of studies in the literature that identify effects using similar procedures (QRPs and p-hacking, as we now know, are pervasive; see e.g., John, Loewenstein, & Prelec, 2012). On the assumption that we can draw strong conclusions about replicability rates from 16 effects (see point 8 below), the conclusion we would then presumably have to draw, given the much lower rate of replication observed for previous findings published using similar selection procedures, is that there is some other difference between the present set of effects and the broader literature (e.g., perhaps the content domain is different, or the effects are more obvious, etc.).

Conversely, suppose that all 16 effects had been selected for inclusion in this project only after repeated large-sample, direct internal replications. In that case, it would be misleading to present these effects as if they were representative exemplars of the broader psychology or social science literature; after all, very few novel findings are initially reported with multiple direct replications using large samples, etc. So, an appropriate conclusion in such a case would be something more like "we show that when we use rigorous methods to try to replicate findings that were selected on the basis of multiple previous successful direct replications, we observe a high rate of success." This would still potentially be an informative finding, but I wager that few would be surprised by it, and it seems clear that no generalization would be warranted to the vast majority of studies that do *not* conduct large internal replications of novel findings before seeking to publish them.

To be clear, I am not speculating as to how these effects actually *were* selected. I am simply pointing out that, absent the authors' disclosure of the processes that led them to select these particular effects for inclusion, it is impossible to determine—even in a ballpark sense—what population of studies the present results are meant to apply to. I would strongly encourage the authors to provide (a) a supplement detailing how each effect was identified, and (b) some explicit indication of what population of studies the authors think the assembled effects is a plausible sample from. Note that this *cannot* be inferred just from knowledge of the effects/designs themselves, because selection of effects into the present study was at least partly conditional on the methodological procedures used to produce the results in the first place. I.e., the standard argument for preregistration (or failing that, for complete transparency) applies here: unless the reader knows how effects were selected, they are unable to gauge what the actual data-generating process underlying the data is.

(3) Inadequate description of effects. Many of the effects summarized in Table 1 are in my view incorrectly described, and this has a number of implications for the authors' conclusions. While I didn't read the methods for every effect in detail, in the 3 cases I randomly examined (including a careful perusal of the documents on the corresponding OSF repository), the single-sentence description did not seem to me to accurately summarize what was actually tested. For example, for the "Ads" study, Table 1 describes the central result as "Watching a short ad within a soap-operate episode increases one's likelihood to recommend and promote the company in the ad". This is incorrect in at least two ways. First, the authors did not measure likelihood of recommendation or promotion of companies; they measured *self-reported* likelihoods of these behaviors. That there is at best a very weak relationship between these things should be obvious, or else McDonald's would experience a significant boost in revenue every time it aired a single ad on TV, which is obviously not the case (indeed, there is a cottage industry within marketing research questioning whether TV ad campaigns have *any* meaningful effect on sales). Second, the current wording implies that the effect applies to companies in general, when actually the authors only asked about McDonald's, and used only a single ad comparison (McDonald's vs. Prudential). This design does not license any general conclusions about "the company in the ad"; it licenses conclusions only about McDonald's. The description in Table 1 should in my view read something like "Watching an ad for McDonald's within a soap-opera episode increases one's self-reported likelihood of recommmending and promoting McDonald's, as compared to a Prudential ad." This is only slightly longer than the current description, but provides a much more accurate summary of the actual effect. If the authors wish to make the broader claim, they need to revise their design to include multiple stimuli and a statistical model that licenses generalization over a population (see Yarkoni, 2019 for extended discussion of this point).

In the "Prediction", study, the authors sumarize the effect as "People make more complicated sets of predictions when asked to do so without having the opportunity to explore data". But the authors only studied a single set of predictions related to a single dataset involving results of congressional elections. This provides no basis for the sweeping generalization the authors draw. It seems unlikely they would have felt comfortable drawing the described conclusion on the basis of a single subject tested using many different datasets, so why draw it on the basis of many subjects tested using only a single dataset? A priori, there is little reason to suppose that variability between subjects is much larger than variability between potential datasets. A more appropriate description of the effect would be something like "People asked to identify variables they predict would influence the outcomes of Republican congressional elections identify more factors when making predictions before seeing a relevant dataset than after it."

Of the "Cookies" study, the authors write: "People will be seen as greedier when they take three of the same kind of (free) cookie than when they take three different (free) cookies". This is an inaccurate description, as no cookie-takers were actually observed; participants were asked to *imagine* how they would feel if they observed people taking cookies. A more accurate description would be: "Participants directed to imagine a specific hypothetical norm-violation scenario rate it as greedier to take three of the same kind of (free) cookie than to take three different (free) cookies." There is also a separate concern about the construct validity of this manipulation that I discuss below, and that in my view may be critical to explaining the observed effect, but I was not able to come up with a single-sentence summary that accurately captured the concern.

I did not look at the methods for the other 13 effects, but based on the above, I expect there are likely to be similar concerns with at least some of them.

In addition to simply being inaccurate, the excessively broad descriptions of the studied effects have direct implications for the conclusions the authors draw. The authors appropriately note in the Discussion that "An uninteresting reason for high replicability would be if the discoveries, although novel, are obviously true. Trivial findings might be particularly easy to replicate." They go on to report that a sample of independent raters could not very reliably predict the direction of the effect, despite the consistent replications in their sample. The problem here is that, as I noted above, the summaries the authors provide for the effects go far beyond the actual operationalizations in question (at least for the 3 I looked at, though most of the others also seem extremely broad).

For example, it is true that it is not intuitively obvious (at least to me) from the "Cookies" description in Table 1 whether participants would rate it greedier to take 3 different cookies of 1 type than 1 cookie of each of 3 different types. However, examining the actual wording of the stimuli leads to a rather different conclusion. The wording used in the presented vignette (in both experimental conditions) is "Even though there are no official rules, the norm is that people only take one cookie." It is unclear from this phrasing whether "one cookie" means a limit of one cookie per type, or one cookie in total. Consequently, it seems very intuitive that participants would rate the behavior described in the "Similar" condition as greedier than in the "Variety" condition: the former unambiguously violates the described norm (there is no way in which taking 3 cookies of one type could fail to violate the "one cookie" limit), whereas the latter is ambiguous (depending on interpretation, taking 1 cookie of each type could be consistent with the norm). The upshot is that, upon close inspection, the observed effect does in fact start to seem rather obvious—it's just that the correct attribution of the effect may be to a quirk of the stimulus, and not to any general effect of behavior on greed perception.

Viewed this way, it's not at all clear that the survey results the authors report speak to the non-obviousness of the tested effects. While hindsight is of course 20/20, at least for the 3 effects I looked at in detail, I am fairly confident that (a) I would not have been able to confidently predict the sign of effect based on 2 of the 3 provided descriptions in Table 1, but (b) I *would* have been able to confidently predict the sign of effect in all 3 cases based on a careful reading of the actual procedures used. So I think the authors should not be so quick to reject the possibility that at least some of the effects that replicated successfully were indeed more obvious than they might appear from the short summaries provided in the survey.

To be clear, I am not suggesting that the effects included in this set are so obvious as to not merit any empirical evaluation. I am simply pointing out that describing these as "16 novel experimental findings" is arguably misleading, because under some fairly natural interpretations of the phenomena in question, many are probably better described as novel operationalizations of rather obvious psychological and linguistic principles. I suspect that it would not be hard to describe many/most of the 16 effects in a way that is at least as accurate, but elicits very different assessments of predictability from respondents.

(4) Unclear justification for several analyses. A number of the key design elements and analyses in the paper are not well-motivated and potentially confusing. First, the authors' distinction between "confirmation" and "self-replication" studies is confusing and potentially misleading. It appears that they collected two separate samples known to be drawn from exactly the same population of subjects, using identical procedures, and with only a short temporal delay between samples. Given that it is hard to imagine any meaningful difference between samples arising due to a gap of a few days, what possible rationale is there for treating these as different studies, rather than collapsing them into a single study of n=1,500? It seems misleading to call the second n=750 sample a "replication" of the first. This terminology would make sense in the context of a discovery/replication design of the type found in, e.g., statistical genetics, because in that context, the discovery sample is used to explore the data and generate novel hypotheses, which are then tested in the independent replication sample. In the present context, by contrast, *all* of the tests, in all samples, are strictly confirmatory. So splitting the sample seems to accomplish nothing except a reduction in power, while decreasing the clarity of the presentation.

Second, the blinding manipulation is not clearly explained in the text, and seems to have no clear scientific motivation. If, as the authors state, the 16 effects were submitted for replication in advance, based on independent pilot studies, then there was no exploratory element at all in the present study: the authors simply applied the predetermined analyses to all datasets and recorded the results as they came out. The only way the blinding manipulation could possibly have had any effect at all is if the authors did *not* in fact follow a preregistered set of procedures in all cases—i.e., if there were undocumented degrees of freedom. But this seems to be quite clearly ruled out by the description of the procedures.

The only hint I can find as to what's going on here comes from the following sentence in the supplementary methods: "If observer effects cause the decline effect, then whichever 750 was analyzed first should yield larger effect sizes than the 750 that was analyzed second". This would seem to imply that the actual motivation for the blinding was to test for some apparently supernatural effect of human observation on the results of their analyses. On its face, this would seem to constitute a blatant violation of the laws of physics, so I am honestly not sure what more to say about this. I won't go so far as to say that there can be no utility whatsoever in subjecting such a hypothesis to scientific test, but at the very least if this is indeed what the authors are doing, I think they should be clear about that in the main text, otherwise readers are likely to misunderstand what the blinding manipulation is supposed to accomplish, and are at risk of drawing incorrect conclusions (e.g., I initially just assumed the blinding referred to a manipulation of whether or not the experimenters of certain studies were aware of the hypothesis—which *would* have been a very sensible thing to assess, had the studies been delivered in a laboratory setting rather than online).

Third, I found the recurring "decline" theme throughout the article fairly puzzling. It seems poorly motivated and at odds with the overall framing of the manuscript as a test of whether or not replicability is possible in psychology (or, as I have suggested reinterpreting it above, as an effort to quantify the impact of certain methodological choices on replicability). It's unclear what grounds there could be for the hypothesis that effects would decline between the self-confirmatory tests and the independent replications—especially given that, as noted above, the self-confirmatory and self-replication samples are in essence the very same study! Testing for a decline over time makes sense in the presence of a non-preregistered literature with heterogeneous methods, where rewards for exciting initial findings are high, and all of the standard selection and publication biases are in play. But in the context of a series of preregistered, independently conducted studies, what basis could there be for positing any change in effect size over time? Again, if the intent here is to test for some kind of supernatural observer effect, then that should be clearly disclosed, and the reader can decide for themselves how to interpret that.

The one comparison where I think we clearly *should* expect substantial decreases in effect size would be between the pilot effects observed *within* each lab (prior to submitting the effect for replication) and the (presumably larger) preregistered replications. However, the effect size estimates obtained in the pilot studies were not included in this report, so it is not possible to assess this prediction. Given the stated objective of the study, this seems like an unfortunate omission, and I would encourage the authors to include (as part of a more detailed description of the effect discovery process, as suggested above) a report of the effect sizes identified in the original pilot studies that prompted follow-up.

(5) Misattribution of high replication rates to rigorous procedures during effect discovery. In several places, the authors explicitly state that their findings demonstrate that high replicability is achievable when effect discovery proceeds in a careful way:

* "Thus, the outset for this investigation was 16 new experimental discoveries free from p-hacking or questionable research practices"
* "When novel findings were preregistered, highly powered, and done completely transparently ... the observed rate of replication was extraordinarily high."
* "It is more likely that we observed high replicability because of the rigor-enhancing methodological standards adopted in both the original research leading to discovery and the rigor in replication."
* "This eliminated publication bias that is particularly pernicious when the selective reporting systematically ignores null results"

So far as I can see, none of these claims are supported by the study design. As noted above, the authors did not actually report how any of the effects were discovered, so there is no basis for saying that the discovery procedures were preregistered, free from p-hacking, rigorous, etc. What is true is that the *replication* attempts adopted these procedures. But this is presently true of most replication attempts in psychology—the majority of which nevertheless fail to replicate (see e.g., Scheel, Schijen, & Lakens, 2020, who found that only 31 of 71 preregistered reports successfully replicated the target finding). Thus, the difference between the authors' results and the prior literature cannot plausibly be explained by procedures used in the replications, and is more more likely to reflect differences in the procedures used to select effects for replication in the first place. But the authors do not indicate how the 16 effects were initially identified, so nothing can really be said about the discovery process.

MINOR CONCERNS
In addition to the above major concerns, I had a number of more minor concerns:

(6) The failure to model labs as random effects in the multi-level meta-analysis means that it is not possible to distinguish between-study variance from between-lab variance. This modeling decision is not unreasonable given practical constraints (i.e., the small samples and nested structure means the uncertainty surrounding these estimates would be very high), but the authors should explicitly acknowledge its implications in the manuscript—e.g., that if there are sizeable between-lab differences in how studies were selected for replication, the results might not be generalizable to the broader population of investigators (see also my related earlier point regarding the importance of understanding the study selection process).

(7) Perhaps I missed it, but given the strong emphasis on preregistration and its benefits, I found it a bit worrisome that the authors did not prominently link to a preregistration document for *this* project as a whole. In the supplement, the authors have a section titled "Pre-registered Analyses of Declining Effect Sizes", which suggests that a preregistration document exists, but I don't think a link is ever provided.

(8) The authors draw strong conclusions about differences between the present study and previous studies in terms of replication rate, but the basis for these conclusions is ultimately a set of just 16 effects. Unless I'm mistaken, no formal analysis is reported comparing the replicability estimate found for the present 16 analyses with the prior literature (is an 81% replication rate in 16 studies different from, say, a population baseline of 50%?). In view of the issues with generalization discussed above (point 2), I'm not sure that a formal comparison would be wise in any case, but it does seem worth noting that the relevant unit of analysis here is study, and it's unclear whether the authors would have been willing to make a similar claim in a case where the unit of analysis was, say, subject. I.e., would the authors have felt comfortable asserting, upon observing that 13 of 16 subjects show a given effect, that the observed rate of success is meaningfully different from prior studies that used different subjects and produced a lower estimate? I don't have a definitive answer, obviously, but I worry that there is something of a double standard in play, in that I'm not sure the authors' conclusions would seem very compelling if the unit of analysis were different, even though the logic would be identical.

Signed,

Tal Yarkoni

Referee #2 (Remarks to the Author):

The authors examine whether previously reported failures to replicate are “due to suboptimal implementation of optimal methods or whether presumptively optimal methods are not, in fact, optimal.” I think as a field we have developed a rather high prior that the suboptimal implementation of methods were the key player in the replication crisis. Nevertheless, it is extremely comforting to see that when these factors are removed, studies indeed replicate as one would expect. I found the study interesting, but have three issues that I think deserve some attention, related to the research question, the conceptualization of the underlying mechanism of a ‘decline effect’, and the reported results based on the percentage of significant results.

Main goal of the project

From the introduction, it does not become clear what the reasons for failure to replicate are when optimal methods would not be optimal, beyond that “such rates and declines are intrinsic to social-behavioral scientific investigation”. This is not really an explanation, or a mechanism. What would make failures to replicate ‘intrinsic’ to social-behavioral science? As this is the core reason for the authors to perform this extensive project, it is important to clarify what the goal actually is.

If the goal is to examine whether solid science should produce reliable effects, as measured based on certain statistical concepts such as significant results or accurate effect sizes in replication studies, in line with predictions by statistical theory, in the absence of an alternative hypothesis (i.e., without any reason to assume that findings would not replicate), then the study is a good demonstration of how scientific methods work in practice as it works in theory, but it is not clear why this should be surprising. The authors could have simulated a line of research, and observed the same results. Statistical theory predicts the outcome of this line of research, and as the authors tried their best to not violate any test assumptions (e.g., attempts to reduce heterogeneity between studies), I think most readers should have a high prior that this is what happens. I know that in the past some scientists have doubted whether psychology should be considered more like history (e.g., the work of Gergen comes to mind) but I think even Gergen would not have been surprised by seeing findings replicate across a short time span, under very similar conditions. The authors seem to have some reasons to have doubted that optimal methods are not optimal – but those reasons are not clearly communicated.

Alternatively, it is possible that the study was mainly a test of the 'decline effect'. When I looked through the OSF materials, study names included 'Decline Effect Wave 1', and in the supplement, mentioning that "The prospective replication project originated from a 2012 meeting about the decline effect organized by Jonathan Schooler at UC Santa Barbara and funded by the Fetzer Franklin Fund") makes me believe the main goal of these studies was, indeed to test a hypothesis about the decline effect in an extremely rigorous manner, but that the studies did not have the goal to test if failures to replicate are 'intrinsic' to psychology. If n examination of the decline effect was indeed the goal, I don't understand why the title is not something like "No evidence for a Decline Effect in Online Social Psychology Studies". This would also explain the vagueness in the alternative I mentioned in the previous 2 paragraphs. In that case, the work simply presents a null result of a what I personally believe to be a badly operationalized concept (see below), and the authors should make a stronger case why a decline effect would be theoretically predicted in the type of original and replication studies that they performed (that convinces people who have a high prior that there is no such thing as a decline effect).

Improve conceptualization

The concept of a 'decline effect' has not been developed sufficiently to make it useful in scientific papers. In general, when authors need to put a concept in quotes, such as the current authors do with the 'decline effect', it is best to remove it from a paper. The decline effect is, as far as I know after reading up the references following it's mention in the paper, not conceptualized any better than that 'effects decline'. The 'decline effect' is an explanandum, not the explanans – and it is the latter we are interested in.

The authors point out bad methods are problematic for reliable findings – but so is bad conceptualization of the topic under investigation, and the 'decline effect' is (after reading the references in the article) in my view a posited effect without clear explanations, that makes it unfalsifiable, and relatively uninteresting. I would recommend to replace it with better conceptualized terms, such as the selection of effects that are published based on their size, their confirmation of our beliefs, or statistical significance. The benefit of using the term selection effects is that we can quantify the bias through statistical selection models – commonly used to examine the effects of publication bias, low power, and p-hacking, which makes this terminology much closer to what the researchers are doing (removing selection effects in research lines).

The authors note in the discussion that "We did not observe a decline effect due to idiosyncrasies of different laboratory practices, different sampling conditions, or the passage of time." But sampling conditions did not differ substantially, the time window was very short, so how severely was heterogeneity tested? Furthermore, some effects should change over time – it's the mechanisms that should remain stable, but not all auxiliary assumptions remain constant across replications, and therefore, it is not clear how interesting it would even have been to show a decline in effects, without also showing *why* effects decline (e.g., which auxiliaries are no longer correct). As it stands, the ‘decline effect’ is, as Pauli would say, ‘not even wrong’.

Power analysis, and evaluation of results based on statistical significance

At several points in the paper the authors remark the performed studies with ‘high statistical power’ (e.g., abstract). But this is not true. We can see this in Figure 1 (which is an excellent figure). There are 2 significant studies (out of the 13) where we have reason to believe these 2 studies did not have high power. It was very difficult to find information about the sample size justification, and the decisions about which study was performed when for the study on in group favoritism: [https://osf.io/adrbe/.](https://osf.io/adrbe/) I can see there was a study 1, a study 2, and then a confirmatory study, and then a self-replication. I can not easily find information about the results in Study 1 and 2, but I see that in study 2 p-augmented is reported, as the sampling plan was not perfectly controlled. In the confirmatory study, the sampling plan was controlled. Here, we see a significant result in one sub-sample (the 2nd wave, although this depends a bit on whether the authors used a 0.005 alpha level or not, in which case the result might also not have been significant, depending on the rounding) but not in the first wave (regardless of the choice of alpha level). We see the three replications yielded very similar results, and together, there seems a non-zero effect. However, we can not assume that 1500 participants were sufficient. With 750 participants in each group, a sensitivity analysis shows the study had 80% power (already quite low) for a d = 0.13. If this confirmatory study happens to be a study where, as a fluke an effect size was observed that was rather extreme (as should happen, by change, in 16 studies), and the true effect was smaller (as it seems to be, in the replication studies), then this was not a sufficiently powered study. Instead, it *seemed to be* a sufficiently powered study, based on the *observed* effect size. But if we take the four replications as an estimate of the true effect size, the study had low power. Of course, all of this requires some speculation, as we never know the true effect size, but the point is, the authors can not argue the studies all had high power, and the absence of a power analysis (let alone a conservative power analysis, such as a safeguard power analysis) should make the authors even more careful about claiming they had ‘high power’. Power is a curve, and it is high for some effects, but low for other effects.
The FSD study clearly did not have very high power. The reported results in the published Psych Science paper show d = 0.142, 95% CI = [0.243, 0.04]. An observed power analysis for the observed effect size shows that for a one-sided test the power is 0.86. I should not need to point the attention to the 95% CI = [0.243, 0.04] to highlight that there is a very realistic possibility the true effect size is smaller than d = 0.14 (indeed, this happens in 50% of studies, on average). There is also some heterogeneity in these results it seems.

In general, using an observed power analysis is not best practice, but it seems the authors in these studies did not even perform a power analysis, and just seem to have thought that 1500 participants (750 in each group) was a large enough sample size not to worry. This seems true for most studies, but not for the two above. Statements such as “the average replication power was 0.96” are not very informative for reasons above, and the statement is also formally incorrect – the average power assuming the observed effect sizes are the true effect sizes is better not summarized by ‘replication power’ but by ‘observed power’ – that should make people more cautious. The statement “The observed replication rate of 90% is slightly smaller than expected based on these power estimates” might be true – but the rate is probably less surprising based on the observed meta-analytic power in the 4 replication studies, and they are also not surprising based on the true power, had we known it.

The bigger point here is that analysis of the percentage of statistically significant replication studies (lines 107 to 121) is relatively uninteresting. If authors had included a sample size justification based on powering all studies to have 95% power, than the proportion of significant results would have been interesting, as it would have followed from statistical theory and best practices in experimental design. However, the authors did not follow best practices in experimental design, and did not include a power analysis, and just sampled 1500 datapoints. Note that in general, with 13 or 16 studies, we can only very roughly approximate the expected long run results. In 13 studies, we do not have a 5% Type 1 error rate (because 0/13, 1/3, 2/13, etc do not round to 5%, see https://daniellakens.blogspot.com/2020/01/observed-alpha-levels-why-statistical.html for an explanation) and similarly, our *observed* power has a rather wide confidence interval in 13 (or even 52, if we include the replications) studies.

The interesting analysis is where the similarity in effect sizes is examined. After all, the sample size justification of 1500 observations, had it been written down, must have been that, irrespective of the power it would provide, the effect sizes were relatively accurate. This is true, and these comparisons are relevant and can be interpreted based on the experimental design. I would suggest the authors to only keep these analyses in the main document. These analyses are interesting, and with Figure 1, easy enough to interpret.

Here, we see some minor violations of the assumptions that there is a single true effect size. There is some heterogeneity in the ‘worse’ and ‘redemption’ studies, where the confirmatory study and the self-replication yield very consistent effect sizes, but the 3 replications yield smaller effect sizes. I did not see a good discussion why this heterogeneity exists. Some follow up studies to explain these differences would have been welcome. After all, the expectations that most effects will replicate is based on the idea that there is a homogenous true effect size. If this is violated, we should examine carefully why, as the expected success rates becomes more uncertain.

Minor points

In the supplemental file, it says “Overall, for 14 of the 48 (29%) independent replications there was no interaction between the originating lab and the replication lab beyond sharing the methods section and key materials.” I feel this is important enough to mention. The authors do not seem to suggest communication between labs is a necessary ‘best practice’ – but then why was there so much communication anyway? In one case, it says “Helped Stanford design materials and ran a pilot for them.” So, in some cases there really was a lot more assistance than just using the materials and code and the written report, as I got the impression in the main document. I am adding this as a relatively minor point, but I do feel this is quite important.

The authors nicely note the limitations in their studies (highly similar samples, methods, and limitations to between paradigms that can be done online). They note how not all online studies replicate – sure, but that is often due to p-hacking or publication bias, and online studies are by far the safest social psychological studies to use if you would *not* want to show a decline effect.

Spacing is absent in all statistical tests – maybe this is due to a word limit issue or Nature reporting guidelines, but it is a bit annoying.

There was no direct link to the decline effect analyses – but I found https://osf.io/rqwn2/ on the OSF, requested access, and received it. I would suggest to add a link to the analysis code and data of the meta-analyses somewhere in the document or supplemental file. The files in in RMarkdown, and I think these analyses are very well done. I did not check the analyses for sub-projects, but the main analyses reported in the paper are very transparently reported.

Signed,

Daniel Lakens

Referee #3 (Remarks to the Author):

Thank you for the opportunity to review this interesting work. I have some comments. I should also declare that, as a participant at the MetaScience2019 conference I participated in the survey of expected results for 12 of the 16 experiments described in the supplementary materials.

A. Summary of the key results

B. The authorship team have conducted an interesting study: 4 participating labs were asked to nominate 4 findings each. While these are described as relating to discovery related basic research, they all are based on the core methodology of internet based research using panels of volunteer participants, described later as “social behavioural scientific investigation”. First, each lab conducted a pre-registered test of their findings. Then – whether or not that initial confirmatory experiment found a significant effect – the finding was tested in 4 replication studies, 3 in other labs and one in the original lab. There are therefore 80 experiments, 16 self confirmatory, 16 “auto-replication” and 52 “allo-replication”. Essentially they find successful replication at around the rate expected according to the statistical power of their replication experiments, with no substantial differences between auto- and allo- replication attempts. They attribute this to the

rigor of the original research claims coupled with the rigor of the replication attempts. They go on to show no effects of whether investigators were blinded to emerging data during the conduct of the research, no effects of partitioning study populations and no pattern of changing effect sizes in the sequences of data collection.

C. Originality and significance: if not novel, please include reference
This is an important and novel set of findings. I am not aware of similar work conducted elsewhere.
D. Data & methodology: validity of approach, quality of data, quality of presentation
The approach is valid, and the quality of data and of its presentation are good. I have some suggestions for clarifications below.
E. Appropriate use of statistics and treatment of uncertainties
Good
F. Conclusions: robustness, validity, reliability
The conclusions are sound in as much as they relate to internet panel based research with very high statistical power. Their generalisability beyond this is unknown, and an interesting topic for future research. Specifically, the populations accessed are highly homologous (adults volunteering for internet based research), and while the providers are different between labs, it may be that participants in such research belong to multiple panels. I don't know enough about the field to know, but it would be good to know if their ~120000 participants (80 x 1500) could include the same individual contributing to different experiments, or to the same experiment conducted by different labs. That is, how truly independent are the samples? With regard to power, I don't think they present power in the self confirmatory experiments (l149 et seq is confusing – does this relate to the power, based on self confirmation, for the 16 replication efforts? Or the power in those studies to detect the effects (and then, is it against the observation in that study, or the ES observed in meta-analysis of the 5 studies?)). But if they are anything like the replication studies, the power of the self confirmatory study approaches 90%. This is virtually unheard of in many research domains, and limits the generalisability of the findings.
G. Suggested improvements: experiments, data for possible revision
No new experiments or analyses required

H. References: appropriate credit to previous work?
Appropriate

I. Clarity and context: lucidity of abstract/summary, appropriateness of abstract, introduction and conclusions
1. L56 effect sizes (ES)s … a little clumsy .. perhaps effect size(ES)s
2. L58 et seq .. in clinical measurement, regression to the mean is very well recognised, for instance in blood pressure recordings – "abnormally high" readings are likely, on retest, to be nearer the population mean (see for instance https://onlinelibrary.wiley.com/doi/full/10.1111/jch.13933). How does that play in the reproducibility debate?
3. L72 – readers of Nature will likely consider discovery orientated basic research as involving different approaches to those deployed here – your later description of social behaviour scientific investigation is I think more informative.

4. L72 et seq: Could you be clearer, earlier, that this is 4 x 4 self confirmations, then 4 x 4 auto-replications and 4 x 4 x 3 allo-replications? At first (and second) reading I didn't get this until later in the MS.
5. L107 et seq: Interestingly, for the 3/16 null self confirmatory experiments, one might consider a statistically significant effect in the hypothesised direction as a failure of replication rather than a success. I get (from the Supplementary materials) why you did it this way, but perhaps promote that argumentation to the main paper?
6. L129 – [tau hat within] is I think a measure of tau sq from the meta-analysis, and is I think a REML estimator, but not described in the supplement in detail.
7. L135 et seq: I consider a difference between judging successful replication on the basis (a) of 95% Cis overlapping (L137) versus (b) the point estimate in the replication lying within the 95% CI of the originator. Perhaps you could be clearer on this? Also, L138 could be clearer that the replication rates are smaller than if you take the criteria to be same direction of effect at <0.05.
8. L150: is this the average power in the attempted replication? Please clarify. Also, I think you mean the average of all self confirmatory, or "Considering all self confirmatory tests, the predicted power was .. "; and given 3 of these were null, the power of the confirmatory test was I guess zero, which explains the wide range.
9. Fig 2: in sequence, the top 4 panels place independent replication first, and the bottom panel places independent replication first.
10. Fig 2: the prediction interval appears wider for the independent replications. Presumably this is adding more information about the range of possible truths, compared with self replication; and so while the point estimate of ESs are pretty similar in replication (and so why bother), do we in fact get important new information by multicentre replications?
11. L263: I think there is a problem of logic here: the failure to replicate findings in different domains says nothing about whether remediation of the problem in your domain is easier, or more difficult, than it would be in these other domains.
12. L301: "only when the replicating labs" … I think you need to say here that this "only" happened 71% of the time, not bury it in the supplementary materials.
13. L362: You mention in the supplement the version of metafor, but not the version of R … this would be helpful.
14. General: it may be the formatting of the submission, but it would be helpful to include ORCID IDs.

I sign my reviews: Malcolm Macleod