nature research

Peer Review Information

Journal: Nature Human Behaviour
Manuscript Title: High Replicability of Newly-Discovered Social-behavioral Findings is Achievable
Corresponding author name(s): John Protzko

Editorial Notes:

Transferred manuscripts	This manuscript has been previously reviewed at another journal that is not operating a transparent peer review scheme. This document only contains reviewer comments, rebuttal and decision letters for versions considered at Nature Human Behaviour.
Redactions – transferred manuscripts (mention of previous referee reports from elsewhere)	This manuscript has been previously reviewed at another journal. This document only contains reviewer comments, rebuttal and decision letters for versions considered at Nature Human Behaviour. Mentions of prior referee reports have been redacted
Redactions – transferred manuscripts (mention of the other journal)	This manuscript has been previously reviewed at another journal. This document only contains reviewer comments, rebuttal and decision letters for versions considered at Nature Human Behaviour. Mentions of the other journal have been redacted.

Reviewer Comments & Decisions:

Decision Letter, initial version:

7th August 2023

Dear Dr. Protzko,

Thank you for your patience as we've prepared the guidelines for final submission of your Nature Human Behaviour manuscript, "High Replicability of Newly-Discovered Social-behavioral Findings is Achievable" (NATHUMBEHAV-23041313A). Please carefully follow the step-by-step instructions provided in the attached file, and add a response in each row of the table to indicate the changes that you have made. Please also address the additional marked-up edits we have proposed within the reporting summary.

Ensuring that each point is addressed will help to ensure that your revised manuscript can be swiftly handed over to our production team.

We would hope to receive your revised paper, with all of the requested files and forms within two-three weeks. Please get in contact with us if you anticipate delays.

When you upload your final materials, please include a point-by-point response to any remaining reviewer comments.

If you have not done so already, please alert us to any related manuscripts from your group that are under consideration or in press at other journals, or are being written up for submission to other journals (see: https://www.nature.com/nature-research/editorial-policies/plagiarism#policy-on-duplicate-publication for details).

Nature Human Behaviour offers a Transparent Peer Review option for new original research manuscripts submitted after December 1st, 2019. As part of this initiative, we encourage our authors to support increased transparency into the peer review process by agreeing to have the reviewer comments, author rebuttal letters, and editorial decision letters published as a Supplementary item. When you submit your final files please clearly state in your cover letter whether or not you would like to participate in this initiative. Please note that failure to state your preference will result in delays in accepting your manuscript for publication.

In recognition of the time and expertise our reviewers provide to Nature Human Behaviour's editorial process, we would like to formally acknowledge their contribution to the external peer review of your manuscript entitled "High Replicability of Newly-Discovered Social-behavioral Findings is Achievable". For those reviewers who give their assent, we will be publishing their names alongside the published article.

Cover suggestions

As you prepare your final files we encourage you to consider whether you have any images or illustrations that may be appropriate for use on the cover of Nature Human Behaviour.

Covers should be both aesthetically appealing and scientifically relevant, and should be supplied at the best quality available. Due to the prominence of these images, we do not generally select images featuring faces, children, text, graphs, schematic drawings, or collages on our covers.

We accept TIFF, JPEG, PNG or PSD file formats (a layered PSD file would be ideal), and the image should be at least 300ppi resolution (preferably 600-1200 ppi), in CMYK colour mode.

If your image is selected, we may also use it on the journal website as a banner image, and may need to make artistic alterations to fit our journal style.

Please submit your suggestions, clearly labeled, along with your final files. We'll be in touch if more information is needed.

ORCID

Non-corresponding authors do not have to link their ORCIDs but are encouraged to do so. Please note that it will not be possible to add/modify ORCIDs at proof. Thus, please let your co-authors know that if they wish to have their ORCID added to the paper they must follow the procedure described in the following link prior to acceptance:
https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research

Nature Human Behaviour has now transitioned to a unified Rights Collection system which will allow our Author Services team to quickly and easily collect the rights and permissions required to publish your work. Approximately 10 days after your paper is formally accepted, you will receive an email in providing you with a link to complete the grant of rights. If your paper is eligible for Open Access, our Author Services team will also be in touch regarding any additional information that may be required to arrange payment for your article.

Please note that *Nature Human Behaviour* is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

For information regarding our different publishing models please see our Transformative Journals page. If you have any questions about costs, Open Access requirements, or our legal forms, please contact

ASJournals@springernature.com.

Please use the following link for uploading these materials:
[REDACTED]

If you have any further questions, please feel free to contact me.

Best regards,
Alex McKay
Editorial Assistant
Nature Human Behaviour

On behalf of

Samantha Antusch

Samantha Antusch, PhD
Senior Editor
Nature Human Behaviour

Reviewer #1:
Remarks to the Author:
Thank you: The authors have responded appropriately to my previous comments

Reviewer #2:
Remarks to the Author:
This manuscript describes a meta-scientific study involving multiple labs across the social and behavioral sciences first nominating, then self-replicating, and then and replicating each other's studies. The replication rate was relatively high (compared to previous replication efforts in the psychological literature), the authors conclude that rigor enhancing methods should be implemented on a larger scale to increase the replicability of new discoveries.

Before I turn to my review, I will add that I was only invited as a reviewer in round 2. While I did not read the originally submitted version of the manuscript, I read the authors' response letter, which I includes three reviews with comments on the original version, and the author replies. I personally find the role of

"rear review" difficult, but this discussion is for another time. In brief, I do not think it is my place to comment on whether the authors have satisfactorily addressed the remarks by the original reviewers, nor to introduce entirely new points previously not raised. Instead, my comments build on earlier points, and hope they support the decision making process by the action editor.

DECLINE

As has already been remarked on by reviewers #1 and #2, it is not very obvious how a "decline effect" could possibly have been observed in the data presented. Putting aside cosmic forces beyond our comprehension, two sensible explanations for a decline effect remain: (1) that the phenomenon (e.g., the strength of a correlation between two variables) itself actually changes from earlier studies to later studies, (2) that the strength of selection of studies into the published literature based on observed effect sizes is stronger in earlier studies than later studies. Although I am sure we can find examples of both, the cause of psychology's low replicability is probably mostly the latter, and in essence, describes publication bias followed by regression to the mean. However, both explanations cannot possibly be an "inevitable feature" (l. 220) of the replication chain reported in this manuscript:

RE: (1) Yes, all effects in the social and behavioral sciences are beautiful flowers* that will wither one day**, but as the studies were conducted within a short timeframe, substantive changes to the phenomena explored seem extremely unlikely.

RE: (2) We know there was no publication bias because publication status did not play a role in this prospective design. In their response letter, the authors state that they retain the term "decline" as they "focus on evaluating whether replication effect sizes are smaller than original findings". But they provide no argument whatsoever what makes replication studies different from original studies (in an unbiased literature!). Here, I find it useful to borrow Gelman's "time reversal heuristic": Surely, the authors would not suggest that it had been plausible to expect an elevation of effect sizes in the counterfactual, i.e. that the replication studies in their design had been conducted before the self-confirmation studies (which would have been possible given they were based on the same preregistration)? This is exactly because their population of 16 self-confirmation studies, unlike "the literature", was not selected for effect size, and is essentially unbiased (more on this below).

Again, as reviewer #1 remarked, the only opportunity for systematic decline would be between the selected original effects and all subsequent confirmations and replications, and like him, I am somewhat puzzled why these are not included in the main paper (including all figures) given their obvious relevance to the authors' arguments.

SELECTION AND "THE LITERATURE"

My second major point regards the selection of original studies for the replication chain. I will not reiterate remarks by reviewer #1 on generalizability issues arising from the nature of the research

questions pursued by the author teams as well as the diversity of stimulus materials. Instead, my concern regards the extraordinariness of the authors themselves.

An important part of the argument for the benefits of rigor enhancing methods for replicability are comparisons of the authors' findings to study characteristics observed in "the literature" (average ESs, replication rates, etc). However, the authors do not at all consider that the research they produce is dramatically different from "the literature". Protzko et al., each in their own way, are role models of behavioral researchers with dramatically positive effects on "the literature", and, if I may say so, my own research practices. They present a strong case for these rigor enhancing methods in replication research, but they do not consider how – in my expectation at least – much more rigorous their original research is compared to "the literature". Case in point: They set out to write a paper reporting the results from 80 replications.

The authors report other evidence for this too. Although they regrettably omit details from the original studies in the main manuscript, the positivity rate of the self-confirmatory studies was 81% (l. 106). This is exceptionally, crazy high, and we know this is from, e.g. Scheel et al., who observed a much lower positivity rate, even in registered reports, in "the literature". Similarly, they highlight that the ESs in self-confirmatory tests were much higher than ESs observed in prior systematic replication attempts of "the literature". A reverse decline effect after all?

No, the authors are probably just better scientists than the average person contributing to "the literature". I understand that modesty forbids making this point with these words, but it would probably truthful to state that the authors use a lot of the suggested rigor enhancing methods not only in replications, but in their original research too, which of course dramatically increases the confirmation and replication success rate. I honestly found it a bit amusing when they described the confirmatory tests as "free from p-hacking or questionable research practices, unlikely to be artifacts of low statistical power, and fully documented" (l. 207-209), which is exactly true – but I would bet that the same is true for the authors' original studies.

In other words: It is easy (relatively speaking) to obtain high replicability when the original research is of high quality. This point might also deserve some further appreciation: There are several papers in the literature trying to predict replication success from various study characteristics (such as discipline, research design, sample size, measurement validity, and so forth), using data from large-scale replication efforts (RP:P, ManyLabs, etc). One recurring issue that these meta-studies face is the difficulty of predicting a signal when the rate of replication failures is high. The authors, relying on their own original research, have created the opposite problem: When basically every replication attempt is a success, it is difficult to identify predictive study characteristics or research practices.

Given these points, I am not convinced of the value of the two survey studies they report. The authors

observe that both researchers and laypeople correctly predict the direction and significance of findings in approximately 40% of the cases, and consider this evidence that "the sample of discoveries used here were not biased towards yielding high replication rates" (p. 249-250). But they were! We know this! The replication rate was over 80%!!! The authors cannot possibly suggest that this was an accident, and to counterbalance this rate, their next 16 studies will all fail to replicate just to meet what the average researcher or layperson thinks of the literature.

I would like to explore two possible explanations here, and both have to do with the stimulus materials of the synopses too:

One is, again, the "exceptionality hypothesis". It might be true that, to the average researcher, "the sample of discoveries used here […] were neither more obvious or predictable than similar findings with low replication rates" (l. 249-251) – but they were more obvious to the authors because I can only assume the authors don't randomly pick things they study (and that they do not study) and then pick for a large-scale replication attempt. Now, from what I was able to puzzle together from the OSF repository, the synopses merely described the research question and basic design. But I bet had they also highlighted that these studies were conducted by Protzko, Krosnick, Nelson, Nosek, etc. and not, for example, Malte Elson, there would have been a lot fewer null predictions. Had they further provided the authors' predictions, even more researchers would have agreed with them. And they should, because these authors are more trustworthy than "the literature", which is the reference category that in absence of any other details the researchers in the sample probably used.

Even leaving aside the author details, information regarding the rigor of the original studies were also absent. I would assume it makes a difference to the study participants to know that the original studies might have been preregistered, that the sample size was large, etc. Again, in absence of these information, they probably used the average practices from "the literature" as a reference.

MINOR REMARKS

"81% (13/16) of the self-confirmatory tests produced statistically-significant results (d=0.265 (95%CI=0.165-0.37; with 0.18SD estimated between-study heterogeneity)." (l. 106-107)

Does this mean that 16/16 of the original findings were significant, or were there nonsignificant findings too that entered the self-confirmatory test phase?

"There was modest evidence that one lab produced slightly smaller ESs in replications compared to one other lab, controlling for the average size of original effects from each lab (robust Approximate Hotelling's T2 (12.31)=3.51, p=.048)"

I'm not sure I quite understand why the original effects were included here, and only here. Maybe the authors can explain this in the SOM?

Signed,
Malte Elson

*except Stroop
**except Stroop

Author Rebuttal to Initial comments

Reviewer #1 (Remarks to the Author):

Thank you: The authors have responded appropriately to my previous comments

Reviewer #2 (Remarks to the Author):

This manuscript describes a meta-scientific study involving multiple labs across the social and behavioral sciences first nominating, then self-replicating, and then and replicating each other's studies. The replication rate was relatively high (compared to previous replication efforts in the psychological literature), the authors conclude that rigor enhancing methods should be implemented on a larger scale to increase the replicability of new discoveries.

Before I turn to my review, I will add that I was only invited as a reviewer in round 2. While I did not read the originally submitted version of the manuscript, I read the authors' response letter, which I includes three reviews with comments on the original version, and the author replies. I personally find the role of "rear review" difficult, but this discussion is for another time. In brief, I do not think it is my place to comment on whether the authors have satisfactorily addressed the remarks by the original reviewers, nor to introduce entirely new points previously not raised. Instead, my comments build on earlier points, and hope they support the decision making process by the action editor.

DECLINE

As has already been remarked on by reviewers #1 and #2, it is not very obvious how a "decline effect" could possibly have been observed in the data presented. Putting aside cosmic forces beyond our comprehension, two sensible explanations for a decline effect remain: (1) that the phenomenon (e.g., the strength of a correlation between two variables) itself actually changes from earlier studies to later studies, (2) that the strength of selection of studies into the published literature based on observed effect sizes is stronger in earlier studies than later studies. Although I am sure we can find examples of both, the cause of psychology's low replicability is probably mostly the latter, and in essence, describes publication bias followed by regression to the mean. However, both explanations cannot possibly be an "inevitable feature" (l. 220) of the replication chain reported in this manuscript:

RE: (1) Yes, all effects in the social and behavioral sciences are beautiful flowers* that will wither one day**, but as the studies were conducted within a short timeframe, substantive changes to the phenomena explored seem extremely unlikely.

RE: (2) We know there was no publication bias because publication status did not play a role in this prospective design. In their response letter, the authors state that they retain the term "decline" as they "focus on evaluating whether replication effect sizes are smaller than original findings". But they provide no argument whatsoever what makes replication studies different from original studies (in an unbiased literature!). Here, I find it useful to borrow Gelman's "time reversal heuristic": Surely, the authors would not suggest that it had been plausible to expect an elevation of effect sizes in the counterfactual, i.e. that the replication studies in their design had been conducted before the self-confirmation studies (which would have been possible given they were based on the same preregistration)? This is exactly because their population of 16 self-confirmation studies, unlike "the literature", was not selected for effect size, and is essentially unbiased (more on this below).

Again, as reviewer #1 remarked, the only opportunity for systematic decline would be between the selected original effects and all subsequent confirmations and replications, and like him, I am somewhat puzzled why these are not included in the main paper (including all figures) given their obvious relevance to the authors' arguments.

We thank Dr. Elson for this comment. As we outlined in the prior interaction with the reviewers, we share the philosophical point, but aim to empirically address a widespread conception here that declining effect sizes and non-replicability are inevitable features of the scientific enterprise when using NHST. Publication bias is a substantial part of those arguments, but so is context sensitivity that replicators do not attend to in conducting replications, tacit knowledge that is not present in replications, and other factors that our "best practice" approach aimed to address. We hope to have addressed this along with citing such articles on p. 13-14:

"We investigated whether low replicability and declining ESs should be expected from the social-behavioral sciences because of the complexity of the phenomena, hidden moderators18, and other factors that might be intrinsic to the phenomena being studied or the replication process[24,25]—instead, we found a high replicability rate. The present results are reassuring about the effectiveness of what we think of as best practices in scientific investigations. When novel findings were transparently subjected to preregistered, large-sample confirmatory tests—and when replications involved similar materials and were implemented with a commitment to faithfulness to testing the same hypothesis with fidelity to the original procedure—the observed rate of replication was high. Furthermore, we saw no statistically-significant evidence of declining effect sizes over replications, either when holding materials, procedures, and sample source constant (except for sampling error) or when materials and procedures and sample sources varied but were faithful to the original studies."

And on p. 15-16, where we also agree that it is a good suggestion to cite Gelman's time-reversal heuristic (citation 46):

"It is 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. First, rather than using exploratory discoveries as the basis for claiming a finding, all discoveries were subjected to preregistered self-confirmatory tests. This eliminated inflation of false positives and ESs by pre-commitment to research designs and analysis plans29. Second, once a discovery was submitted for a self-confirmatory test, we committed to reporting the outcomes. This eliminated publication bias that is particularly pernicious when selective reporting of study findings systematically ignores null results[30-31, 46]. Third, all self-confirmatory tests and replications were conducted with large sample sizes (Ns ≳ 1500), resulting in relatively precise estimates. Fourth, each lab was part of the process of both discovering and replicating findings. This may have motivated teams to be especially careful in both characterizing their methods and in carrying out their replications. Fifth, if there were essential specialized materials for the

experimental design, the discovering lab made them available as supplementary materials. Sharing original materials should increase understanding and adherence to critical features of original experimental methodologies. We expect that all of these features contributed to improving replicability to varying degrees. Future investigations could manipulate these features to learn more about their causal contribution to replicability."

SELECTION AND "THE LITERATURE"

My second major point regards the selection of original studies for the replication chain. I will not reiterate remarks by reviewer #1 on generalizability issues arising from the nature of the research questions pursued by the author teams as well as the diversity of stimulus materials. Instead, my concern regards the extraordinariness of the authors themselves.

An important part of the argument for the benefits of rigor enhancing methods for replicability are comparisons of the authors' findings to study characteristics observed in "the literature" (average ESs, replication rates, etc). However, the authors do not at all consider that the research they produce is dramatically different from "the literature". Protzko et al., each in their own way, are role models of behavioral researchers with dramatically positive effects on "the literature", and, if I may say so, my own research practices. They present a strong case for these rigor enhancing methods in replication research, but they do not consider how – in my expectation at least – much more rigorous their original research is compared to "the literature". Case in point: They set out to write a paper reporting the results from 80 replications.

The authors report other evidence for this too. Although they regrettably omit details from the original studies in the main manuscript, the positivity rate of the self-confirmatory studies was 81% (l. 106). This is exceptionally, crazy high, and we know this is from, e.g. Scheel et al., who observed a much lower positivity rate, even in registered reports, in "the literature". Similarly, they highlight that the ESs in self-confirmatory tests were much higher than ESs observed in prior systematic replication attempts of "the literature". A reverse decline effect after all?

No, the authors are probably just better scientists than the average person contributing to "the literature". I understand that modesty forbids making this point with these words, but it would probably truthful to state that the authors use a lot of the suggested rigor enhancing methods not only in replications, but in their original research too, which of course dramatically increases the confirmation and replication success rate. I honestly found it a bit amusing when they described the confirmatory tests as "free from p-hacking or questionable research practices, unlikely to be artifacts of low statistical power, and fully documented" (l. 207-209), which is exactly true – but I would bet that the same is true for the authors' original studies.

In other words: It is easy (relatively speaking) to obtain high replicability when the original research is of high quality. This point might also deserve some further appreciation: There are several papers in the literature trying to predict replication success from various study characteristics (such as discipline, research design, sample size, measurement validity, and so forth), using data from large-scale replication efforts (RP:P, ManyLabs, etc). One recurring issue that these meta-studies face is the difficulty of predicting a signal when the rate of replication failures is high. The authors, relying on their own original research, have created the opposite problem: When basically every replication attempt is a success, it is difficult to identify predictive study characteristics or research practices.

Given these points, I am not convinced of the value of the two survey studies they report. The authors observe that both researchers and laypeople correctly predict the direction and significance of findings in approximately 40% of the cases, and consider this evidence that "the sample of discoveries used here were not biased towards yielding high replication rates" (p. 249-250). But they were! We know this! The replication rate was over 80%!!! The authors cannot possibly suggest that this was an accident, and to counterbalance this rate, their next 16 studies will all fail to replicate just to meet what the average researcher or layperson thinks of the literature.

I would like to explore two possible explanations here, and both have to do with the stimulus materials of the synopses too:

One is, again, the "exceptionality hypothesis". It might be true that, to the average researcher, "the sample of discoveries used here […] were neither more obvious or predictable than similar findings with low replication rates" (l. 249-251) – but they were more obvious to the authors because I can only assume the authors don't randomly pick things they study (and that they do not study) and then pick for a large-scale replication attempt. Now, from what I was able to puzzle together from the OSF repository,

the synopses merely described the research question and basic design. But I bet had they also highlighted that these studies were conducted by Protzko, Krosnick, Nelson, Nosek, etc. and not, for example, Malte Elson, there would have been a lot fewer null predictions. Had they further provided the authors' predictions, even more researchers would have agreed with them. And they should, because these authors are more trustworthy than "the literature", which is the reference category that in absence of any other details the researchers in the sample probably used.

We appreciate the reviewer's praise of our characteristics as researchers, we aspire to earning it. It is true that we are not going to claim in the paper that we achieve high replicability because we are great researchers, not just because of modesty, but also because we don't think such a claim is particularly an alternative explanation for our findings. However, Dr. Elson does point out an important feature of our original research that is not detailed in the paper -- that it is likely conducted with high rigor as well. Indeed, we believe that it was as we have strong cultures in our labs and expectations of careful, rigorous work. Because we don't have the direct evidence to provide, we don't make a strong claim in this regard, but we did revise the paper to point out that even in the original research, our labs have strong norms for rigor--a feature that is achievable by anyone, not just us. We address this on p. 15, complete with a citation to a paper of some of ours showing traditional metrics of researcher expertise are unrelated to whether one can actually replicate a study.

"It is also possible to imagine we observed higher replicability than other investigations because of the qualities of the researchers involved in this project, such as being better at imagining and discovering new, replicable phenomena. While we could be motivated to believe this possibility, the PIs in this project all have direct experience with their own published findings failing to replicate. Also, in this, and other research, the participating labs have established practices of making risky predictions, most of which fail to materialize into reliable phenomena. If there is an investigator influence on the observed findings, we believe that it is aligned with our interpretation of the present evidence as being due to the adoption of rigor-enhancing practices as lab norms rather than individual exceptionality.[45]"

One more minor point, we did not address Dr. Elson's last point that survey respondents might have increased their estimates of replicability if we had attached the names of the researchers to the hypothesized findings. While an interesting idea, basing assessments of credibility on reputation rather than the plausibility of the claims themselves is a separate research question than what we investigated, and the purpose of the paper. The relevant section summarizing our

purpose and approach is on p. 14-15:

"An uninteresting reason for high replicability would be if the discoveries, although novel, are obviously true. Trivial findings might be particularly easy to replicate. To assess this, we conducted two additional studies (see SI for additional details). In the first, 72 researchers reviewed a synopsis of most of the research designs and predicted the direction of each finding. On average, raters correctly predicted the direction and significance of the self-confirmatory tests 42% of the time, incorrectly predicted null results 38% of the time, and incorrectly predicted the direction of the findings 20% of the time. In the second, 1,180 participants reviewed synopses of the research designs from this study that showed high replicability and from a prior study of published findings from the same fields with similar methodologies that showed low replicability[6, 8-9]. The synopses were generated by independent researchers with experience in this design. On multiple preregistered criteria, participants were no better at predicting the outcomes of the highly replicable discoveries presented here ($M_{present\ studies}$ = 40.7% correct prediction) than at predicting the other less replicable findings from the prior investigation ($M_{comparison\_studies}$ = 41.1%, = -1.65, $p < .001$ for a pre-registered equivalence test of H_0:). Notably, the average accuracy rate of researchers in the first study was nearly identical to the average accuracy among lay people in the second. Additionally, the accuracy of predictions for specific findings was significantly associated with the absolute magnitude of the average ESs from independent replications ($b = 2.79$, z = 2.95, $p = .003$ for the findings in the present study; $b = 0.66$, z = 3.05, $p = .002$ for the comparison findings); absolute ES explained 35% of the variance in predictability rates. These findings indicate the sample of discoveries used here were not of a prima facie

different type of content that would yield high replication rates. Nor were the content or hypotheses more obvious or predictable than similar findings with low replication rates."

Even leaving aside the author details, information regarding the rigor of the original studies were also absent. I would assume it makes a difference to the study participants to know that the original studies might have been preregistered, that the sample size was large, etc. Again, in absence of these information, they probably used the average practices from "the literature" as a reference.

We believe the way we wrote the original manuscript may lead to a belief that each team only ran one pilot study, put it forward for replication, then was successful the overwhelming majority of the time. The fault is ours if that is the reading. We have made additions to the manuscript throughout, most importantly on p. 4-5:

"Four laboratories conducting discovery-oriented social-behavioral research participated in a prospective replication study. Over five years, the labs conducted their typical research, examining topics covering psychology, marketing, advertising, political science, communication, and judgment and decision-making (Table 1). Each lab engaged in pilot testing of new effects based on their laboratories business as usual practices. These practices could involve collecting data with different sample providers, and with any sample size the lab saw fit. All pilots were required to have materials, procedure, hypotheses, analysis plan, and exclusions preregistered prior to data collection. The main criterion for moving from piloting and exploration into the confirmation and replication protocol was the lab believed they had discovered a new effect that was statistically distinguishable from zero during the piloting phase. Each of the four labs submitted four new candidate discoveries for a self-confirmatory test and four replications, for a total of 16 confirmatory tests and 64 replications. In the self-confirmatory test, the discovering lab

conducted a preregistered study with a large sample (N≳1500) and shared a report of the methodology. Regardless of the outcome of the self-confirmatory test, in the replication phase, all labs conducted independent preregistered replications using the written methodology and any specialized study materials shared by the discovering lab (e.g., videos constructed for delivering interventions). Ordinarily, we would promote strong communication between labs to maximize sharing of tacit knowledge about the methodology, but in this case, to maintain the independence of each replication, we opted to discourage communication with the discovering lab outside of the documented protocols except for critical methodology clarifications (see SI). The replicating labs used equally large sample sizes (N's≳1500) and each lab used a different sample provider.

Preregistration, reporting all outcomes, large sample sizes, transparent archiving, and sharing of materials, and commitment to high fidelity replication procedures should minimize irreplicability or declining effect sizes stemming from questionable research practices, selective reporting, low-powered research, or poorly implemented replication procedures. Such optimizing might promote higher replicability than previously reported in the literature. If—despite these rigor-enhancing practices—low replicability rates or declining effects are observed—such rates and/or declines could be intrinsic to social-behavioral scientific investigation [24,25,40-43].

Each of the 16 ostensible discoveries were obtained through pilot and exploratory research conducted independently in each laboratory. Not every pilot study the labs conducted was put forward for confirmation and replication. Like all exploratory research, labs sometimes found errors, did not find signals of potential effects, or just lost interest in pursuing it further. 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 confirmatory tests, replications, and analyses were preregistered in both the individual studies (see Table S2) and for this meta-project (https://osf.io/6t9vm)."

MINOR REMARKS

"81% (13/16) of the self-confirmatory tests produced statistically-significant results (d=0.265 (95%CI=0.165-0.37; with 0.18SD estimated between-study heterogeneity)." (l. 106-107)

Does this mean that 16/16 of the original findings were significant, or were there nonsignificant findings too that entered the self-confirmatory test phase?

We hope we have made it more clear now on p. 5 that 3 of the studies put forward did not produce statistically significant results during the Confirmation testing:

"Of the 16 studies put forward for replication, 81% (13/16) produced statistically-significant results during the confirmation phase (d = 0.265 (95%CI = 0.165 to 0.37; with 0.18SD estimated between-study heterogeneity). The average ES of the self-confirmatory tests was smaller than the estimated average ES of the published psychological literature (d = 0.43)[27], even when only considering the 13 statistically-significant findings (d = 0.32). No lab produced self-confirmatory tests with larger average ESs than the other labs (robust Approximate Hotelling's $T^2(6.01) = 0.6$, $p = .638$)."

"There was modest evidence that one lab produced slightly smaller ESs in replications compared to one other lab, controlling for the average size of original effects from each lab (robust Approximate Hotelling's T2 (12.31)=3.51, p=.048)"

I'm not sure I quite understand why the original effects were included here, and only here. Maybe the authors can explain this in the SOM?

This was simply sloppy language on our part. We were not referring to the original exploratory findings from each lab, but rather to the size of effects in the initial self-confirmatory tests, as described above. We have revised the text to clarify this on p. 8-9:

Within-study heterogeneity across replications was estimated to 0.06SD, suggesting little heterogeneity overall, despite 75% of the replications being conducted independently using different sample providers. There was modest evidence that one lab produced slightly smaller ESs in replications compared to one other lab, controlling for the average size of effects in the initial self-confirmatory tests from each lab (robust Approximate Hotelling's T^2 (12.31) = 3.51, p = .048).

Signed,

Malte Elson

*except Stroop

**except Stroop

Final Decision Letter:

Message:

Dear Dr Protzko,

We are pleased to inform you that your Article "High Replicability of Newly-Discovered Social-behavioral Findings is Achievable", has now been accepted for publication in Nature Human Behaviour.

Please note that Nature Human Behaviour is a Transformative Journal (TJ). Authors may publish their research with us through the traditional subscription access route or make their paper immediately open access through payment of an article-processing charge (APC). Authors will not be required to make a final decision about access to their article until it has been accepted. Find out more about Transformative Journals

Authors may need to take specific actions to achieve compliance with funder and institutional open access mandates. If your research is supported by a funder that requires immediate open access (e.g. according to Plan S principles) then you should select the gold OA route, and we will direct you to the compliant route where possible. For authors selecting the subscription publication route, the journal's standard licensing terms will need to be accepted, including self-archiving policies. Those licensing terms will supersede any other terms that the author or any third party may assert apply to any version of the manuscript.

Once your manuscript is typeset and you have completed the appropriate grant of rights, you will receive a link to your electronic proof via email with a request to make any corrections within 48 hours. If, when you receive your proof, you cannot meet this deadline, please inform us at rjsproduction@springernature.com immediately. Once your paper has been scheduled for online publication, the Nature press office will be in touch to confirm the details.

Acceptance of your manuscript is conditional on all authors' agreement with our publication policies (see http://www.nature.com/nathumbehav/info/gta). In particular your manuscript must not be published elsewhere and there must be no announcement of the work to any media outlet until the publication date (the day on which it is uploaded onto our web site).

If you have posted a preprint on any preprint server, please ensure that the preprint details are updated with a publication reference, including the DOI and a URL to the published version of the article on the journal website.

An online order form for reprints of your paper is available at <a

href="https://www.nature.com/reprints/author-reprints.html">https://www.nature.com/reprints/author-reprints.html. All co-authors, authors' institutions and authors' funding agencies can order reprints using the form appropriate to their geographical region.

We welcome the submission of potential cover material (including a short caption of around 40 words) related to your manuscript; suggestions should be sent to Nature Human Behaviour as electronic files (the image should be 300 dpi at 210 x 297 mm in either TIFF or JPEG format). Please note that such pictures should be selected more for their aesthetic appeal than for their scientific content, and that colour images work better than black and white or grayscale images. Please do not try to design a cover with the Nature Human Behaviour logo etc., and please do not submit composites of images related to your work. I am sure you will understand that we cannot make any promise as to whether any of your suggestions might be selected for the cover of the journal.

You can now use a single sign-on for all your accounts, view the status of all your manuscript submissions and reviews, access usage statistics for your published articles and download a record of your refereeing activity for the Nature journals.

To assist our authors in disseminating their research to the broader community, our SharedIt initiative provides you with a unique shareable link that will allow anyone (with or without a subscription) to read the published article. Recipients of the link with a subscription will also be able to download and print the PDF.

As soon as your article is published, you will receive an automated email with your shareable link.

In approximately 10 business days you will receive an email with a link to choose the appropriate publishing options for your paper and our Author Services team will be in touch regarding any additional information that may be required.

If you have any questions about our publishing options, costs, Open Access requirements, or our legal forms, please contact ASJournals@springernature.com

We look forward to publishing your paper.

[redacted]

P.S. Click on the following link if you would like to recommend Nature Human Behaviour to your librarian http://www.nature.com/subscriptions/recommend.html#forms

** Visit the Springer Nature Editorial and Publishing website at <a

href="http://editorial-jobs.springernature.com?utm_source=ejP_NHumB_email&utm_medium=ejP_NHumB_email&utm_campaign=ejp_NHumB">www.springernature.com/editorial-and-publishing-jobs for more information about our career opportunities. If you have any questions please click here.**