Science AMA Series: We are authors of "Estimating the Reproducibility of Psychological Science" coordinated by the Center for Open Science AUA

CenterForOpenScience [1] and r/Science AMAs[1]

[1]Affiliation not available

April 17, 2023

Abstract

Last Thursday, our article "Estimating the Reproducibility of Psychological Science" was published in Science. Coordinated by the Center for Open Science, we conducted 100 replications of published results in psychology with 270 authors and additional volunteers. We observed a substantial decline effect between the original result and the replications. This community-driven project was conducted transparently, and all data, materials, analysis code, and reports are available openly on the Open Science Framework. Ask us anything about our process and findings from the Reproducibility Project: Psychology, or the initiatives to improve transparency and reproducibility in science more generally. We will be back at 12pm EDT (9 am PT, 4 pm UTC), AUA! Responding are: Brian Nosek, Center for Open Science & University of Virginia Johanna Cohoon, Center for Open Science Mallory Kidwell, Center for Open Science [EDITED BELOW] Some links for context: PDF of the paper: http://www.sciencemag.org/content/349/6251/aac4716.full.pdf OSF project page with data, materials, code, reports, and supplementary information: https://osf.io/ezcuj/wiki/home/ Open Science Framework: http://osf.io/ Center for Open Science: http://cos.io/ TOP Guidelines: http://cos.io/top/ Registered Reports: https://osf.io/8mpji/wiki/home/ 12:04. Hi everyone! Mallory, Brian, and Johanna here to answer your questions! 12:45. Our in house statistical consultant, Courtney Soderberg, has joined us in responding to your methodological and statistical questions. 3:50. Thanks everyone for all your questions! We're closing up shop for the holiday weekend but will check back in over the next few days to give a few more responses. Thanks to all the RPP authors who participated in the discussion!

the WINNOWER

REDDIT

Science AMA Series: We are authors of "Estimating the Reproducibility of Psychological Science" coordinated by the Center for Open Science AUA

CENTERFOROPENSCIENCE R/SCIENCE

ABSTRACT

Last Thursday, our article "Estimating the Reproducibility of Psychological Science" was published in Science. Coordinated by the Center for Open Science, we conducted 100 replications of published results in psychology with 270 authors and additional volunteers. We observed a substantial decline effect between the original result and the replications. This community-driven project was conducted transparently, and all data, materials, analysis code, and reports are available openly on the Open Science Framework.

Ask us anything about our process and findings from the Reproducibility Project: Psychology, or the initiatives to improve transparency and reproducibility in science more generally.

We will be back at 12pm EDT (9 am PT, 4 pm UTC), AUA!

Responding are:

Brian Nosek, Center for Open Science & University of Virginia

Johanna Cohoon, Center for Open Science

Mallory Kidwell, Center for Open Science

[EDITED BELOW] Some links for context:

PDF of the paper: http://www.sciencemag.org/content/349/6251/aac4716.full.pdf

OSF project page with data, materials, code, reports, and supplementary information: https://osf.io/ezcuj/wiki/home/

Open Science Framework: http://osf.io/

Center for Open Science: http://cos.io/

TOP Guidelines: http://cos.io/top/

Registered Reports: https://osf.io/8mpji/wiki/home/

12:04. Hi everyone! Mallory, Brian, and Johanna here to answer your questions!

12:45. Our in house statistical consultant, Courtney Soderberg, has joined us in responding to your methodological and statistical questions.

3:50. Thanks everyone for all your questions! We're closing up shop for the holiday weekend but will check back in over the next few days to give a few more responses. Thanks to all the RPP authors who participated in the discussion!

CORRESPONDENCE:

DATE RECEIVED:
September 05, 2015

DOI:
10.15200/winn.144136.67798

ARCHIVED:

In the wake of your study, do you think the problem is in the experimental methods preferred by psychology, or in the criteria for publication used by major journals (and the role such publications play in tenure)? What kind of reform would you propose to the field of psychology?

ParkerAdderson

Great questions. There are many factors contributing to the challenges of reproducibility.

The challenge is this. My lab does a lot of studies, only a subset of those get published. The ones that

September 04, 2015

CITATION: CenterForOpenScience , r/Science , Science AMA Series: We are authors of "Estimating the Reproducibility of Psychological Science" coordinated by the Center for Open Science AUA, *The Winnower* 2:e144136.67798 , 2015 , DOI: 10.15200/winn.144136.67798

are more likely to get published obtain significant results, show something novel, and have a tidy story. The rest may have been perfectly competently conducted, but they don't meet peer review standards that are focused on results rather than methods.

This issue is compounded by underpowered research designs. With underpowered designs, in order for me to obtain positive results that are publishable, I need to leverage chance variation. As a consequence, the effect size estimates in the published literature are necessarily exaggerated - called the winner's curse.

Reforms? A next post...

In the wake of your study, do you think the problem is in the experimental methods preferred by psychology, or in the criteria for publication used by major journals (and the role such publications play in tenure)? What kind of reform would you propose to the field of psychology?

ParkerAdderson

On reforms. A general answer is that the full research process should be transparent. Science depends on transparency in order for the community to evaluate the evidence for each others' claims. Right now, with just the publication you only observe the products of my research, not my process, data, or materials. Being able to examine how I arrived at my claims will make my process more reproducible, and the threats to my inferences more apparent.

Some specific ways that we can nudge the incentives so that what is good for science is also good for my success as a scientist: (1) TOP Guidelines: http://cos.io/top - journals can incentivize or require more transparency in the research process to be published

(2) Registered Reports: https://osf.io/8mpji/wiki/home/ - journals can conduct peer review in advance of data collection so that research is evaluated on the importance of the question and the quality of the design to test the question. This complements peer review that occurs after the results are known.

(3) Registration: The file-drawer can be eliminated if there is a log of what research has been conducted so that the results that do not survive peer review are still discoverable. Further, confirmatory research can be distinguished from exploratory research by preregistering analysis plans. Both confirmatory (hypothesis testing) and exploratory (hypothesis generating) approaches are vitally important for science. But, it is critical that the distinction be clear - one cannot confidently generate and test a hypothesis with the same data. EDIT: adding link to our effort to support registration: the Open Science Framework: http://osf.io/

Thank you for your work!

It seems that psychology can very quickly attract a lot of negativity, and since your reproduction results have come out, there's been some attacks on psychology as a weak pseudo-science. However, other scientific fields seem to have their own problems: a reproduction of 53 "landmark" papers in oncology was only able to replicate 6 (11%) of their findings (Begley & Ellis, 2012), and the current' 'record holder' for the most fabricated data is anaesthesiologist Yoshitaka Fujii with 183 falsified papers (Retraction Watch, 2015).

Do you believe that psychology has more serious reproduction problems than other research fields? And do you believe that psychology is unfairly targeted?

Refs:

- **Begley, C Glenn, & Ellis, Lee M. (2012). Drug development: Raise standards for preclinical cancer research. Nature, 483(7391), 531-533.**

- **Retraction Watch (2015) "The Retraction Watch Leaderboard" http://retractionwatch.com/the-retraction-watch-leaderboard/**

DrunkDylanThomas
Very important questions. The short answer is that we do not know. This is the first systematic effort to get an estimate in a field, and it is not even definitive for psychology. As you note, the closest other estimates based on empirical data are reports of replication efforts in cancer biology by Amgen and Bayer. As crude estimates, those observed 11% and 25% replication rates. However, we don't know how to interpret those reports because none of the studies, methods, or data were made available to review. In partnership with Science Exchange, we are now conducting a Reproducibility Project in cancer biology (detail here: https://osf.io/e81xl/wiki/home/). We are also working with folks in a few other fields to develop grant proposals for Reproducibility Projects in their disciplines. Our hope is that this project will stimulate systematic efforts across disciplines so that we can have better data on the rate of reproducibility and why it might vary across research applications.

There are reasons to think that the challenges for reproducibility are pervasive across disciplines. There are many reasons, but a common one is that the incentives for researcher success are pushing toward getting published, not getting it right. We talk about this in depth here: http://pps.sagepub.com/content/7/6/615.full .

Is psychology unfairly targeted? Well if people are using the Reproducibility Project to conclude that psychology's reproducibility problem is worse than other disciplines, then, yes, because we don't yet have evidence about differences in reproducibility rates. As for myself, I think the main story of the Reproducibility Project is a positive one for the field - the project was a self-critical, community effort to address an important issue that people in the field care about. Science is not a publicity campaign, it is a distributed effort to figure out how the world works. The Reproducibility Project is just an illustration of that in action.

How did you determine which studies to select for replication? Were they randomly selected and selected in an unbiased manner?

have_a_laugh
The articles were from three top psychology journals: Psych Science, Journal of Personality and Social Psychology, and Journal of Experimental Psychology: Learning, Memory, and Cognition. From these, we created an "available studies pool" from which replicators could choose a study that was relevant to them and their interests. Replicators were asked to reproduce the last study from their article, and to select one key effect as the target of replication.

This isn't a random sample, but it was designed to reduce bias. We didn't make the whole year of studies available at once because then all the easy ones would be snatched up and the more difficult procedures would be left behind. Instead, we released them in batches in chronological order. About 60 were left unclaimed from the available pool.

Edit: Like others mentioned, there is more information in the article. You can see a PDF here: https://osf.io/phtye/

Thank you for addressing these issues. Are you familiar with any current clinical approach that was formulated, at its root, from one of the studies that you have brought into question?

I understand that most of these studies are not necessarily clinical in focus, but I'm curious as to whether there are any current clinicians that would need to rethink their use of materials based on the unreliability that you've uncovered.

leontes

Good question. It is conceivable that there are some original studies in this set that have made it to influencing some kind of clinical or social application, but I doubt it. These were almost entirely basic research studies that do not have direct clinical relevance. Over time, they may accumulate to evidence for clinical application, but that is a few steps away.

Hi guys!

I'm an algorithm guy that currently is collaborating a lot with neuroscientists and neuroscientists, and all of their statistics methodologies. I'm pretty baffled by the statistics methodologies that they use and are standard in their field. It seems to me that all of them are explicitly intended to generate false positive results. This is especially true when more and more advanced analysis techniques are introduced, which I feel that are basically just data-dredging (regularizations, sparsity constraints, *omics).

What's your opinion on this?

Also, I often find some approximations used to be especially bad in inflating statistical significance. Like assuming some data measures, like points on a scale, to be gaussian-distributed when they have very long tails. This usually results in very-strongly over-estimated correlation coefficients/underestimated covariances, boosting p-values.

Can we reform this?

lucaxx85
Yes, there are many analytic practices that inflate the likelihood of false positives. A solution most relevant to your comments is having a two-phase analysis process. Split the data in two and put one part aside. With the first half of the data, conduct explicitly exploratory analysis for hypothesis-generation. Once there are well-defined hypotheses and models, use the second half of the data for confirmatory hypothesis-testing. This way, one can take full opportunity of learning from one's data, and then apply constraint so that the confirmatory tests are as diagnostic as possible for inference.

As a co-author, I'm sure you've seen the way the media is reporting this (Eg: http://www.independent.co.uk/news/science/study-reveals-that-a-lot-of-psychology-research-really-is-just-psychobabble-10474646.html). It seems hugely unfair for them to report the findings as an attack on the entire psychological discipline and also to frame it consistantly as "other scientists" failing to replicate the data rather than other Psychologists, just to hammer their message home.

I don't know much about publication and the media, so I have a genuine question: Is there any way to hit back against this kind of reporting? I feel like it's purposely misleading but do the authors or journal have any sway over reporting or is the media allowed to spin it any way they please? As someone who will be starting their own research come October, this is something I'd love to have insight on.

SealingPandora
We've definitely seen some varied summaries of the findings and their implications. Our best way to exercise control over how the media reports our findings is to be involved in crafting the message. Our goal has been to be clear about what the results suggest and what the takeaways should be. We cannot control what other people say or how our findings are conveyed to the public, but we can be clear in our own speech. One way to help reduce misleading headlines is to avoid overgeneralizing yourself.

As a co-author, I'm sure you've seen the way the media is reporting this (Eg:

http://www.independent.co.uk/news/science/study-reveals-that-a-lot-of-psychology-research-really-is-just-psychobabble-10474646.html). It seems hugely unfair for them to report the findings as an attack on the entire psychological discipline and also to frame it consistantly as "other scientists" failing to replicate the data rather than other Psychologists, just to hammer their message home.

I don't know much about publication and the media, so I have a genuine question: Is there any way to hit back against this kind of reporting? I feel like it's purposely misleading but do the authors or journal have any sway over reporting or is the media allowed to spin it any way they please? As someone who will be starting their own research come October, this is something I'd love to have insight on.

SealingPandora
And, as a follow-up, we have been genuinely pleased that many science writers have written beautifully about the project, the questions it raises, and - just as important - what it does NOT show. There are many examples. Here are a few: http://www.theatlantic.com/health/archive/2015/08/psychology-studies-reliability-reproducability-nosek/402466/ , http://www.vox.com/2015/8/27/9212161/psychology-replication , http://fivethirtyeight.com/datalab/psychology-is-starting-to-deal-with-its-replication-problem/?ex_cid=538twitter , http://www.buzzfeed.com/catferguson/red-wine-is-good-for-you

Due to the file drawer problem, it seems as though there's no way to get an accurate sense for the strength and reliability of effects with traditional article publishing. Linking replication data to original effects would be one way to deal with this. Is the Center for Open Science considering some sort of way to link data sets, so that if they are all related to a specific effect, they could all easily come up in a search?

Also, it seems like ethics reviews for the most part are fairly useless. Registering hypotheses and methods with an ethics committee, who then helps make sure the results are available (with the Center for Open Science, for example) could almost completely eliminate the file drawer problem, when combined with a flexible tagging system to group findings, methods, etc, and would make ethics reviews extremely useful. Is the Center for Open Science considering advocating this, or for any other top-down changes to encourage our field to be fully open and transparent? It would also allow for easy meta-analyses, enable us to find smaller effects, determine moderators, and stop wasting time, as many studies right now (even with significant effects) never see the light of day.

Edit: Also, for this particular replication effort, even for the studies that did not reach statistical significance, the majority of replication effect sizes were still positive. This seems to suggest that the pool of non-significant studies in-aggregate provides evidence that some of those effects are probably real. Have you considered any analyses to see in-aggregate, what proportion of the original studies probably have real effects, and which are probably type I errors?

One final thing: as "significant" replication effects in the opposite direction would be considered a failure to replicate, it seems like this really would be a case where one-tailed statistical tests should be used. There are very few papers in the original analyses where this would come in to play, but there are a couple that were rated as "non-replicated", when they did replicate with a directional hypothesis.

Zorander22
The Reproducibility Project used a web app that the Center for Open Science developed to manage the project and all of its data—the Open Science Framework (OSF) (https://osf.io). The OSF is a resource for scientists to share their research and organize it. In using tools like the OSF, scientists

should be able to make their work publicly available and connected in such a way that it makes obvious what findings are related to what data.

Like /u/octern mentioned, we did discuss and employ several methods of analyses. See the paper (https://osf.io/phtye/), pages 9-16. I would add that the replications attempted to mimic the original analyses as closely as possible. When it was stated that the test was one tailed, the replicators did the same thing. In many cases replicators and original authors identified alternative means of analysis that often times seemed to be more up-to-date or preferable to those that were published in 2008. In our dataset (https://osf.io/fgjvw/) that we only report there results of the replications of the original analyses. The individual reports (https://osf.io/ezcuj/wiki/Replicated%20Studies/) , however, provide additional detail on the supplementary analyses.

Finally, and crucially, these replications are only singular attempts and can't truly evaluate if an effect is "real." The RPP set out to estimate the rate of reproducibility in psychology, not to determine whether or not certain findings were true. In the process, we've learned a lot about what it takes to conduct a replication and how difficult they can be. The low reproducibility rate is a reflection of the barriers to reproducibility.

Have you identified certain "trends" when it comes to studies which cannot be reproduced?

Akesgeroth
Yes, we did exploratory analysis of a number of factors that might be correlated with replication success. The details are in Tables 1 and 2 of the paper: http://www.sciencemag.org/content/349/6251/aac4716.full.pdf . Some that stood out were: p-value of the original study outcome (closer to p=.05 were less likely to reproduce), rated challenge of conducting the replication (harder to conduct were less likely to reproduce), whether the test was of a main effect or interaction (interactions half as likely to reproduce), whether the result was surpising or not (more surprising were less likely to reproduce), and whether it examined a cognitive or social topic (cognitive twice as likely to reproduce). Some that showed little to no relation to replication success were the expertise of the original authors or the replication teams and rated importance of the original result.

I think it would be both unfortunate and incorrect if people concluded from this study that psychology is "less scientifically rigorous" or a "soft science" (whatever that means) because many effects failed to replicate. As the authors stated in the paper, a failed replication does not necessarily deem the original finding "false." The irony is that, despite many studies failing to replicate, psychology's effort to measure this issue (no other field has done this) is evidence in itself of the field's dedication to the scientific method in its proper form. Moreover, a reproducibility problem is not unique to psychology.

to the authors - do you believe this paper will change the incentive structure for researchers to perform replications of past work?

benign_b
I don't think the paper itself will address that cultural challenge, but it may help provide a basis of evidence for efforts that do tackle it. Our results show a striking difference between effect sizes and p-values from the original work and their replications—see Figure 1 from our paper: https://osf.io/7js8c/. We also saw a relationship between how surprising the key effect was perceived to be and the likelihood of replication. We hope that the discussion around these issues will prompt further action, but this one paper is unlikely to be the basis of a cultural shift.

This kind of research can inform our actions and proposed solutions to issues like publication bias and an imperfect incentive structure. The Center for Open Science, for example, has helped develop

guidelines for journals and funders to help promote openness and transparency in the research they publish and fund: https://osf.io/9f6gx/wiki/home/

Have you had any interesting responses from the authors of the original published results? Were they defensive, critical, appreciative?

dity4u
On the whole, the original authors were very helpful and had a positive outlook. The vast majority were able to provide the original materials or at least assist in the creation of new ones. On some occasions we were unable to get in touch with the authors.

I would say that the biggest concern from the original authors was that they wanted to be sure that their work was accurately represented. As this was also the goal of the replicators, most conversations went very well. The discussion between original and replicating researchers was vital to ensuring high quality replications. In some cases there were disagreements over the interpretation of a replication's findings, and in situations where the dispute could not be resolved, we hoped to offer a means for the original authors to provide their own commentary. Links to original author responses are included in our list of replications: https://osf.io/ezcuj/wiki/Replicated%20Studies/

Can you please do more replications of Construal Level Theory? I've spend the better part of my phd working on that theory and it simply doesn't work. Nothing replicates. It's since become one of the leading social psych theories of the decade and really needs more debunking.

Palmsiepoo
The goal for our project was not to select any particular study or area of research, but to examine the reproducibility rate for a quasi-random sample of studies in the literature. As for your experiences, I recommend that you post your study designs and data to the OSF (http://osf.io/) and make it available for others. That could be useful in two ways: (1) you might receive useful feedback from other experts in the community about changes that you could make to your designs that may be important for the effect, and (2) your data might help others who are conducting meta-analyses or novel research to evaluate CLT more generally.

Would you be open to answer some questions for my students via Skype conference to an audience of 100+?

I could also make it so that JUST the research methods people join

This would be for November

Jose_Monteverde
We'd be happy to talk to you about that. We give free onsite and virtual workshops related to reproducible stats and methods (http://centerforopenscience.org/stats_consulting/), so just send us an email at contact@cos.io and we can try to work out details.

Do you think (the editor-in-chief of The Lancet) Richard Horton's comments on reproducibility in science adequately explain your results?

Much of the scientific literature, perhaps half, may simply be untrue. Afflicted by studies with small sample sizes, tiny effects, invalid exploratory analyses, and flagrant conflicts of interest, together with an obsession for pursuing fashionable trends of dubious importance, science has taken a turn towards darkness. As one participant put it, “poor methods get results”.

Asshole_PhD

We don't know. Low-powered designs and a strong selection bias for positive (p<.05) results does increase the likelihood of false positives. But, we do not have sufficient evidence about any particular effect in our study being a false positive. All of the factors that Horton, Ioannidis, and others have discussed are challenges for maximizing the credibility of obtained results. A more detailed discussion of those is in a few recent papers. Here are two: http://europepmc.org/articles/pmc4078993 , http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4103621/

In your research article in Science, you have detailed some interesting findings. To me, it looks like the evidence in Figure 1B, suggests that while most findings regress to the mean, about 10-20% have a reasonably robust effect size and p-value. One can imagine that as most results vanish over time and repeated study, a "droplet" would detach from the bulk "chess piece". A droplet of stable results. Is this a reasonable conclusion?

Another thing that is striking is that having the right hypothesis and testing it the right way seems to be more important than relative expertise.

helm

I am not quite sure that I understand the idea, but a couple of thoughts.

(1) Yes, there is an overall decline effect. 83% of the replication results were weaker (closer to zero) than the original results.

(2) It is possible that there are two distributions - a distribution in which there is an effect to obtain, and a distribution in which there is no effect to obtain. But, it is difficult to identify those separate distributions in what we have here.

(3) We do have some evidence that significant versus non-significant is NOT sufficient to distinguish these two distributions. The distribution of p-values in the non-significant effects was not uniform (though the test p=.048 doesn't inspire a strong inference). That suggests that one or more of the non-significant effects was just underpowered. Looking at Figure 3, it is easy to see some likely candidates for that.

(4) It is also quite possible that the distributions of significant and non-significant results would be different if we conducted the exact same procedures again, and even more different if we did another independent round of design of the replication protocols. These data provide little conclusive evidence about any one of the effects.

How much of the lack of reproducibility is the result of poor statistical analysis by the original researchers? Have you seen any blatant attempts at purposeful manipulation (p-hacking, etc.) of the data to obtain a specific result?

Would requiring researchers to pre-register their experimental plans with a third party prior to conducting any experiments help reduce the cherry-picking of data? If they originally planned n=20 and then only had n=16 they would need to explain the discrepancy.

shiruken

We didn't evaluate the quality of the analyses in the original article for this project.

Yes, preregistration can make clear when design or analysis changes occur from original plans to what is reported. Many times those changes are highly defensible, the key is to make it transparent so that the reader can decide. In the case of the Reproducibility Project, all designs and analysis plans were prepared in advance and registered publicly on the Open Science Framework (https://osf.io/ezcuj/wiki/home/) so that we would have strong confirmatory tests.

Also, the Center for Open Science will soon launch an initiative to encourage people to try out preregistration call the Pre-Reg Challenge (http://cos.io/prereg/). It includes $1,000,000 of awards for conducting and publishing preregistered research. We expect to learn a lot about the promise and challenges of pre-registration in basic science research through that initiative.

Thank you very much for this AmA, this is of great interest to the whole scientific community and science enthusiasts.

My question is related to statistical significance in not only psychology, but science as a whole. With the recent multiple retractions and reproducibility issues (which happnened not only in psychology, but also in preliminary cancer studies), and with the Basic and Applied Social Psychology journal refusing to publish p-values due to significance issues arising from its usage, I wonder: is there any statistical alternative to solve the issue of reproducibility once and for all?

As a medical student and science enthusiast, the reproducibility matter in science is no doubt of great importance to my present and future life, as well as for all others that are directly and indirectly related to science (which is probably everyone in the planet!)

Thank you so much.

vasavasorum
Great question. I'm not sure there is one statistical approach to rule them all. Bayesian analyses have been gaining popularity as an alternative to NHST, but more recent research has shown that bayesian inferences can also be invalidated by researcher degrees of freedom (http://datacolada.org/2014/01/13/13-posterior-hacking/), so no one technique is completely fullproof. I think there are a few things that would help a great deal, whether NHST or Bayesian inferences are being used:

One is a more holistic approach. So, don't just look a p-values or bayes factors. Look at p-values, effect sizes, measures of precision, etc. and also take things like statistical power into account and then make judgments in light of all these pieces of information.

A second is to increase the transparency in reporting of analyses. So make the distinction between exploratory and confirmatory analyses more clear in studies. This can be accomplished through behaviors like pre-registration of studies and analysis plans for confirmatory research, as well as more openness to the explicit reporting of exploratory findings.

Finally, though this is by no means an exhaustive list, I think it's important to make it easier to publish null results to both decrease publication bias as well as to decrease incentives to find statistical signifiance that can often, unconsciously, lead ot researcher degrees of freedom behaviors which can increase rates of false positives. The Registered Reports format (https://osf.io/8mpji/wiki/home/) is being adopted by some journals to help with this.

I know little about how to evaluate statistical methods... is there a general misapplication of complex statistics, or is social science research problematic even with its use of t-tests and ANOVAs?

Thanks for your time!

Bruleur
There are a host of research practices that can lead to problems for both simple and complex statistical models. For example, in situations where statistically significant findings are more likely to be published, the reported effect sizes from statistical models, whether it be a t-test, anova, or something

more complex, will be likely to over-estimate the true effect size if the study is underpowered (https://osf.io/sf9cv/). Researcher degrees of freedom (this is also sometimes referred to as p-hacking) can also lead to an increase in false positives in both more simple and complex models. So, it is less about the complexity of the model and more about general research practices (under powered studies, researcher degrees of freedom, current incentive structures, and publication bias) that can lead to issues with replicability. Note that many of these practices are seen in a broad range of scientifici disciplines, not just social sciences.

Hi, clinical psychologist here. Even though my job is mostly clinical, I do get to spend some of it in research and am really appreciative of what you're trying to do here. I'm wondering if you're planning to reproduce studies in my field? Given that clinical psychology is one of (if not the largest) branch of specialty psychology, it would be interesting to have some of our evidence-based treatments re-examined by your project, especially treatments like prolonged exposure and CPT, which has many articles published by their founders claiming efficacy, but anecdotally have been criticized (see Slate's article of "Trauma Post Trauma"). Is delving into the clinical psychology literature a plan you have for your future project?

robinsena
While we are working with researchers in several other disciplines to begin more Reproducibility Projects, clinical psychology is not one of them. Presently, the Reproducibility Project: Cancer Biology is ongoing (https://osf.io/e81xl/wiki/home). We are happy to help support the development of additional projects in other fields, if there are interested researchers.

How many of the studies do you think were results of bad science compared to how many were falsified such as much of the work of Diederick Stapel?

schrodingers_beaver
This is an important question. We have no reason to believe that bad science or fraud was involved in any of the original studies that did not replicate. While you are correct that less than half of the 100 studies included in this project were considered successful replications, there are four likely explanations for this: 1) the original effect was a false positive, 2) the replication outcome was a false negative and the original effect observed was correct, 3) the findings from both the original and replication are accurate, but the methodologies differ in a systematically important way (which we tried to minimize by contacting the original authors and creating a set protocol for all replicators (https://osf.io/ru689/)), or 4) the phenomenon under study is not well known enough to anticipate differences in the sample or environment. None of these possibilities indicate bad science or fraud was involved, but instead that replication is difficult to achieve and the factors that influence reproducibility are uncertain. Much more research on reproducibility will need to be conducted before we have any evidence to support why an experiment's replication was unsuccessful.

How many of the studies do you think were results of bad science compared to how many were falsified such as much of the work of Diederick Stapel?

schrodingers_beaver
One additional note: One of Stapel's papers was in the sampling frame but it was removed from consideration for the Reproducibility Project because it had been retracted due to fraud.

Hi! I'm a recent psych graduate who's done some work on trial registration in clinical psychology. I recently came across the Registered Reports (RR) format that you promote and I think that it is an absolutely fantastic idea. I have two(ish) questions:

1. **What has the overarching response been to the RR format from editorial boards of journals? (I know that it is available in some journals but some of these journals are directly affiliated to the RR project (e.g., Chris Chambers in an editor of both Aims Neuroscience and Cortex))**
2. **Do you have plans on expanding and opening an office in europe?**

lukezndr

Registered Reports (https://osf.io/8mpji/wiki/home/) is presently in use by 16 journals across a number of disciplines. Journal editors that have not yet adopted the format find it intriguing, but many are interested in seeing what comes from the experiences of the initial adopting journals. Here are two examples, a special issue of Social Psychology with 15 Registered Reports (https://osf.io/hxeza/wiki/home/) and eLife's publishing of our Reproducibility Project: Cancer Biology studies as Registered Reports (http://elifesciences.org/collections/reproducibility-project-cancer-biology).

We do not have immediate plans to add an office in Europe, but if we continue to receive funding support to operate and grow, then we expect that we will need to do so to better support the research community in Europe.

[RPP Author] Are there plans to replicate the Reproducibility Project: Psychology? A replication that sampled from 2015 publications could inform how these problems may have changed in the past 7 years. A replication years from now would hopefully illustrate the self-corrective nature of science and the importance of this project.

BenjaminTBrown

No plans (by us) yet. Our current priority is broadening to have similar investigations in other disciplines. But, I agree that this would be very interesting. I think I'd wait until 2018 or so for sampling purposes though. Many changes to research process are emerging right now, the effect of those process may take some time to have a meaningful shift in the literature.

I'm currently a graduate student in psychology and have seen many of my peers rush to defend psychology and try to make these results out to seem as if they are 'not so bad after all.'

I have a hard time seeing how one would come to psychology's defense (or JPSP/the original authors defense), and I have a hard time seeing how this could be interpreted as anything other than a strong negative mark. When the inability to replicate is combined with statistical techniques that are already criticized (e.g. NHST), I don't see how one can come to the interpretation of 'This is not so bad,' instead of 'Ok this is pretty serious, we ought to do something about it and fast.'

I suppose I'm curious as to what your opinions are in regards to those two different opinions. In your view, what are the implications of this work for authors, journals, and psychology as a whole?

Demon_Slut

I agree that this paper provides some support for calls to action rather than calls for complacency. We can do better than this. And, there are steps that we can take that are not so difficult to do that could have a substantial impact.

That said, change is hard, and anticipating the impact of new approaches is harder. The advantage of the current system is that we know its demons (of course I don't need to tell you this, demon_slut). So, when trying out new efforts to nudge incentives and improve reproducibility, we must evaluate their impact, refine to improve, and dump what doesn't improve. I think we can get a long way in improving reproducibility while not damaging (or even improving) innovation.

What organization or person funds your work, primarily?

TygerPanzy
Here is a list of the Center for Open Science's funders along with $ from each funder: http://centerforopenscience.org/about_sponsors/ . For the Reproducibility Project itself, it started with no funding outside of funds personally available. 1.5 years in, we received a $250,000 grant from the Laura and John Arnold Foundation. That made a HUGE difference for the project.

The overall take away message from your publication by the general public is going to be that psychology, in general, is not trustworthy. I think it's extremely important for science to have a reproducibility feedback system. Is there any way we can salvage the extremely negative press and get the general public's trust back?

Formic-and-Pikachu
I don't have as negative a perception of the media coverage of the Reproducibility Project as you do here. There have been some crummy articles, but many have been quite good in identifying the challenges of reproducibility and why they exist. In particular, some have been excellent at pointing out the positive steps that psychology has been advancing over the last few years such as the TOP Guidelines (http://cos.io/top/), increased support for publishing replications, and novel initiatives like Registered Reports (https://osf.io/8mpji/wiki/home/). I think the key effort for earning the public's trust is to keep working toward a science that practices its values.

Thank you for conducting this exceedingly important study and publishing it so everyone has access to to it.

The reports on this study are generally very doomsday... All hope is lost in psychology research. To me, your study is the opposite of that. But what does the field need to do to gain back trust, become rigorous and move forward. And how do we interpret the previous literature?

Jobediah
You are correct that our paper had the opposite goal: we intend this investigation to be a conversation starter, not the final word on reproducibility in psychological literature. While a great deal of the coverage has been responsible but realistic (example: goo.gl/Hqkmf5), you make a great point that this investigation was conducted in order to instigate progress. Moving forward, we should keep two points in mind: 1) Conducting a replication was made immeasurably easier when original materials (and/or original analysis scripts) were readily available. Making your work publicly accessible assists replications by removing assumption out of the preparation process. To this point, we made all of our materials, data, analysis scripts, and supplemental materials available here: https://osf.io/ezcuj/wiki/home/. Making your work available is the easiest way to foster collaboration and constructive critique. 2) You can be your own worst critic. Challenge yourself to be tough on your work prior to submitting for review, and you will find that you may address many of the criticisms would have received ahead of time. As far as interpreting the previous literature, this project does not provide a definitive answer on how to assess reproducibility for individual studies. It does indicate that, overall, we may need to be more skeptical of what we interpret. Although the replications had effect sizes that were on average half the size of the original, we did see that the larger the effect size of the original, the more likely the replication was successful. As you look at previous literature, readers may need to be slightly more skeptical of effects with small effect sizes and p-values closer to 0.05.

One thing that you didn't really talk about in your paper that I'd like to discuss is a cultural change in how we present exploratory data. While pre-registering is useful, I think it's actually targeting a symptom and not a cause. There would be no incentive to p-hack or adjust your

hypotheses if it were culturally appropriate to talk about exploratory results.

If I do a study looking for X and instead find Y, I can't get published unless I say that I was looking for Y all along. If we move towards a system where I can say "I didn't find X but possibly found an interesting novel result that we should try to replicate" we still get the potentially novel and useful information while retaining intellectual and scientific integrity. Do you think a shift like that is possible?

ImNotJesus
Yes, this is important. I wrote a longer reply in another thread about this. The key is that both confirmatory and exploratory approaches are vitally important for science. Also vitally important? Knowing which is which. There is no shame in not knowing the answer in advance. We need to explore data because we are studying things that we do not understand. We SHOULD be surprised by some outcomes. But, can also express freely that these results were obtained via discovery. That is, the discovery process helped generate a new hypothesis. Testing it requires new data. I have rarely had push back in reporting exploratory results as exploratory. In the other thread I pointed out that we made this explicit in my first project (master's thesis) and in my most recent publication (earlier this week). Try being straightforward with it; you might be surprised by reviewers' responsiveness to it.

Can you envision a future where psychologists have the sort of unified goals and agreement on methods that characterize physics and enable large-scale international collaborations on projects like the Large Hadron Collider? Are there any projects currently underway or in development that you think have the potential to unify psychologists (at least in a given subfield such as cognitive psychology)? What would have to change in our field to allow that level of cooperation?

seitanicverses
Yes, I think that there is huge potential in collective effort for psychology (and other disciplines). The Reproducibility Project is one example, but the Many Labs projects are additional examples that this is not just a one-off: https://osf.io/wx7ck/ . Moreover, there is now an effort called The Many Lab to help organize these collaborative efforts: https://osf.io/89vqh/