Gmail

Rick Klein <raklein22@gmail.com>

Psychological Science - Decision on Manuscript PSCI-19-1597

Psychological Science <onbehalfof@manuscriptcentral.com> Wed, Feb 12, 2020 at 11:08 PM
Reply-To: psci@psychologicalscience.org
To: raklein22@gmail.com

12-Feb-2020

Dear Mr. Klein:

I thank you for submitting "Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement" to Psychological Science. Thanks also to anonymous R1 and R2, each of whom provided detailed, insightful, constructive critiques, as did Tom Pyszczynski & Armand Chatard as R3.

I admire the ambitions of this project. I am sure that a horrific amount of work has already gone into it. In my view, a revised version of this manuscript will very likely warrant publication in a good journal. But I don't think Psych Science is the right outlet. There are just too many gliches and the meaning of the findings is too murky. I have therefore decided to decline the manuscript without inviting a revision.

Let me emphasize again that I believe a revised version of this work is quite likely to succeed with the right journal. I think it is desirable to do a thorough and careful revision expeditiously and get it reviewed by experts. It seems to me that the revision might work as an AMPPS submission, with a fair amount of the emphasis being on methodological concerns. In any case, wherever you submit a revision, please feel free to include with it a copy of this letter and the reviews. Sometimes editors use such information in ways that expedite the review process. I could on request provide such an editor with contact info for R1 and R2.

I am keenly aware of the disappointment that a letter of this nature brings. I am sorry that the outcome was not positive. As you know, Psychological Science must be extraordinarily selective due to the large number of manuscripts that are submitted to the journal (nearly 2,000 new submissions are expected this year). It doesn't help that my term is already over and my time very tight these days. The reviewers and I are hopeful you will find the comments in this action letter and appended reviews useful as you consider the next steps in this research program.

Thank you again for considering Psychological Science as an outlet for your work. I hope you will do so again in the future.

Sincerely,

Steve

D. Stephen Lindsay
Erstwhile Editor, Psychological Science
psci@psychologicalscience.org

REVIEWER(S)' COMMENTS:

Reviewer: 1

Comments to the Author
Many Labs 4 explores whether input from original authors increases the replicability of an effect. Unfortunately, neither author-advised studies or in-house studies (where replicators tried made their own version of an original study) showed an effect predicted by Terror Management Theory. The replication failures for TMT are interesting and of some service to the field. However, the failures mean that they key investigation of ML4 is unaddressed (since there seems to be no effect regardless of author-advice, there is no possibility to find an impact of the author's advice).

I have to admit to be quite skeptical of TMT. Every time I read about the theory (or its predictions), I laugh out loud. It is hard to believe that anyone takes it seriously (I mean, the obvious things are obvious: yes people are afraid/concerned about dying. The predictions are obviously false: that priming about death would make people more pro-US is rather ridiculous.) Given the influence of the theory, I clearly hold a minority view, so that is not a reason to reject the paper. I just wanted to make my view explicit.

Along those lines, explicitness is a strength of current paper. The authors go to great lengths to be open and honest about what was done and why. Even the flaws are clearly laid out for the reader, so we get an accurate presentation of what was done.

That being said, the manuscript falls well short of its goals. The key problem is one of methodology. To investigate the impact of original author advice, the replicators needed to use an experiment that investigated a real effect. It seems that they did not do so, and thus the key point of ML4 fails. To their credit, the replicators do not "spin" the result as if they simply asked whether they could replicate a TMT effect. But being open and honest does not mean that the replicators did a good job (or good science). The study of author advice was just poorly done. One might argue that it is not the replicators' fault that the TMT experiment did not work; but I would say that it is their fault to pick a study that was not first validated as working.

Don't get me wrong, I think the study results should be published to share the knowledge about the replication failures and I think the authors should be honest about their original plans to investigate the impact of author advice. However, I think Psych Science should be publishing the best scientific work, and this study does not satisfy that goal.

Some specific points:

-Pages 16-17 discuss power and effect sizes for various sample sizes. The numbers do not match up (at least if sample sizes are split equally across the two conditions).

-Several places in the text refer to "high power", but power is only defined relative a specified effect size. In particular, the conclusion section refers to ML4 as being a high power investigation of TMT, but if there is no effect then there cannot be high power for an experiment. In general, any reference to power has to simultaneously refer to an effect size.

-When the text first described a three-level meta-analysis, I thought that it seemed like overkill and potentially problematic. Sure enough, there was insufficient data to support such an analysis. But why? If the authors think the three-level meta-analysis is necessary, then they should have gathered enough data to do it. One gets the sense that the replicators are as much in the dark as the original authors about how to get/analyze data to investigate this topic.

-The author advice seems to be taken without any justification (at least, none is given). This seems weird, as it is the reason for a method that matters, not the source (although they might be correlated). This is especially noticeable for the exclusion sets. The author-advice involves two exclusions sets. Well, if they don't know how to exclude subjects, then why follow their advice at all? Indeed, the original authors do not seem to know what they are doing, but then neither do the replicators. (That sounds harsher than I intend, I do not mean that the original authors, or replicators, have _no_ idea what they are doing; but neither group seems to know exactly what should be done in this situation; and that knowledge seems critical for what they set out to investigate.)

-The analysis for research question 3 (effect of standardization) seems ad hoc and unjustified. It is basically a step-wise regression approach, which is not considered a very good method of model comparison. Why not use an AIC or BIC approach?

While reading the manuscript, I marked some additional comments/corrections. A pdf copy is attached.

Reviewer: 2

Comments to the Author
Review

Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement

Summary

The present MS reports the efforts of 21 labs to determine whether “silent” or non-published or knowledge contributes to the success of replication attempts.
The high importance of such an endeavor is clear: If replications in psychological science depend to some extent on non-accessible knowledge or let call it a “je ne sais quoi”, this would provide a challenge for replication attempts and, in my opinion, a challenge for psychological science as a science.
To do so, the MS chose to replicate the “mortality salience” effect (MS; I will use “paper” to describe the current manuscript), derived from Terror Management Theory (TMT), based on procedures published by Greenberg and colleagues (1994). In short, participants either have to think about dying and their body decomposing, or about watching TV. The DV is the world-view defense, measure by the dislike for an Anti-American essay’s author.
Half of the labs were assigned to a condition that tried to replicate the MS effect based on the description in Greenberg et

al. (1994), and half of the labs had (some of?) the original authors as consultants.
The MS does not find evidence for differential author evaluation, based on the mortality salience manipulation.

Evaluation

I was excited about the MS. It is also easy-to-read and well-balanced. I am also happy and proud that psychological science is now at a state where we may collectively do this kind of research. However, upon reading and re-reading, I also have substantial concerns. Please take these concerns with a grain of salt - I really appreciate the attempt, and I am fully aware of the frustration that follows when such an endeavor is not successful.

First, independent from effort and context and specific topic, this is akin to a paper that has a failed manipulation check (here: no MS effects whatsoever); there is consequently not much to learn for the original research question. The conclusion from a high-powered and well-done manipulation that fails the manipulation check would be not to use this manipulation any longer (e.g., the Velten technique as a means to induce mood changes). Clear conclusions regarding a hypothesis and/or theory are not possible (e.g., does good mood foster creativity). Here, there is simply no MS effect, and therefore, the main question cannot be answered.

This shifts the research question to something else: Should we doubt MS effects, or TMT in general? From my own experience with TMT experiments, this question is justified, but the body of (published) evidence suggests otherwise. The meta-analysis by Burke et al. (2010) reports an average effect size of r = .35 with 164 papers and 277 experiments; and there are many “successful” TMT experiments since 2010. As the present paper reports, about 75% of these 277 experiments used the mortality salience manipulations similar to the ones employed here. Are these all false positives? If not, it will be difficult defending publishing the present MS. As much as I would like to suggest otherwise, the present result may be not that informative beyond a reminder that MS effects derived from TMT are fragile.

Second, I am really astonished that there is no effect whatsoever. This could be ameliorated by some additional analyses that show some effects (e.g., on the PANAS or any of the questionnaires used by the “in-house” labs). I have done my share of TMT experiments, and the problem was never to find an effect, but to predict the direction of the effect. In hindsight, it always made sense, but to predict it a priori was in my experience difficult to predict (and we never published these findings). There is little evidence in such personal anecdotes, but to be clear, the MS manipulation used here is anything but subtle. To find *no* effects between conditions in which people think about their rotting corpse or watching TV is really surprising (again – the paper should at least report the available PANAS data).

Third, there are reasonable voices that one should hold replication research to the same standards as original research. As replication research is still a “young” discipline, I believe it should be given more leeway. Yet, if the original research question would have been a simple comparison between two conditions, I would have criticized the following:
The design uses only one stimulus (here, the experiment by Greenberg et al, 1994) to investigate the role of original author involvement; generalizing to other stimuli (i.e., experiments or paradigms) is logically not possible (see Wells & Windschtil, 1999; Judd et al., 2017). A better design (in hindsight obviously) would have been to randomly sample the stimuli (i.e., different experiments); if one stimulus does not “work”, it does not matter so much. Imagine if the present attempt had resulted in strong MS effects in the author advised condition, and weak MS effects in the In-House condition. Would we have concluded that replication attempts benefit from original author involvement? No – this would be an inference error. One would have only been able to derive that author involvement matters for this specific TMT manipulation – no more, no less. For single experiments, authors sometimes avoid this by using prototypical stimuli. Yet, I would be hard-pressed to see a TMT study as the “typical” psychology experiment.
Even without considering the role of stimulus sampling, the design is incomplete. Again, in hindsight, one should have selected a “difficult” effect (e.g., the MS effect used here) and an “easy” effect (e.g., an IAT effect). As stated in the previous paragraph, generalizations would have been impossible. But using an informed selection of studies, one might have been able to make some inferences (e.g., with more complex procedures, author advice might help).
From this perspective, the present outcome might even be GOOD NEWS. A clear failure to replicate might have more informational value than an ambiguous result with regards to the role of author involvement. Now, the paper may argue that it is at least a prototypical TMT experiment.

Fourth, what are potential reasons for the observed null effect? The MS offers some suggestions, but refutes them. In my opinion, there are two missing: a) there might be a ceiling effect due to the high negativity of the Anti-American essay. All the studies deviated in this critical point from the original procedure. b) (apologies for the cynical view) In my view, the present incentive structure in replication research is “not to replicate” rather than to “replicate” (i.e., non-replications are more prominent than replications). I observe an almost morbid curiosity to see prominent effects fail, and replication failures garner currently more attention than success (alas, I lack a good source to back this up; this is more of a general observation).
One of the main accusations against psychology was that researchers were rather unscientific in their positive testing hypothesis strategy of finding “significant” results. Potentially, the pendulum is now swinging in the other direction. Such biases might happen on the level of data collection and generation, not on the level of data analyses (i.e., unmotivated research assistants, lack of training, sloppy procedures). I fully acknowledge that the problem is thereby of an almost untestable nature, but there is one solution, and that is that the same test session shows a (trivial?) significant effect

together with the failed replication attempt. Maybe one may count the higher ratings of the Pro-American essay as such an effect?

For the present paper, the situation is somewhat different, as the paper aimed to find an effect and moderate it, but given the setup, I am not sure if the single sites were aware of the goal. I thus want to urge the main authors to include some evidence that something systematic was going on in the procedures.

Where does this leave me in my recommendation? There are two possible interpretations, and obviously, both are judgment calls.

a) If there is evidence that the sites did not only produce random noise, the paper could be published as a cautionary note for TMT research. I believe the “Handbook of TMT research” appeared last year (could also be cited), and this might be a critical qualification. The paper would also need to argue that this is a prototypical TMT experiment, and some more background work on TMT is necessary.

b) There is no evidence for systematic variations across the sites due to the conditions. Then, the data holds little informational value for TMT research as well, and it should not be published.

I am leaning towards solution a), but I see also the arguments against publication.

In any case, here are some smaller points in order of appearance:

On p. 4, the MS states that one original author doubted that MS effects can be done in a ManyLabs style investigation. That should be commented on or addressed.

On p. 6, under point 3., I would at least like to see a rough estimate to how many experiments this applies. Maybe I am naïve, but as an experimental psychologist, so many effects come to mind that work effortlessly across labs.

On p. 7: “Participants in the “subtle own death salient” condition wrote about the emotions they experienced when thinking about their own death, and about what would happen to their physical body as they were dying and once they were dead.” This is really a pet peeve of mine (see above)…how is this a SUBTLE manipulation? This is one of the most heavy-handed manipulations I know to induce “mortality salience”.

On p. 7, bottom: Make clear that this is the data from Greenberg et al (1994).

Table 1: This is an interesting example how my personal standards have shifted. I immediately thought that all the Ns with less than 3 digits are underpowered.

On p. 16, very small detail: is the calculated power under the assumption of equal sample sizes in the two different groups or based on the factual distribution? Please specify.

On p. 18, the experimenter survey might appear to provide data against my “cynical” interpretation. Yet, it is clear that such explicit questions provide almost no remedy here. Experimenter effects are subtle, and there is good reason that experimenters should be blind to conditions.

Table 2: The p-value of zero for location 13 must be a typo (it would be the single case when the logic of null-hypothesis testing would be fully valid.

Reviewer: 3

Comments to the Author

This paper reports a Many Labs (ML) attempt to replicate a classic terror management theory (TMT) effect, in which mortality salience (MS; reminders of death) increases worldview defense in the form of pro-US bias. Beyond that, the study also attempts to assess the effect of input from those with experience and expertise in the research domain on whether ML teams are able to replicate classic findings. A project like this has the potential to make a useful contribution to the literature. We should say at the outset that, given our long-standing association with TMT, we are not impartial evaluators of this paper. On the other hand, we likely have at least somewhat more knowledge about these issues than most reviewers and are more motivated to look carefully into the data reported here. This paper has been posted online prior to peer review and has already received a considerable amount of attention, but none of the comments on the paper posted thus far noticed some rather serious problems with it. This is one reason that widespread dissemination of unpublished research prior to peer review is problematic.

Having said that, we have mixed feelings about whether this paper ought to be published. On one hand, we’re pleased

that ML researchers are interested in conducting a large-scale study of mortality salience effects and think that replication studies like this have the potential to both increase confidence in existing literature and point to areas where refinement of current understanding is needed. We also agree that the ML hypothesis about the influence of expertise in replication is interesting and deserves to be pursued. However, there are major problems with the way the data from this study were analyzed that led the authors to inappropriate conclusions. The biggest problem is that the data were not analyzed as specified in the preregistration plan. Our reanalysis of the ML data indicate that when the preregistered plan is followed, the study actually does replicate the original effect when the advice from those with expertise and experience in his area are followed. Clearly, the problems we point out need to be remedied, and we assume the authors would wish to do so. Indeed, the differences in findings that emerge when the preregistered data analysis plan are and are not followed could be viewed as demonstrating the importance of following preregistered data plans.

In what follows, we summarize the main problems we see with this study. In the interest of full transparency, and because this paper has received considerable attention on twitter and online blogs, we posted a reply on one of the websites where it was posted that provides more detail regarding our reanalysis of the ML4 data (https://osf.io/6v4kf/?view_only=fb2429209c3e436da12bc3f4930c088a), along with the data and code used in our reanalysis. We hope this will be useful in evaluating the conclusions we've drawn in our evaluation of this study.

Major concerns

The major problem with this work is that the authors did not follow their own data analysis plan as stated in either the paper itself or in their preregistration – and they said nothing in their paper about these deviations. Given that the ML proponents are major champions of the preregistration cause, we found this extremely surprising. The paper, and the preregistered data plan, state that all labs would collect data from at least 80 participants (40 per cell) and that this is the minimum sample size needed to provide an adequate test of their hypotheses. Because the ML researchers planned to exclude a lot of participants in some analyses (all participants who are not White Americans and those strongly identified to the US), we agree with them that 80 participants per study is a bare minimum to have minimally informative power to replicate the original effect within each site.

Inspection of Table 1 shows that 8 samples (38%) did not meet the preregistered minimum sample size for inclusion into the ML study. Contrary to the preregistered plan, data from all of these 8 labs were included in the analyses. Even if small studies are given less weight than large studies in meta-analysis, the inclusion of a large proportion of small studies in a meta-analysis can seriously impact the precision of the meta-analysis results (Lin, 2018; Nüesch et al., 2010) because small studies are often more heterogeneous and less precise than large ones (IntHout, Ioannidis, Borm, & Goeman, 2015). This may lead researchers to misleadingly conclude that there is an effect when there is not, or that there is no effect when in fact there is one. An examination of the ML4 data confirmed this fear. The overall effect size was greater for the 13 studies that met the inclusion criteria of at least 80 participants per lab (Hedges' g = 0.10, SE = 0.06) than for the 8 studies that did not (Hedges' g = –0.06, SE = 0.10). A Mann-Whitney test showed that this difference was significant, U = 24.0, p = .046, with a large effect size, Rank-Biserial Correlation = .54, 95CI[.08, .80]. This shows that the alleged failure to replicate was mainly due to the inclusion of underpowered, heterogeneous and imprecise small studies in the meta-analysis. Importantly, these studies would not have been included into the analyses if the study was conducted and analyzed as planned.

If one is going to tout preregistration as a major virtue of one's research, and a necessary benchmark for trusting other research findings, following one's own preregistration plan is essential. As Nosek et al. (2018) wrote: "An analysis plan is necessary to specify how the prediction will be tested with the observed data." (PNAS paper, p 10518). We understand that problems sometimes arise when collecting data and one might later decide to take a different approach. In that case, this should be clearly explained in the paper and the alternate analyses should be presented and interpreted as exploratory. But nowhere in this paper do the authors mention this deviation – which again strikes us as a very strange violation of the logic for preregistration that is central to the Open Science movement.

Given these concerns, we accessed the data sets and scripts and redid the analyses using the criteria for sample size stated in the paper and preregistration (with N > 80 participants per study, as preregistered). This yielded a very different picture. When analyzed according to the authors' plan, the findings replicate the original study, even when collapsing across expert and in house versions, at least among White American participants (under data exclusion criteria 2). More interestingly and pertinent to the stated major purpose of this research, the effect is stronger in the expert than in the in lab versions, and when the two versions are analyzed separately it emerges only in the expert version. In the expert version, but not in the in house lab version, the ML4 project replicated the original effect among White American participants (under data exclusion criteria 2) and among participants strongly identified with the US (under data exclusion criteria 3). In both cases, the effect is significant in the expected direction (p < .05, one-tailed), with a small effect size (Hedges' g > . 20). The authors did not specify whether they would use one- or two-tailed tests. Given that the hypothesis that MS increases worldview defense is clearly directional in nature and that a finding in the opposite direction would be viewed as a failure to replicate, a one-tailed test is clearly appropriate. We can imagine some might quibble with this, but the effects in the expert condition would have been significant with either one- or two-tailed tests.

This difference in findings, depending on whether or not the preregistered data plan is followed, suggests that not following one's preregistered criteria can lead to Type 2 errors (concluding an effect is not there when it actually is), in

addition to the more commonly discussed potential for Type 1 errors. Minimally, these results show that the authors' acceptance of the null hypothesis is clearly inappropriate.

Other concerns
Another surprising aspect of the ML4 data plan is that toward the end of the preregistration, in a section labeled "data exclusion," they state, "Samples will be included as long as they collect at least 60 participants by the time data collection ends." This exclusion criterion was not followed in any reported analyses, nor was it mentioned anywhere in the paper. Nonetheless, we went back and reanalyzed their data with this more liberal criteria (at least 30 per cell). These analyses yielded the same general results (though of course the values of statistical tests are different) as those reported above, replicating the basic effect with Exclusion Criteria 2 and 3, and finding this effect in the expert but not in lab versions. Having two criteria for inclusion in a preregistered study strikes us as defeating the purpose of preregistration.

We're really not sure how to understand why the authors deviated from their preregistered plan. The fact that these deviations are not mentioned anywhere in the paper makes us suspect an innocent mistake rather than reverse p-hacking or intentional obfuscation. It might be that, given the number of investigators involved, researchers conducting ML studies need to be especially vigilant for possible errors. This raises concerns about the overall level of quality control in projects that involve so many different sites. This is one reason many researchers are skeptical of such efforts, though of course many others view them very positively. We think it is important to balance discussion of the positive potential of this type of research with consideration of the potential problems.

Concerns about the DV and literal vs conceptual replication
As the authors note, the TMT researchers who advised on this study (including ourselves) were skeptical about using this particular dependent measure (evaluations of persons who praised or criticized the US) in the year after Donald Trump was elected president. We think there are sound reasons this is a risky measure in the current climate of political divisiveness. We thought using this measure was ill-advised because it seemed likely to be less sensitive to these processes than it was in previous years because of the anger, despair, and disbelief that about roughly half of Americans were experiencing around that time, concerns that persist to this day This issue notwithstanding, using this measure does provide a literal replication of the procedures used in previous studies, which is of interest as long as one is clear about the ambiguity that results from using an operationalization that may or may not tap into the conceptual variables central to the theory from which hypotheses are derived. Based on our reanalysis of the data, it appears that the MS effect did emerge when expert advice was followed, suggesting this particular effect might be more robust than we suspected. Still, we think it important to take into account the current social zeitgeist when planning and interpreting research regarding attitudes that seem likely to be affected by recent historical events. Though the authors mention this concern, we think it, and the more general issue that many of the social attitudes that social psychologists study are not static and depend on the current social zeitgeist, ought to be more thoroughly discussed.

In sum, there are major problems with the current paper. Perhaps these problems could be resolved with a thorough and fully transparent revision that presents the analyses that were initially planned and preregistered. A thorough presentation of this study could be informative and illustrative regarding a variety of issues that arise in many lab replication efforts.

Tom Pyszczynski & Armand Chatard

---

2 attachments

ML4.PSYCHSCIENCE.pdf
84K

PSCI-19-1597_Proof_hi.pdf
634K