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

Richard A. Klein, Tilburg University
Corey L. Cook, Pacific Lutheran University
Charles R. Ebersole, University of Virginia
Christine Vitiello, University of Florida
Brian A. Nosek, University of Virginia
Paul Ahn, University of Wisconsin-Madison
Abbie J. Brady, Palo Alto University, Ashland University
Christopher R. Chartier, Ashland University
Cody D. Christopherson, Southern Oregon University
Samuel Clay, Brigham Young University – Idaho
Brian Collisson, Azusa Pacific University
Jarret T. Crawford, The College of New Jersey
Ryan Cromar, Brigham Young University – Idaho
Gwendolyn Gardiner, University of California, Riverside
Courtney L. Gosnell, Pace University
Jon Grahe, Pacific Lutheran University
Calvin Hall, Virginia Commonwealth University
Irene Howard, Ithaca College
Jennifer A. Joy-Gaba, Virginia Commonwealth University
Miranda Kolb, University of Wisconsin-Madison
Angela M. Legg, Pace University
Carmel A. Levitan, Occidental College
Anthony D. Mancini, Pace University
Dylan Manfredi, The Wharton School of the University of Pennsylvania
Jason Miller, University of Kansas
Gideon Nave, The Wharton School of the University of Pennsylvania
Liz Redford, Healthy Minds Innovations
Ilaria Schlitz, Behavioral Scientist
Kathleen Schmidt, Southern Illinois University
Jeanine L. M. Skorinko, Worcester Polytechnic Institute
Daniel Storage, University of Denver
Trevor Swanson, University of Kansas
Lyn M. Van Swol, University of Wisconsin-Madison
DeVere Vidamurte, California State University, Northridge
Leigh Ann Vaughn, Ithaca College
Brady Wiggins, Brigham Young University – Idaho
Kate A. Ratliff, University of Florida

Abstract

Interpreting a failure to replicate is complicated by the fact that the failure could be due to the original finding being a false positive, unrecognized moderating influences between the original and replication procedures, or faulty implementation of the procedures in the replication. One strategy to maximize replication quality is involving the original authors in study design. We (*N* = 21 Labs and *N* = 2,220 participants) experimentally tested whether original author involvement improved replicability of a classic finding from Terror Management Theory (Greenberg et al., 1994). Our results were non-diagnostic of whether original author involvement improves replicability because we were unable to replicate the finding under any conditions. This suggests that the original finding was either a false positive or the conditions necessary to obtain it are not yet understood or no longer exist. Data, materials, analysis code, preregistration, and supplementary documents can be found on the OSF page: https://osf.io/8ccnw/

Keywords: Terror Management Theory, mortality salience, Many Labs, replication, metascience

Author note: This project was supported in part by a French National Research Agency "Investissements d'avenir" grant (ANR-15-IDEX-02), the John Templeton Foundation, Templeton World Charity Foundation, Templeton Religion Trust, and Arnold Ventures. We thank Jeff Greenberg, Tom Pyszczynski, Sheldon Solomon, and Armand Chatard for helping develop and review materials. Conflict of interest: B. A. Nosek is Executive Director of the Center for Open Science a non-profit organization with a mission to increase openness, integrity, and reproducibility of research.

Author contributions: R. A. Klein, C. L. Cook, C. R. Ebersole, B. A. Nosek, and K. A. Ratliff conceived and designed the study idea. R. A. Klein, C. L. Cook, C. R. Ebersole, B. A. Nosek, and K. A. Ratliff developed study materials. P. Ahn, C. R. Chartier, C. D. Christopherson, S. Clay, B. Collisson, J. T. Crawford, R. Cromar, D. Dudley, G. Gardiner, C. L. Gosnell, J. Grahe, C. Hall, I. Howard, J. A. Joy-Gaba, M. Kolb, A. M. Legg, C. A. Levitan, A. D. Mancini, D. Manfredi, J. Miller, G. Nave, L. Redford, I. Schlitz, K. Schmidt, J. L. M. Skorinko, D. Storage, T. Swanson L. M. Van Swol, L. A. Vaughn, B. Wiggins, and A. J. Brady adapted materials for their individual site and collected data. R. A. Klein analyzed the data. R. A. Klein, C. L. Cook, C. R. Ebersole, C. Vitiello, B. A. Nosek, and K. A. Ratliff drafted the report. All authors reviewed, edited, and approved the manuscript for submission.

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

A substantial proportion of replications in recent large-scale efforts have failed to support the original finding. Less than 40% of 100 replications from three top psychology journals were considered successful across a variety of criteria (Open Science Collaboration, 2015). Likewise, 13 of 21 social science experiments published in the journals *Nature* or *Science* were interpreted as replicating successfully based on observing statistical evidence ($p < .05$) in the same direction as the original study (Camerer et al., 2018). Using similar criteria, the "Many Labs" projects conducted high-powered replications of non-randomly sampled studies, yielding variable rates of success: Many Labs 1 (Klein et al., 2014): 10/13, Many Labs 2 (Klein et al., 2018): 14/28, Many Labs 3 (Ebersole et al., 2016): 3/10. There are many possible contributors to failures to replicate, including publication bias or *p*-hacking in original research, unidentified moderators that differ between the original and replication, and failures to effectively implement the replication studies. Shortcomings in implementation could occur by failing to transfer knowledge of key features of the study methods to replication teams.

We sought to evaluate whether involving original authors in selection and design could improve replication success. To do so, we selected an important theory that is also considered sensitive to the procedures necessary to elicit its effects: Terror Management Theory (TMT; Greenberg et al., 1986, 1994). TMT considers the psychological consequences of being reminded of one's impending death. This topic has spawned hundreds of publications, some with more than 1,000 citations. TMT states that as humans evolved self-awareness, they also came to know that their death is inevitable. To avoid preoccupation with thoughts of death or feelings of meaninglessness, one must manage the potential terror caused by this knowledge (Becker, 1973;

1975). Greenberg, Pyszczynski, and Solomon (1986) proposed that self-esteem is the buffer to these intrusive thoughts, and that the purpose of self-esteem is to manage terror related to mortality. TMT has been applied to understand human activities such as religion (e.g., belief in an afterlife grants literal immortality, alleviating mortality terror; Jonas & Fischer, 2006), cultural identity (e.g., belief in being part of a greater good that will persist after death; Greenberg et al., 1990), and inter-group conflict (Burke, Martens, & Faucher, 2010). TMT experts also indicated that there was substantial nuance required in implementing a successful TMT study, and at least one doubted it could be captured in a "Many Labs" style project at all. These nuances include how the experimenter delivers the experimental script (tone, manner), precise design of materials, accounting for contextual details, and other aspects which were refined over years. However, there has never been a systematic investigation of the replicability of TMT findings.

To investigate whether author involvement could improve replicability, we engaged experts in three phases of research design. First, we conducted a community-based search for an important area of research that requires expertise for effective implementation and had experts willing to contribute to replication. TMT met these criteria.

Second, we consulted with TMT experts -- Tom Pyszczynski and Sheldon Solomon -- to identify a seminal study appropriate for replication (Jeff Greenberg later contributed to the design of the study materials). Candidate studies had to take 30 minutes or less, be administered on pencil-and-paper or computerized, and preferably have two or fewer between-subjects conditions to maximize statistical power. We sought studies that had room for researcher flexibility in design, were theoretically central to TMT, and had reasonably high expectations of replicability. In consultation with experts, we selected Study 1 from Greenberg, Pyszczynski,

Solomon, Simon and Breus (1994). This paper is highly influential (985 citations on Google Scholar as of May 24, 2019), and Study 1 used a prototypical mortality salience manipulation: writing about one's own death, which 79.8% of studies in the field of TMT have used (Burke, Martens, & Faucher, 2010).

Third, original authors were instrumental for designing the "Author Advised" version of the replication protocol. The goal was to leverage all possible expertise to design a study with the greatest chance of replicating the original effect. All parts of the description, implementation, and analysis plan for this Author Advised protocol were reviewed by at least one original author before data collection began.

To investigate whether author involvement increases replicability, we recruited 21 teams and randomly assigned them to either administer the Author Advised protocol or to develop and administer their own "In House" protocol to samples that they recruited. In House protocols were designed without contact with the original authors, other teams participating in the project, or other domain experts. Each lab was responsible for collecting data from at least 80 participants to ensure a high degree of statistical power when aggregating across sites[1]. At some labs, when separate principal investigators were identified to administer independent replications, both In House and Author Advised protocols were conducted.

With this design, we examined the following questions:

1. **Can we successfully replicate a central finding supporting TMT?** We replicated Study 1 of Greenberg et al. (1994) in 21 labs with 2,220 participants. At minimum, this provides a high-powered test of this important finding.

[1] One site was excluded for collecting substantially less than this target (N = 19 participants).

2. **Does original author involvement improve replicability?** The selected study (Greenberg et al., 1994, Study 1) left ample room for non-experts to miss potentially critical aspects of implementation. For example, the original paper deviated somewhat in materials from the eventual Author Advised replication implementation and contained little of the advice regarding lab context and setup. We tested whether the Author Advised protocol was more effective at eliciting statistically significant and larger effect magnitudes than In House protocols.
3. **Does a standardized, author-reviewed protocol produce less variability across data collection sites?** We surmised that a standardized implementation of the Author Advised protocol across sites would result in less variation in results compared to the In House protocols because the In House protocols would likely have substantial variability in procedures.

Method

Study 1 of Greenberg et al. (1994) provided evidence that reminders of death induce worldview defense, and that this effect was stronger when the reminders of death were subtle than when reminders of death were more obvious. In the original study, a total of 59 introductory psychology students in the United States were randomly assigned between five conditions. The replication focused on the two conditions that were most effective in the original paper: the "subtle own death salient" condition and the "TV salient" (control) condition. The outcome of interest was evaluations of pro- vs. anti-American essay authors. Specifically, participants in the death salient condition reported a greater preference for the pro-American essay author over the anti-American essay author, compared to participants in the control condition.

Participants in both conditions were told that the study session was comprised of two separate studies. In the "first study," participants completed two filler measures and the mortality salience or control manipulation. 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. Participants in the "TV salient" condition received similar writing prompts, but instead described the emotions they would experience while watching television, and what they thought happened to their physical body as they watched television. Participants then completed the Positive and Negative Affect Schedule (PANAS; Watson, Clark, & Tellegen, 1988).

In a so-called "unrelated second study" during the same session, participants read two brief essays ostensibly written by international students. One essay was pro-American and the other was anti-American. For the dependent variables, participants answered three questions evaluating each essay's author, and two questions evaluating the essay itself. The present replication focused on evaluations of the author because that showed the strongest effect in the original study. A composite rating of the author was created by subtracting the mean of the three anti-American items from the mean of the three pro-American items. A pairwise comparison revealed that participants in the "subtle own death salient" condition showed a greater preference for the pro-American author than for the anti-American author (*Mean Diff* = 12.25). This was compared to participants in the "TV salient" (control) condition (*Mean Diff* = 1.64), $t(53) = 4.87$, $p < .001$, Cohen's $d = 1.34$.

Notable Differences from Original. The replication included just two of the five conditions from the original ("subtle own death salient" and "TV salient") and focused on just

the evaluations of the essay authors (disregarding any evaluations of the essays themselves). The replication studies were administered in the US Fall of 2016 and Spring of 2017.

Procedure for Author Advised Protocol. All materials and instructions used in the study are available on the OSF page (https://osf.io/bq4n4/), as well as instructions and detailed procedures and advice provided to Author Advised sites. For the replication, we made several changes to the original protocol as requested by the original authors:

1. The Anti-American essay was adjusted to be more forceful and extreme, to ensure it conflicted with the average participant's worldview.
2. The filler-task portion of the study was extended to include additional surveys and increase the delay between the mortality salience induction and essay evaluation.
3. The procedure tried to put participants into a relaxed, experiential mood. Precise language was included in the expert instructions packets and included recommendations such as selectively choosing relaxed research assistants to administer the study, using a covered box for handing in packets to ensure a feeling of confidentiality, separating participants into different cubicles or rooms, and having experimenters act and dress casually.
4. Additional demographic items were added to the end of the survey to apply exclusion criteria.

The Author Advised protocol was always administered in person and in individual cubicles or small groups. When participants first entered the room, they were read a prepared script that included a cover story that the study consisted of two parts, and statements assuring participants of the anonymity of their responses to reduce demand characteristics. Participants were instructed to complete the first half of the survey, and then return the completed form to a

covered box before being given the second survey. Then, participants were seated and given the second packet of materials.

The first materials packet started by reassuring participants that their responses were confidential, and that they should respond naturally: “On the following pages you will find a series of personality, attitude and judgment questionnaires. There are no right or wrong, or good or bad answers; rather different responses reflect different personalities, attitudes and judgment styles. Please respond honestly and naturally to each question and complete the questionnaires in the order that they appear in the packet. Your responses to these questions are completely anonymous and will be used for research purposes only.” Then, participants completed a 23-item “personality inventory” included as a distractor, in which participants answered “Yes” or “No” to items such as, “Does your mood often go up or down?” They also completed a 12-item measure modeled after the Personal Need for Structure Scale (Thompson et al., 2001), which was also included primarily as a distractor. Participants responded on a 6-point scale (1 – strongly disagree to 6 – strongly agree) to items like “It upsets me to go into a situation without knowing what I can expect from it.” After this, participants completed the mortality salience or TV (control) induction.

Prior to the induction text, participants were again reminded to respond naturally: “On the following page there are a couple of open-ended questions. Please respond to them with your first, natural response. We are just looking for people's gut-level reactions to these questions.” Participants in the mortality salience condition then responded to two open-ended items disguised as “The Projective Life Attitudes Assessment.” The first item asked participants to “Please briefly describe the emotions that the thought of your own death arouses in you,” while the second item asked participants to “Jot down, as specifically as you can, what you think will

happen to you physically as you die and once you are physically dead." Participants in the TV (control) condition responded to nearly identical items, which instead asked about their emotions experienced while watching television and what happens to them as they watch television. Participants then completed the PANAS-X (Watson & Clark, 1994), a 60-item measure of current emotional state in which they indicated the degree to which they currently felt each of 60 emotions on a 5-point scale (1 – very slightly or not at all to 5 – extremely), and the Morningness-Eveningness Questionnaire (Horne & Ostberg, 1975) which is a 19-item scale assessing the degree to which participants performed better in the morning or evening. Example items included: "What time would you get up if you were entirely free to plan your day?" and, "What time would you go to bed if you were entirely free to plan your evening?" The purpose of these two scales was to provide "filler" time between the induction and the essay evaluations. This concluded the "first half" of the study and the first materials packets.

After the participant handed in the first packet, they were provided the second packet of materials, matched by a participant number. Participants read a cover story that indicated the university had collected writing samples of impressions of America from international students attending the university. Participants then read two essays, in counterbalanced order across participants. One essay was relatively pro-U.S. and the other was relatively critical of the U.S. Participants rated each essay and its author on a five-item, 9-point scale (1 = "not at all" to 9 = "extremely). Example items include "To what extent do you think the essay makes valid points?" and "How intelligent do you think the person who wrote this essay is?".

Procedures for In House Protocols. The In House protocols were created independently by each participating lab. Instructions provided to these labs are available on the OSF page (https://osf.io/drfg2/). To the best of their ability, and using only the original paper, relevant

literature, and any other publicly available resources, each lab designed their own replication protocol of the assigned study. The labs were prohibited from contacting the original authors, other experts, and the other participating labs. The only outside review was by the project leaders to confirm that the correct study was being replicated and that the resulting data collection would yield all variables necessary for the basic analysis plan.

The In House protocols differed substantially, both from one another and as compared to the standardized Author Advised protocol. For example, 7 In House labs used the PANAS as a filler task between the manipulation and the dependent variables; the other 5 labs included no filler task at all. None of the In House labs used the same set of filler tasks as the Author Advised version, which included the PANAS and second unrelated questionnaire. All 12 In House labs collected data using a computer instead of pencil-and-paper, a design explicitly discouraged by the original authors for the Author Advised version. We summarize major attributes of the designs of all In House sites in Table 2. All materials used at In House sites and videos documenting their implementation are available at http://osf.io/8ccnw/ to facilitate review of the varying procedural implementations.

Study Administration. Within each lab, participants were randomly assigned to either the mortality salience or control condition of the given protocol. At three universities, both protocols were administered by independent groups of researchers (to maintain blinding to the Author Advised materials). At those sites, the groups recruited participants separately, but randomized participants between conditions in their given protocol. Each collaborating lab was responsible for recruiting at least 80 participants for their assigned protocol. Participants were prohibited from participating in the study more than once.

Provisions for Quality Control

As in previous iterations of the "Many Labs" projects (e.g., Klein et al., 2014; Ebersole et al., 2016; Klein et al., 2018), we employed quality control and accountability standards for each lab's data collection. Each site was required to film a mock session of the data collection in which the researcher filmed a mock participant walking through the steps as if they were actually taking the study. These videos allowed a more thorough documentation of the lab spaces and any idiosyncratic differences in protocol. In addition, sites were responsible for ensuring all experimenters in direct contact with participants were blind to the participant's condition assignment. Collaborators also completed an experimenter survey assessing their expertise, expectations, and motivations for joining the collaboration. The survey and videos are available at https://osf.io/8ccnw/.

At the end of the Author Advised protocol, participants filled out demographic information. Three of these items were included at the request of original authors: importance of American identity, country of birth, and race. In House protocols were asked to collect similar demographic information (see https://osf.io/drfg2/), but these were described minimally to avoid influencing design decisions. The American identity item was omitted entirely. As a result, demographic data from In House sites were not entirely comparable to demographic data from Author Advised sites.

Participants

A total of 21 labs participated and provided a total sample of 2,281 participants. In accordance with the pre-registration (https://osf.io/4xx6w), we immediately excluded from all analyses participants who either failed to complete all 6 ratings of the essay authors, or who failed to complete both writing prompts within the mortality salience or control conditions (e.g.,

the between-subjects manipulation). The latter exclusion criteria applied only to participants from Author Advised sites, because the necessary data was not always available for In House sites. Thus, the usable sample included 2,220 participants (see Table 1 for a summary of sites). 1,157 participants (52.12%) reported being female and 708 participants (31.89%) reported being male; the remaining participants did not respond to the item, were asked about gender in a non-standard way, or chose a different response. The mean age was 19.87 years (*SD* = 2.79). Participant reported race was 910 (40.99%) White, 221 (9.95%) Asian, 120 (5.41%) Black or African American, 36 (1.62%) American Indian or Alaska Native, 20 (0.90%) Native Hawaiian or Pacific Islander, 114 (5.14%) Other. The remaining participants did not report their race, or responses were not easily recoded to match these categories.

Table 1

Data collection sites

Location	Site Identifier	Author Advised (AA) or In House (IH)	*N*
1. Ashland University, OH	Ashland	AA	56
2. Azusa Pacific University, CA	Azusa	IH	30
3. Brigham Young University – Idaho, ID	Byui	AA	81
4. Ithaca College, NY	Ithaca	IH	177
5. Occidental College, CA	Occid	AA	88
6. Pace University, NY	Pace_expert	AA	106
7. Pace University, NY	Pace_inhouse	IH	58
8. Pacific Lutheran University, WA	Plu	IH	246
9. Southern Oregon University, OR	Sou_inhouse	IH	29
10. The College of New Jersey, NJ	Cnj	AA	136
11. University of California, Riverside, CA	Riverside	AA	107
12. University of Florida, FL	Ufl	IH	252
13. University of Illinois at Urbana-Champaign, IL	Illinois	IH	87
14. University of Kansas, KS	Kansas_expert	AA	43
15. University of Kansas, KS	Kansas_inhouse	IH	75
16. University of Pennsylvania, PA	Upenn	IH	86
17. University of Wisconsin, Madison, WI	Uwmadison_expert	AA	79
18. University of Wisconsin, Madison, WI	Uwmadison_inhouse	IH	68
19. Virginia Commonwealth University, VA	Vcu	AA	103
20. Wesleyan University, CT	Wesleyan_inhouse	IH	174
21. Worcester Polytechnic Institute, MA	Wpi	IH	139

Note: Location numbers rather than site names will be used on subsequent tables.

Table 2

Variability in methods across In House labs compared to the Author Advised standard (AA)

	University												
Protocol	2	4	7	8	9	12	13	15	16	18	20	21	AA
Filler task(s) before mortality salience:	X	X	X		X		X		X	X	X	X	
Rosenburg Self Esteem scale	X	X	X		X		X		X	X	X		
Neuroticism scale[a]	X				X				X	X	X		
TIPI[b]		X	X										
BFI or shortened BFI[c]							X					X	
Dot task[d]												X	
PNS scale[e]													
Filler task after mortality salience:		X		X				X	X	X		X	
Mood scale									X				
PANAS		X	X	X				X	X	X		X	
MEQ[f]													
Picture presented after mortality salience		X							X				
Essays attributed to student author	X	X	X	X	X		X	X	X	X	X		X
Computer data collection	X	X	X	X	X	X	X	X	X	X	X	X	
Paper data collection													
Completed in the lab		X		X			X		X	X	X		X
Completed online	X		X		X	X		X				X	X
College student sample	X	X	X	X	X	X	X	X	X	X	X	X	X
Used essays from Greenberg et al. (1992)				X		X		X	X		X	X	

[a]From the Eysenck Personality Inventory [b]Ten Item Personality Inventory [c]Big Five Inventory [d]Participants instructed to place 10 dots among shapes [e]Personal Need for Structure Scale [f]Morning-Eveningness Questionnaire

Analysis Plan

The primary finding of interest from Greenberg et al., (1994) was that participants who underwent the mortality salience treatment showed greater preference for the pro-US essay author over the anti-US essay author compared to the control condition. To assess whether the replication results support the original, we followed a similar analysis plan as in the original article. Scores from the three items evaluating the authors of the anti-American essays were averaged (α = 0.90) and then subtracted from the average of the three items evaluating authors of the pro-American essays (α = 0.89).[2] An independent-samples *t*-test was then conducted comparing those in the "subtle own death salient" (MS) condition with scores from the "TV salient" (control) condition. Some labs administered both Author Advised and In House protocols. To account for this nesting of effect sizes within labs, a three-level random-effects meta-analysis was conducted using the MetaSEM package (Cheung, 2014) in R (R Core Team, 2019).

Original authors were not entirely in agreement about what exclusions should be implemented. So, we repeated our analyses under different exclusion criteria suggested by original authors:

Exclusion Set 1: Include all participants who completed the materials (e.g., wrote something for both writing prompts, and completed all six items evaluating the essay authors).

[2] Supplemental analyses treating these as two separate dependent variables are available in the online supplement (https://osf.io/xtg4u/), and those outcomes do not qualify the conclusions offered here.

Reduces the usable N from 2,281 to 2,220 participants. This sample size gives us 95% power to detect a condition effect of d = .15 in an independent samples t-test.

Exclusion Set 2: All prior exclusions, and further exclude participants who did not identify as White or who indicated they were born outside the United States. Reduces N to 1,874. This sample size gives us 95% power to detect a condition effect of d = .16.

Exclusion Set 3: All prior exclusions, and further exclude participants who responded lower than 7 on the American Identity item ("How important to you is your identity as an American?" 1 - not at all important; 9 - extremely important). Further reduces the usable N to 1,693 participants. This sample size gives us 95% power to detect a condition effect of d = .18.

Exclusion Sets 2 and 3 were specifically recommended by original authors and these criteria were used to analyze the data from Author Advised labs. However, the data required to make these exclusions were often not collected at In House replication sites because they made independent decisions about design and demographic measures for potential exclusion, and these measures were not in the original article. Thus, for all analyses only Exclusion Set 1 was used for In House participants. All data handling, exclusions, and computation of results within sites followed our pre-registered (prior to data collection) analysis plan on the OSF (https://osf.io/4xx6w).

Results

Researcher Expectations and Characteristics

A total of 28 researchers from 21 participating sites completed an experimenter survey about their motivations and expertise. This survey was administered during data collection, and although no researcher had access to overall project-wide results, ~⅓ of the researchers reported looking at or analyzing their own site's data prior to completing the survey. Psychology research

experience ranged from 0 to 28 years (M = 9.32, SD = 8.80). One (4%) researcher indicated they were an expert in TMT, five (18%) indicated they had "a lot" of TMT knowledge, ten (36%) indicated "some" knowledge, five (18%) indicated little knowledge, six (21%) indicated zero knowledge, and one (4%) did not respond to the question.

When asked what outcome they wanted to happen, 13 (46%) indicated that they hoped for the project to successfully replicate the TMT effect, ten (36%) indicated no preference, and three (11%) hoped the project would result in a failure to replicate, with two (7%) researchers leaving the question blank. On average, the teams estimated a 54% chance of successful replication with a wide range of estimates from 20% to 95% (SD = 22.14).[3]

Deviations from Pre-registered Analytic Plan

Our pre-registered analytic plan specifies the use of a three-level meta-analysis, conducted in the MetaSEM R package (Cheung, 2014), to control for the clustering of effect sizes when independent teams ran both In House and Author Advised versions of the protocol at the same university. However, during data analysis we discovered that we did not have enough data in these clusters for the planned analysis to be accurate (e.g., we had to drop the clustering variable). The results reported below are thus a more common univariate meta-analysis conducted in the same package, which is the model that most closely mirrors our originally planned analysis. The entire results section and analysis code written to report the originally planned three-level model are available on the OSF (https://osf.io/8ccnw/), and no conclusions substantively change between the current model and the three-level model.

[3] Including only sites that had not looked at any data, researchers estimated a 56% chance of successful replication.

Research Question 1: Meta-analytic results across all labs (random effects meta-analysis).

The most basic question is whether we observed the predicted effect of mortality salience on preference for pro- vs anti- American essay authors. To assess this we conducted a random-effects meta-analysis. This analysis produces the grand mean effect size across all sites and versions. Regardless of which exclusion criteria were used, we did not observe the predicted effect and the confidence interval was quite narrow: Exclusion Set 1: *Hedges' g* = 0.03, 95% CI = [-0.06, 0.12], *SE* = 0.05, *Z* = 0.55, *p* = 0.58. Exclusion Set 2: *Hedges' g* = 0.06, 95% CI = [-0.06, 0.17], *SE* = 0.06, *Z* = 0.99, *p* = 0.32. Exclusion Set 3: *Hedges' g* = 0.04, 95% CI = [-0.07, 0.16], *SE* = 0.06, *Z* = 0.71, *p* = 0.48. Forest plots showing the effects for individual sites and the aggregate are available in Figure 1 for Exclusion Set 1 (see https://osf.io/8ccnw/ for the other two Exclusion Sets).

There may have been a mortality salience effect at some sites and not others, so we next examined how much variation was observed among effect sizes (e.g., heterogeneity). For Exclusion Sets 1 and 3, this sort of variation did not exceed variation expected by chance (e.g., sampling variance): Exclusion Set 1: $Q(20) = 25.82$, $p = 0.17$; Exclusion Set 3: $Q(20) = 29.61$, $p = 0.08$. The amount of variation between sites did exceed chance for Exclusion Set 2, $Q(20) = 36.32$, $p = 0.01$, however it was small in magnitude, $Tau^2 = 0.02$. Results across all individual sites using Exclusion Set 1 are presented in Table 3.[4]

In sum, we observed little evidence for an overall effect of mortality salience in these replications. And, overall results suggest that there was minimal or no heterogeneity in effect sizes across sites. This lack of variation suggests that it is unlikely we will observe an effect of

[4] Results for other exclusion criteria are available at https://osf.io/xtg4u/

Author Advised versus In House protocols or other moderators such as differences in samples or TMT knowledge. Even so, the plausible moderation by Author Advised/In House protocol is examined in the following section.

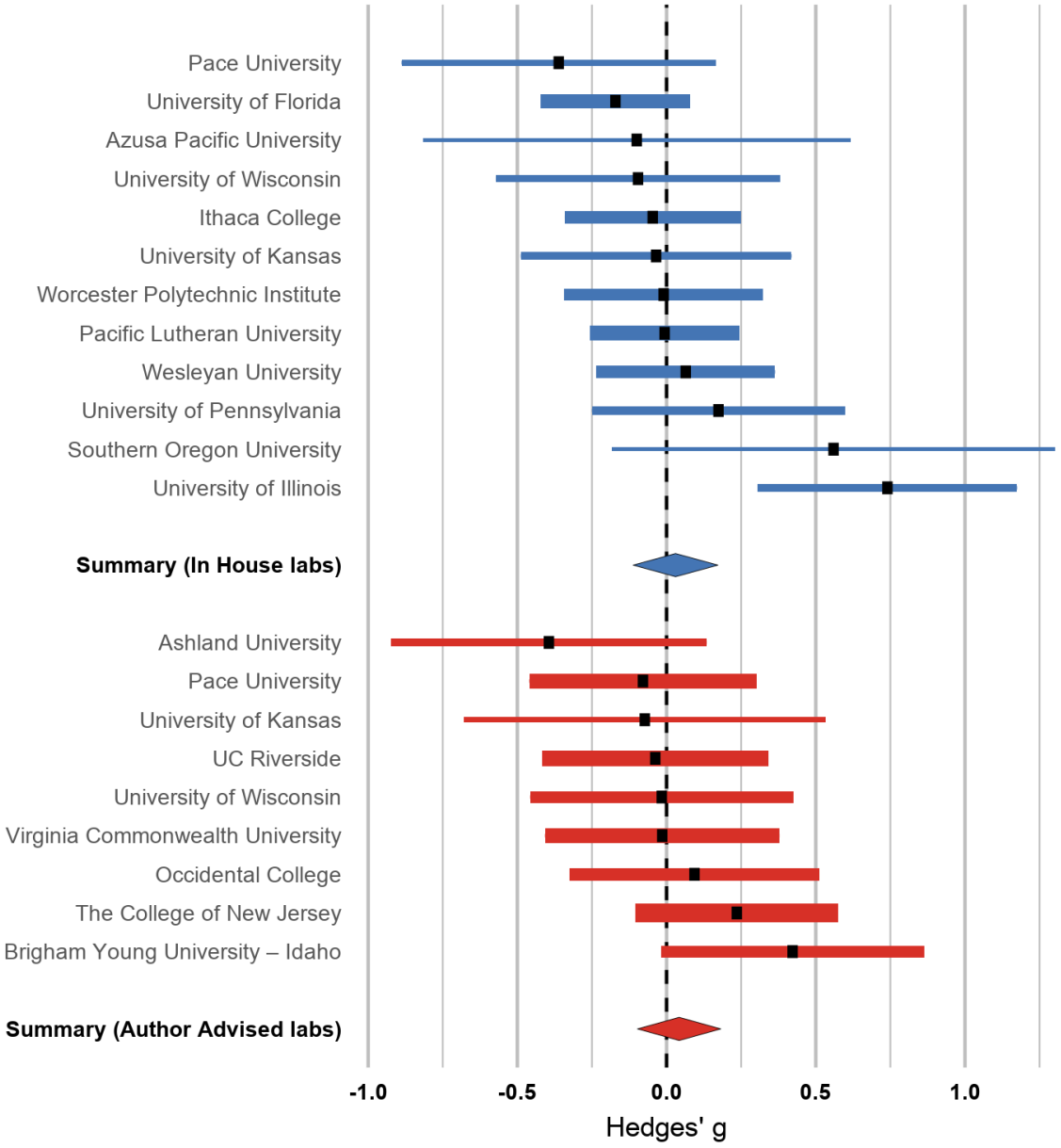

Figure 1. Forest plot summarizing results from all sites using Exclusion Set 1. Error bars indicate the 95% confidence interval, and thickness is scaled by N at that lab. Blue shading indicates an In House site, red shading indicates an Author Advised site. Diamonds indicate the aggregated

result across that subset of labs. Forest plots for the other exclusion sets are available on the OSF page (https://osf.io/xtg4u/).

Table 3

Results per site using Exclusion Set 1

Location	Author Advised (AA) or In House (IH)	N (TV)	N (MS)	Mean (TV)	Mean (MS)	SD (TV)	SD (MS)	Hedges' g	p
1	AA	28	28	1.88	1.14	1.75	1.93	-0.39	0.14
2	IH	15	15	0.22	0.09	1.01	1.54	-0.1	0.78
3	AA	41	40	1.77	2.65	2.06	2.06	0.42	0.06
4	IH	87	90	-0.41	-0.53	2.14	2.78	-0.05	0.76
5	AA	42	46	0.33	0.52	1.66	2.26	0.09	0.66
6	AA	53	53	1.33	1.18	1.89	2.05	-0.08	0.68
7	IH	34	24	2.2	0.11	5.93	5.33	-0.36	0.17
8	IH	125	121	0.37	0.36	1.85	1.97	-0.01	0.96
9	IH	14	15	-1.02	-0.18	1.68	1.25	0.56	0.14
10	AA	60	76	1.14	1.57	1.96	1.69	0.24	0.18
11	AA	52	55	0.79	0.73	1.68	1.72	-0.04	0.84
12	IH	107	145	0.99	0.76	1.47	1.23	-0.17	0.19
13	IH	44	43	-0.78	0.64	1.8	2.01	0.74	0
14	AA	18	25	1.07	0.89	2.31	2.51	-0.07	0.81
15	IH	40	35	1.32	1.21	3.03	3.04	-0.03	0.88
16	IH	45	41	-0.16	0.2	1.75	2.23	0.17	0.42
17	AA	39	40	0.94	0.91	1.72	2.16	-0.02	0.94
18	IH	36	32	-0.74	-0.93	2.05	1.77	-0.1	0.69
19	AA	42	61	1.33	1.3	1.15	1.99	-0.01	0.94
20	IH	97	77	-0.2	-0.09	1.81	1.78	0.06	0.67

21	IH	68	71	0.53	0.51	1.96	1.61	-0.01	0.95

Research Question 2: Moderation by Author Advised/In House protocol

A covariate of protocol type (In House vs Author Advised) was added to the random effects model to create a mixed-effects meta-analysis. This is our primary model of interest, and the model most similar to the three-level mixed-effects meta-analysis that we pre-registered as our primary outcome.

This analysis again produces an overall grand mean effect size, and those were again near zero and relatively precisely estimated across all three Exclusion Sets: Exclusion Set 1: *Hedges' g* = 0.01, 95% CI = [-0.10, 0.13], *SE* = 0.06, *Z* = 0.25, *p* = 0.80. Exclusion Set 2: *Hedges' g* = 0.02, 95% CI = [-0.11, 0.15], *SE* = 0.07, *Z* = 0.35, *p* = 0.73. Exclusion Set 3: *Hedges' g* = 0.02, 95% CI = [-0.10, 0.15], *SE* = 0.06, *Z* = 0.33, *p* = 0.74.

Variation among effect sizes also followed the previously observed pattern. Weak heterogeneity for Exclusion Set 2, *Q*(20) = 36.32, *p* = 0.01, Tau^2 = 0.01; while variation did not meet the statistical significance threshold for Exclusion Set 1 *Q*(20) = 25.82, *p* = 0.17; or Exclusion Set 3: *Q*(20) = 29.61, *p* = 0.08.

Critically, protocol version did not significantly predict replication effect size regardless of which exclusion criteria were used. Exclusion Set 1: *b* = 0.03, *Z* = 0.28, *p* = 0.78; Exclusion Set 2: *b* = 0.11, *Z* = 0.91, *p* = 0.36; Exclusion Set 3: *b* = 0.09, *Z* = 0.67, *p* = 0.50. The Author Advised version did not produce larger effect sizes when compared with the In House versions.

Research Question 3: Effect of Standardization

Finally, we examined whether In House protocols displayed greater variability in effect size than Author Advised protocols. We outlined this hypothesis in our pre-registration, but the methods for testing it are exploratory.

As an initial test, we conducted separate meta-analyses for the In House and Author Advised labs. For each, we conducted both a fixed-effects (with variance between labs constrained to be equal to zero) and random-effects meta-analysis, and then compared the two models with a chi-squared differences test to assess whether the fit significantly changed. If the random-effects model fit significantly better than the fixed-effects model, this would indicate that allowing for variability in effect sizes between sites improved the model.

In this case, neither In House nor Author Advised labs showed a significant benefit of the random effects model over the fixed effects model across any of the Exclusion Sets: In House labs: Exclusion Set 1: χ^2 (1) = 0.29, p = 0.59; Author Advised labs: Exclusion Set 1: χ^2 (1) = 0.00, p = 1; Exclusion Set 2: χ^2 (1) = 0.03, p = 0.87; Exclusion Set 3: χ^2 (1) = 0.00, p = 1. Overall, this evidence indicates that neither In House nor Author Advised labs showed significant variability in effect size across sites, despite the fact that In House labs were unambiguously more variable in their procedural implementation. This does not mean the variances were equal, but based on the present evidence we cannot conclude that they were different.

Follow-Up Exploratory Analyses

Results for TMT-knowledgeable sites. One principal investigator reported being an expert in TMT, while five others indicated having “a lot” of knowledge about TMT. One might expect that these locations would have greater success at replicating the mortality salience effect.

Aggregating across these sites, and using only the first exclusion rule, these sites did not elicit a larger difference between the mortality salience group (M = 1.02, SD = 2.30) and the control group (M = 0.93, SD = 2.30), t(520.81) = 0.43, p = 0.67, *Hedges' g* = 0.04, 95% CI = [-0.13, 0.21].[5]

Results for participants who preferred the pro-US author The present hypothesis that mortality salience would cause a participant to become more favorable to the pro-US author as compared to the anti-US author relies on the participant perceiving the pro-US stance as more similar to their own worldview (and/or the anti-US stance as threatening to their worldview). Original authors anticipated that the essays from the original study may not serve this function in the replication, run in 2016. For this reason, the anti-US essay from the original study was made more extreme in the Author Advised version of the replication. There was a particular concern that in the months leading up to and following the 2016 US Presidential Election of Donald Trump, the generally more liberal-leaning student bodies on college campuses may feel less patriotic and not identify with the pro-US worldview. Indeed, analysis suggests the original authors anticipated and more successfully addressed this issue. Among In House replications, 49% of participants prefered the pro-US essay author, 40% preferred the anti-US essay author, and 11% had no preference. Among Author Advised replications, 68% of participants prefered the pro-US essay author, 22% preferred the anti-US essay author, and 10% had no preference.

However, the predicted mortality salience effect was not larger or detectable via statistical significance when subsetting to only participants at Author Advised sites who preferred the pro-

[5] One site, UW Madison In House, used a 7-point scale. This has been rescaled to a 9-point scale for this analysis to approximately compare it with the others.

US author. In all exclusion sets, the mortality salience and control groups showed similar levels of preference for the pro-US author over the anti-US author: Exclusion Set 1: mortality salience group (M = 1.23, SD = 2.06), control group (M = 1.15, SD = 1.83), t(796.99) = 0.58, p = 0.56, *Hedges' g* = 0.04, 95% CI = [-0.10, 0.18]; Exclusion Set 2: mortality salience group (M = 1.53, SD = 2.15), control group (M = 1.38, SD = 1.97), t(446.51) = 0.79, p = 0.43, *Hedges' g* = 0.07, 95% CI = [-0.11, 0.26]; Exclusion Set 3: mortality salience group (M = 1.96, SD = 2.15), control group (M = 1.83, SD = 2.05), t(264.51) = 0.49, p = 0.62, *Hedges' g* = 0.06, 95% CI = [-0.18, 0.30]. The confidence intervals were wider because of the smaller total sample size, but this evidence is not consistent with the hypothesis that preference for the pro-US author would elicit an effect of mortality salience in this context.

Discussion

We conducted a high-powered investigation of the replicability of a classic finding supporting Terror Management Theory (Greenberg et al., 1994; Study 1). With 21 labs contributing usable data from 2,220 participants, we observed little evidence that priming mortality salience increased worldview defense compared to a control condition (*Hedges' g* = 0.06, 95% CI = [-0.07, 0.16]), and the narrow confidence interval suggests, if anything, that the effect is very small. We intended to use this paradigm to evaluate whether expert advising on the research protocol would increase effect sizes and overall replication success compared to independent In House replication attempts. However, neither the Author Advised or In House protocols successfully replicated the original finding. Moreover, we did not observe greater variability in effect sizes across sites using In House protocols despite their procedural variability compared to the standardized Author Advised protocol. With these protocols, in the context of these labs and time in history, we find little support for this key finding of TMT.

This failure to replicate is quite precisely estimated, but it does not mean that the original effect was necessarily a false positive. It could be that changes in history have substantially altered the observability of the effect (original d = 1.34; replication d = 0.06). This null effect also does not necessarily mean that this method of invoking mortality salience is ineffective. It could be effective on other outcomes or in other circumstances. If so, we would need evidence for why it did not occur under these circumstances, and updating of the theory for this boundary condition. Finally, TMT is supported by a network of evidence using a variety of procedures and testing a variety of claims. A single failure to replicate, no matter how precisely estimated, does not overturn all that prior work. The present evidence does, however, provide an important challenge for TMT to address. The study was designed with feedback from experts who suggested the study to use because of its centrality to TMT. Moreover, the study was highly powered and used a preregistered analysis plan to maximize diagnosticity of the statistical inferences (Nosek et al., 2018). In addition, another recent pre-registered replication also failed to find support for TMT (Sætrevik & Sjåstad, 2019). Effective counterevidence to this challenge to the reliability of TMT would be new evidence to show that the finding can be reliably replicated under other conditions, and direct evidence of a boundary condition that the present study identified inadvertently (Nosek & Errington, 2019).

For example, at least one original author proposed prior to the study that the timing of the replication—September 2016 to May 2017, the period leading up to and following the election of Donald Trump as President of the United States—may result in a failure to observe the mortality salience effect on worldview defense. In essence, students at U.S. universities, which tend to be more liberal, may have been experiencing perpetually threatened worldviews which would decrease the likelihood of demonstrating the mortality salience effect. While it remains possible

that idiosyncrasies of the time period may have decreased the effect, we sought to address this concern in three ways: (1) the anti-US essay was made more extreme in its criticism, making it less likely to appeal to most Americans, (2) the most strict exclusion criteria selected for only highly patriotic Americans, and (3) we analyzed and reported exploratory results including only participants who indicated a preference for the pro-US author. Moreover, the decline in effect size was dramatic (original $d = 1.34$; replication $d = 0.06$). Nevertheless, it is possible to collect new evidence to test this speculation directly.

It is also possible that this effect requires more expertise than the present design could deliver—for instance, one original author indicated that he was not comfortable endorsing any replication attempt that he did not directly supervise. Advice provided to Author Advised labs was sometimes subtle, making it difficult to ensure correct implementation. For example, labs were told to induce a casual, relaxed mindset in participants by, for example, selecting laid back research assistants, dressing casually, and using less formal lab space. Each site recorded a video of data collection to try to document these subtleties, but it is difficult to assess how successfully this context was induced. However, in exploratory analysis, even highly TMT-knowledgeable sites failed to replicate the mortality salience effect in the present project.

The failure to replicate undermined our goal to test for an expertise effect on replicability. We selected this study because of the belief that the finding was likely replicable, but subtle, so we could study the extent to which expertise matters. Without any effect, we cannot evaluate whether expertise matters. We did observe procedural differences between the In House and Author Advised protocols, including features that were deemed critical by the original authors. For example, original authors agreed that in-lab demonstrations would produce larger effects, but

some In House protocols were web-based. Because these differences did not matter in this case we cannot conclude much about the role of expertise for replicable phenomena.

Perhaps most importantly, the present project underscores the need for replication even in areas of research with large bodies of published evidence. If expert advice on central findings of large literatures is insufficient to ensure replicability, then there is still a good deal of work to do to establish replicability of at least some apparently robust findings. Improvements to descriptions and sharing of methods, power of research designs, preregistration of studies and analysis plans, and peer review by experts, may all contribute to improving the robustness and reliability of research findings and their theoretical explanation.

References

Becker, E. (1973). *The denial of death*. New York: Free Press.

Becker, E. (1975). *Escape from evil*. New York: Free Press.

Burke, B. L., Martens, A., & Faucher, E. H. (2010). Two decades of terror management theory: A meta-analysis of mortality salience research. *Personality and Social Psychology Review, 14*(2), 155-195.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637.

Cheung, M. W. L. (2014). MetaSEM: an R package for meta-analysis using structural equation modeling. *Frontiers in Psychology, 5*.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... & Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68-82.

Greenberg, J., Pyszczynski, T., & Solomon, S. (1986). The causes and consequences of a need for self-esteem: A terror management theory. In R. F. Baumeister (Ed.), *Public self and private self* (pp. 189-212). New York, Springer-Verlag.

Greenberg, J., Pyszczynski, T., Solomon, S., Rosenblatt, A., Veeder, M., Kirkland, S., & Lyon, D. (1990). Evidence for terror management theory II: The effects of mortality salience on reactions to those who threaten or bolster the cultural worldview. *Journal of Personality and Social Psychology, 58*(2), 308-318.

Greenberg, J., Pyszczynski, T., Solomon, S., Simon, L., & Breus, M. (1994). Role of consciousness and accessibility of death-related thoughts in mortality salience effects. *Journal of Personality and Social Psychology, 67*(4), 627-637.

Horne, J. A., & Ostberg, O. (1975). A self-assessment questionnaire to determine morningness-eveningness in human circadian rhythms. *International journal of chronobiology*, *4*(2), 97-110.

Jonas, E., & Fischer, P. (2006). Terror management and religion: Evidence that intrinsic religiousness mitigates worldview defense following mortality salience. *Journal of Personality and Social Psychology, 91*(3), 553-567.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, S., Bernstein, M. J., ... Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, *45*(3): 142-152. doi: 10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., . . . Nosek, B. A. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., Mellor, D. T. (2018). Preregistration revolution. *Proceedings of the National Academy of Sciences, 115(11)*, 2600-2606. doi: 10.1073/pnas.1708274114

Nosek, B. A., & Errington, T. E. (2019). What is replication? *MetaArXiv*, https://osf.io/preprints/metaarxiv/u4g6t. Doi: 10.31222/osf.io/u4g6t

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. doi: 10.1126/science.aac4716

R Core Team (2019). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/

Sætrevik, B., & Sjåstad, H. (2019, May 17). Failed pre-registered replication of mortality salience effects in traditional and novel measures. https://doi.org/10.31234/osf.io/dkg53

Thompson, M. M., Naccarato, M. E., Parker, K. C. H., & Moskowitz, G. (2001). *The Personal Need for Structure (PNS) and Personal Fear of Invalidity (PFI) scales: Historical perspectives, present applications and future directions*. In G. Moskowitz (Ed.), Cognitive social psychology: The Princeton symposium on the legacy and future of social cognition (pp. 19-39). Mahwah, NJ: Erlbaum.

Watson, D., & Clark, L. A. (1994). *The PANAS-X: Manual for the Positive and Negative Affect Schedule—Expanded Form* (Unpublished manuscript). Retrieved from https://ir.uiowa.edu/cgi/viewcontent.cgi?article=1011&context=psychology_pubs

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology, 54*(6), 1063-1070.