ML4 Results Section in rMarkdown

Richard A. Klein1

1 Université Grenoble Alpes

Author note

This script generates the participants + results sections for the main ML4 manuscript. To knit this document you must install the papaja package from GitHub.

Correspondence concerning this article should be addressed to Richard A. Klein, . E-mail: [raklein22@gmail.com](mailto:raklein22@gmail.com)

ML4 Results Section in rMarkdown

## Participants

Twenty-one labs from 18 universities participated in the project and collected data from a total of 2,281 participants. Several overall exclusions were made, each as specified by our preregistration (<https://osf.io/4xx6w>). First, data from four labs were excluded for collecting fewer than 60 participants.[[1]](#footnote-22) This excluded 158 participants. Next, some In House sites began data collection before the analysis plan was pre-registered due to their own deadlines. Data collected before the pre-registration were excluded from the confirmatory tests reported in this manuscript. This excluded a further 545 participants.[[2]](#footnote-24)

From the remaining 17 labs and 1,578 participants, in adherence to our pre-registration, we 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 1,550 participants (see Table 1 for a breakdown by lab).

841 participants (54.26%) were women and 442 participants (28.52%) were men; 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 19.91 years (*SD* = 2.49). Participant reported race was 946 (61.03%) White, 252 (16.26%) Asian, 112 (7.23%) Black or African American, 4 (0.26%) American Indian or Alaska Native, 7 (0.45%) Native Hawaiian or Pacific Islander, and 124 (8%) another category. The remaining participants did not report their race or their responses were not easily recoded to match these categories.

## 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 as compared to the control condition. To assess whether the replication results support the original, within each lab we followed a similar analysis plan as in the original article, as specified by our pre-registration (<https://osf.io/4xx6w>). Scores from the three items evaluating the authors of the anti-American essays were averaged ( = 0.89) and then subtracted from the average of the three items evaluating authors of the pro-American essays ( = 0.89). 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. We then analyzed these individual results meta-analytically to get an aggregate effect size across all labs. Supplemental exploratory (non-preregistered) 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.

In addition to the overall exclusions detailed previously in the “Participants” section, original authors requested some additional exclusions based on participant responses. Original authors were not entirely in agreement about what exclusions should be implemented. Therefore, as indicated in the pre-registration, we repeated our analyses under three different exclusion criteria: A minimal set of exclusions (Exclusion Set 1), and two sets of more strict exclusions recommended by original authors (Exclusion Sets 2 and 3).

Because Exclusion Sets 2 and 3 were specifically recommended by original authors and could be considered part of their expertise, we did not inform the In House labs about these exclusions prior to data collection. This masking was to avoid influencing their decisions about data collection procedures, which had to be made independently of any outside help. However, as a result, none of the In House labs collected the data required to make these more strict exclusions because the items required were not present in the original target article.

Therefore, analyses were repeated for each of the three exclusion sets for Author Advised participants. However, Exclusion Set 1 was used consistently across all analyses for In House participants. The exclusion sets consisted of:

*Exclusion Set 1:* Exclude participants who did not complete both writing prompts and all six items evaluating the essay authors. This yields 1,550 participants (*n* = 699 Author Advised, *n* = 851 In House). This aggregate sample size gives us 95% power to detect a mortality salience condition effect of *d* = 0.18 in an independent samples *t*-test.[[3]](#footnote-28)

*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. This reduces the *N* to 1,229 participants (*n* = 378 Author Advised, *n* = 851 In House). This aggregate sample size gives us 95% power to detect a mortality salience condition effect of *d* = 0.21.

*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). This further reduces the usable *N* to 1,076 participants (*n* = 225 Author Advised, *n* = 851 In House). This aggregate sample size gives us 95% power to detect a mortality salience condition effect of *d* = 0.22.

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>). All results are reported with two-tailed tests of significance.[[4]](#footnote-29)

# Results

## 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 these models failed to converge because we did not have enough data within each cluster, as most sites conducted only one study. As such, we had to drop the clustering variable. The results reported below are thus a more common univariate meta-analysis, which is the model that most closely mirrors our originally planned analysis.

## 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* = 0.07, 95% CI = [-0.03, 0.17], *SE* = 0.05, *Z* = 1.32, *p* = .187. Exclusion Set 2: *Hedges’ g* = 0.09, 95% CI = [-0.03, 0.21], *SE* = 0.06, *Z* = 1.49, *p* = .135. Exclusion Set 3: *Hedges’ g* = 0.09, 95% CI = [-0.04, 0.22], *SE* = 0.07, *Z* = 1.30, *p* = .193. 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.) These results indicate that, in the aggregate, we failed to replicate the predicted mortality salience effect.

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). This sort of variation did not exceed variation expected by chance (e.g., sampling variance) regardless of exclusion rule: Exclusion Set 1: *Q*(16) = 19.35, *p* = .251; Exclusion Set 2: *Q*(16) = 21.72, *p* = .152; Exclusion Set 3: *Q*(16) = 18.57, *p* = .292. This result suggests that any observed differences in effect size between sites are likely due to random noise, as opposed to true differences in underlying effect size.

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

## 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* = 0.07, 95% CI = [-0.03, 0.17], *SE* = 0.05, *Z* = 1.33, *p* = .182. Exclusion Set 2: *Hedges’ g* = 0.11, 95% CI = [-0.02, 0.24], *SE* = 0.07, *Z* = 1.71, *p* = .087. Exclusion Set 3: *Hedges’ g* = 0.13, 95% CI = [-0.03, 0.29], *SE* = 0.08, *Z* = 1.57, *p* = .117.

Again, significant heterogeneity was not observed: Exclusion Set 1, *Q*(16) = 19.35, *p* = .251; Exclusion Set 2, *Q*(16) = 21.72, *p* = .152; Exclusion Set 3, *Q*(16) = 18.57, *p* = .292. This again suggests that any observed differences in effect size between sites are likely due to random noise, as opposed to true differences in effect size.

Critically, protocol version did not significantly predict replication effect size regardless of which exclusion criteria were used. Exclusion Set 1: *b* = 0.01, 95% CI = [-0.09, 0.11], *SE* = 0.05, *Z* = 0.20, *p* = .842; Exclusion Set 2: *b* = 0.05, 95% CI = [-0.08, 0.18], *SE* = 0.07, *Z* = 0.82, *p* = .410; Exclusion Set 3: *b* = 0.07, 95% CI = [-0.09, 0.23], *SE* = 0.08, *Z* = 0.86, *p* = .391. The Author Advised version did not produce significantly 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: *²* (1) = 0.67, *p* = 0.41; Author Advised labs: Exclusion Set 1: *²* (1) = 0.00, *p* = 1; Exclusion Set 2: *²* (1) = 0.00, *p* = 1; Exclusion Set 3: *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 (non-preregistered) Analyses

Results reported in this section were not pre-registered and should be considered exploratory.

**Researcher expectations and characteristics** A total of 23 researchers from 17 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, about one third 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* = 9.78, *SD* = 9.50). Five (22%) indicated they had “a lot” of TMT knowledge, eight (35%) indicated “some” knowledge, four (17%) indicated little knowledge, five (22%) indicated zero knowledge, and one (4%) did not respond to the question. One researcher indicated that they were an expert in TMT, but their site did not reach the minimum sample size specified by the preregistration.

When asked what outcome they wanted to happen, ten (43%) indicated that they hoped for the project to successfully replicate the TMT effect, nine (39%) indicated no preference, and two (9%) hoped the project would result in a failure to replicate, with two (9%) researchers leaving the question blank. On average, the teams estimated a 56% chance of successful replication with a wide range of estimates from 20% to 95% (*SD* = 23.15).[[5]](#footnote-35)

**Results for expert-only protocols.** To provide a test of the replicability and average effect size of TMT under ideal circumstances, one could focus only on the effect size within author-advised protocols. Effect sizes are descriptively larger among these sites but still not statistically significant. Exclusion rule 1: *Hedges’ g* = 0.08, 95% CI = [-0.07, 0.23], *SE* = 0.08, *Z* = 1.05, *p* = .292; Exclusion rule 2: *Hedges’ g* = 0.17, 95% CI = [-0.04, 0.37], *SE* = 0.10, *Z* = 1.59, *p* = .111; Exclusion rule 3: *Hedges’ g* = 0.19, 95% CI = [-0.10, 0.49], *SE* = 0.15, *Z* = 1.28, *p* = .200.

**Results for TMT-knowledgeable sites.** Five principal investigators indicated having “a lot” of knowledge about TMT. One might expect that these locations would have greater success at replicating the mortality salience effect. However, in all exclusion sets, when restricting analyses to these five sites, results were not statistically significant: Exclusion rule 1: *Hedges’ g* = 0.02, 95% CI = [-0.18, 0.23], *SE* = 0.10, *Z* = 0.20, *p* = .843; Exclusion rule 2: *Hedges’ g* = 0.03, 95% CI = [-0.18, 0.25], *SE* = 0.11, *Z* = 0.30, *p* = .766; Exclusion rule 3: *Hedges’ g* = -0.03, 95% CI = [-0.27, 0.22], *SE* = 0.13, *Z* = -0.20, *p* = .842.

**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. 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. For this reason, the anti-US essay from the original study was made more extreme in the Author Advised version of the replication.

This did successfully increase participant preferences for the pro-US author over the anti-US author among Author Advised replications as compared to In House replications, (2) = 76.82, < .001. Among In House replications, 48% of participants preferred the pro-US essay author, 42% preferred the anti-US essay author, and 10% had no preference. Among Author Advised replications, 68% of participants preferred the pro-US essay author, 22% preferred the anti-US essay author, and 10% had no preference.

Is the anticipated effect stronger among participants who preferred the pro-US author? We restricted the analysis to only participants at Author Advised sites who preferred the pro-US author. Under this restriction, there was still no significant difference between the mortality salience and control groups in their preference for the pro-US author over the anti-US author. Exclusion Set 1: *Hedges’ g* = 0.15, 95% CI = [-0.03, 0.33], *SE* = 0.09, *Z* = 1.60, *p* = .110; Exclusion Set 2: *Hedges’ g* = 0.13, 95% CI = [-0.12, 0.37], *SE* = 0.12, *Z* = 1.02, *p* = .306; Exclusion Set 3: *Hedges’ g* = 0.07, 95% CI = [-0.23, 0.37], *SE* = 0.15, *Z* = 0.46, *p* = .649.[[6]](#footnote-36)

1. We originally posted a pre-print accidentally omitting this exclusion: <https://psyarxiv.com/vef2c/>. In a critique, Chatard, Hirschberger, and Pyszczynski (2020) noted this as a deviation from the pre-registration. We revised the present manuscript with these comments in mind and ensured strict adherence to the pre-registration. [↑](#footnote-ref-22)
2. Whether these participants are included or excluded has little impact on the results. When these participants are included, some confidence intervals are somewhat narrower, and some results show effect sizes that are slightly smaller. The results presented without these exclusions may be viewed on the OSF page, and all data are available for reanalysis: <https://osf.io/8ccnw/> [↑](#footnote-ref-24)
3. Power calculated with the ‘pwr’ R package (Champely, 2020) assuming equal distribution of participants between mortality salience and control conditions, and two-tailed test with alpha of .05.. [↑](#footnote-ref-28)
4. We did not specifically pre-register whether tests would be one- or two-tailed, and some have argued that one-tailed tests would be more appropriate (e.g., Chatard, Hirschberger, & Pyszczynski, 2020). However, we had always intended to use two-tailed tests and this is consistent with all prior Many Labs projects (Klein et al., 2014; 2018, Ebersole et al., 2016; 2020) and the pre-registered sample analysis code (<https://osf.io/4xx6w>). [↑](#footnote-ref-29)
5. Including only sites that had not looked at any data, researchers still estimated a 56% chance of successful replication. [↑](#footnote-ref-35)
6. The random-effects meta-analysis failed to converge because the heterogeneity parameter, , was estimated as very, very small. To get the model to converge, we restricted to zero, creating a fixed-effects meta-analysis. [↑](#footnote-ref-36)