ML4 Results Section in rMarkdown

Richard A. Klein1

1 Université Grenoble Alpes

Author note

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

# Work in progress

Power calculations in this document are hardcoded, not calculated on the fly using RMarkdown, and can be inaccurate.

## Participants

Twenty-one data collection efforts from 18 unique sites collected data from a total of 2281 participants. In line with our preregistration, data from four labs were discarded for collecting fewer than 60 samples pre-exclusions. This excluded about 160 participants. Some In House sites collected data before the trial’s registration; that data was discarded from this confirmatory test, in line with the replication. This excluded a further 545 participants.

A total of 17 labs participated and provided a total sample of 1,578 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 1,550 participants (see Table 1 for a summary of sites). 841 participants (54.26%) reported being female and 442 participants (28.52%) 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 19.91 years (*SD* = 2.49). Participant reported race was 774 (49.94%) White, 219 (14.13%) Asian, 101 (6.52%) Black or African American, 3 (0.19%) American Indian or Alaska Native, 7 (0.45%) Native Hawaiian or Pacific Islander, 104 (6.71%) Other. The remaining participants did not report their race, or 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 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. 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. 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.

We then analyzed these individual results meta-analytically to get an aggregate effect size across all labs. 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). This gives us 1,550 participants. This sample size gives us 95% power to detect a condition effect of *d* = 0.18 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 1229. This sample size gives us 95% power to detect a 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). Further reduces the usable *N* to 1,076 participants. This sample size gives us 95% power to detect a condition effect of *d* = 0.22.

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 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, ~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* = 9.78, *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.

When asked what outcome they wanted to happen, 10 (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).[[1]](#footnote-28)

## 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.

## 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).

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*() = , *p* = ; Exclusion Set 2, *Q*() = , *p* = ; Exclusion Set 3: *Q*() = , *p* = .

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 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*() = , *p* = ; Exclusion Set 2, *Q*() = , *p* = ; Exclusion Set 3, *Q*() = , *p* = .

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 Analyses

**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. [NOTE: THESE ARE TWO-TAILED P-VALUES AND 95% CIs. I’M NOT SURE IF METASEM CAN EASILY PROVIDE 90% CIS AND ONE-TAILED TESTS.] 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.** 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* = 1.03, *SD* = 2.28) and the control group (*M* = 0.93, *SD* = 2.31), *t*(477.87) = 0.49, *p* = 0.62, *Hedges’ g* = 0.04, 95% CI = [-0.13, 0.22].[[2]](#footnote-35)

**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, 48% of participants prefered the pro-US essay author, 42% preferred the anti-US essay author, and 10% 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-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* = 1.26, *SD* = 2.04), control group (*M* = 1.10, *SD* = 1.80), *t*(697.00) = 1.12, *p* = 0.26, *Hedges’ g* = 0.08, 95% CI = [-0.06, 0.23]; Exclusion Set 2: mortality salience group (*M* = 1.60, *SD* = 2.14), control group (*M* = 1.28, *SD* = 1.95), *t*(370.43) = 1.50, *p* = 0.13, *Hedges’ g* = 0.15, 95% CI = [-0.05, 0.36]; Exclusion Set 3: mortality salience group (*M* = 2.08, *SD* = 2.11), control group (*M* = 1.66, *SD* = 2.02), *t*(216.09) = 1.51, *p* = 0.13, *Hedges’ g* = 0.20, 95% CI = [-0.06, 0.46]. 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.

1. Including only sites that had not looked at any data, researchers estimated a 56% chance of successful replication. [↑](#footnote-ref-28)
2. 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. [↑](#footnote-ref-35)