A Bayesian Multiverse Analysis of Many Labs 4: Quantifying the Evidence against Mortality Salience

Julia M. Haaf
University of Amsterdam

Suzanne Hoogeveen
University of Amsterdam

Sophie Berkhout
University of Amsterdam

Quentin F. Gronau
University of Amsterdam

Eric-Jan Wagenmakers
University of Amsterdam

Abstract

Many Labs projects have become the gold standard for assessing the replicability of key findings in psychological science. The Many Labs 4 project recently failed to replicate the mortality salience effect where being reminded of one's own death strengthens the own cultural identity. Here, we provide a Bayesian reanalysis of Many Labs 4 using meta-analytic and hierarchical modeling approaches and model comparison with Bayes factors. In a multiverse analysis we assess the robustness of the results with varying data inclusion criteria and prior settings. Bayesian model comparison results largely converge to a common conclusion: We find evidence against a mortality salience effect across the majority of our analyses. Even when ignoring the Bayesian model comparison results we estimate overall effect sizes so small (between $d = 0.03$ and $d = 0.18$) that it renders the entire field of mortality salience studies as uninformative.

Keywords: Bayes factor, Bayesian meta-analysis, Replication

Introduction

Many Labs is a crowd-sourcing project that collects data from many different sites across the globe to answer questions about replicability and variability of effects, and it has become the gold standard for assessing the robustness of key findings in the psychological literature. Many Labs 4 (Klein et al., 2019), the most recent implementation of this idea, is a large scale attempt to replicate the Mortality Salience Effect from Terror Management Theory (Greenberg, Pyszczynski, Solomon, Simon, &

This research was supported in part by a Vici grant (016.Vici.170.083) to EJW and by a Netherlands Organisation for Scientific Research (NWO) grant to QFG (406.16.528).
Analysis code is provided at https://github.com/jstbcs/ml4-reanalysis. JMH, SH, and EJW planned the analyses; JMH, SB, and SH conducted the data analysis; all authors contributed to the writing of the manuscript.
Correspondence should be sent to Julia M. Haaf, University of Amsterdam, Nieuwe Achtergracht 129 B, 1018 WT Amsterdam, The Netherlands. E-mail may be sent to j.m.haaf@uva.nl.

Breus, 1994): reminders of one's own death strengthen one's cultural identity. In the classical demonstration of this effect, participants from the United States who were prompted to imagine their own death expressed more pro-American (i.e., in line with their worldview) beliefs than participants who were prompted to imagine watching TV. In addition to the question of replicability, Klein et al. (2019) wanted to assess the impact of involving the original authors in the study design. Therefore, some studies followed a standard protocol that was agreed upon by experts in the field (Author-Advised) while other studies were designed by the labs conducting them (In-House). After data collection from over 2,000 participants in 21 labs with and without involvement of the original authors the project could not replicate the original finding of Study 1 of Greenberg et al. (1994), and reported an overall effect size of $g = 0.03$, $95\%\text{CI} = [-0.06, 0.12]$.

Soon after the preprint of the Many Labs 4 project was posted, a critique of the analysis emerged. Chatard, Hirschberger, and Pyszczynski (2020) pointed out that Klein et al. (2019) did not follow their own preregistered analysis. Chatard et al. (2020) argued that the preregistration specified a minimum of 40 participants per experimental cell as the threshold for sufficient power of any individual study, and therefore determined a total of 80 participants as target sample size for each lab. When reanalyzing the data from the Many Labs 4 project only including studies with 40 participants per condition Chatard et al. (2020) found a significant effect in line with the original results.

Include or Exclude?

Which of these analyses is the correct one? Based on theoretical arguments and (interpretations of) the preregistered plan, there may be several valid answers to this question, and several levels of exclusion criteria that ought to be considered. Both Klein et al. (2019) and Chatard et al. (2020) agreed on three *participant-level* exclusion criteria (the last two are suggested by the original authors – Greenberg, Pyszczynski, and Solomon –, who were consulted by the Many Labs 4 team):

1. Exclude participants who did not respond to all prompts of the dependent variable (leaving $N = 2211$).

2. In addition to exclusion criterion 1, participants who did not self-identify as

white and/or who reported not to be born in the United States were also excluded (leaving $N = 637$).[1]

3. In addition to exclusion criteria 1 and 2, participants who responded below 7 on the 9-point American Identity item were also excluded (leaving $N = 277$).

In addition to these three participant-level exclusion criteria, power considerations motivated three different study-level exclusion criteria. We refer to these exclusion criteria as *N-based* criteria.

1. Include data from all labs (leaving $K = 21$ studies).

2. Exclude labs with fewer than 60 participants (leaving $K = 17$ studies).

3. Exclude labs with fewer than 40 participants per condition (leaving $K = 13$ studies).

Note that N-based exclusion criterion 2 was preregistered by Klein et al. (2019): "Samples will be included as long as they collect at least 60 participants by the time data collection ends" (see preregistration document, osf.io/4xx6w). In contrast, Chatard et al. (2020) derive exclusion criterion 3 from the *target* sample size specified in the preregistration document, although it is never mentioned as a criterion for exclusion. We decided to add both exclusion criteria for the sake of comparison.

Lastly, Greenberg et al. (1994) suggested that the effect may only emerge in Author-Advised studies as the mortality salience effect is highly sensitive to nuances in the study implementation. Therefore, the following distinction may constitute an additional set of *study-level* exclusion criteria. We refer to these exclusion criteria as *Protocol* criteria.

1. Include all studies (leaving $K = 21$).

2. Exclude all In-House studies (leaving $K = 9$).

[1] The argument is that the effect may only be present for participants who strongly identify with pro-American worldviews. We included participants who did identify as white in addition to another ethnicity, i.e., who are multiracial. We consider this the most appropriate interpretation of the preregistered ethnicity criterion.

These three levels of exclusion result in $3 \times 3 \times 2 = 18$ constellations of exclusion criteria. Table 1 shows all constellations, the resulting number of studies and total number of included participants. In the preprint, Klein et al. (2019) based their main conclusions on three of these constellations (blue rows): Including all studies, but varying the participant-level exclusion criteria.[2] Similarly, even though Chatard et al. (2020) conducted a variety of analyses in their comment, they based their key conclusions on three different constellations of criteria (pink rows): Excluding studies with fewer than 40 participants per condition, excluding In-House studies, with varying participant-level exclusion criteria.

In the following we will first report a reanalysis for the three exclusion constellations from Klein et al. (2019), and then for the three exclusion constellations from Chatard et al. (2020). Subsequently, lacking compelling argumentation for or against any of the criteria, we decided to conduct an analysis based on the entire set of 18 constellations as a multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). Note that some of the participant-level and study-level criteria are completely overlapping (e.g., only Author-Advised labs recorded American Identity, hence all In-House labs are excluded for the third participant-level exclusion set). As a result, there are 11 instead of 18 unique constellations (only rows that are not grey in Table 1).

A Bayesian Multiverse Reanalysis

We opt for a Bayesian analysis using Bayes factor model comparison (Jeffreys, 1939; Kass & Raftery, 1995). In short, Bayes factors quantify the relative evidence for a model (e.g., the alternative) over another model (e.g., the null). For an introduction to Bayes factor model comparison we refer the reader to Wagenmakers et al. (2018) and Rouder, Haaf, and Aust (2018).

The main advantage of Bayesian statistics in light of the current debate around the Many Labs 4 results is that it allows us to distinguish between evidence for the absence of the mortality salience effect and the absence of evidence for or against

[2] We note that the eventual published article of Many Labs 4 may adopt different study-level criteria in order to adhere to the preregistration (i.e., exclude labs where $N < 60$). Furthermore, close examination of the preregistration document also revealed that some In-House labs had already started data collection prior to the registration and were therefore solely to "be included in clearly labelled supplemental and exploratory analyses". However, as these data were not accessed by the lead researchers and concerned In-House studies that were free to design their own protocols, we see no reason to exclude these observations.

Table 1
Exclusion constellations and resulting sample sizes

Participant-level	N-based	Protocol	Sample Size	Number of Studies
All	All	All	2,211	21
White & US-born	All	All	637	12
US-Identity > 7	All	All	277	9
All	All	AA	799	9
White & US-born	All	AA	463	9
US-Identity > 7	All	AA	277	9
All	N > 60	All	2,053	17
White & US-born	N > 60	All	549	9
US-Identity > 7	N > 60	All	229	7
All	N > 60	AA	700	7
White & US-born	N > 60	AA	386	7
US-Identity > 7	N > 60	AA	229	7
All	N > 80	All	1,852	14
White & US-born	N > 80	All	549	9
US-Identity > 7	N > 80	All	229	7
All	N > 80	AA	700	7
White & US-born	N > 80	AA	386	7
US-Identity > 7	N > 80	AA	229	7

Note. Blue rows refer to Klein et al.'s key analyses; pink rows refer to Chatard et al.'s key analyses; grey rows are repeated data sets and not included in the multiverse analysis; AA = Author-Advised.

the effect. We decided to conduct two alternative analyses: Bayesian model-averaged meta-analysis (Gronau et al., 2017), and Bayesian hierarchical modeling (Haaf & Rouder, 2017; Rouder, Haaf, Davis-Stober, & Hilgard, 2019). The key distinction between these two approaches is that they operate on two levels of the data: For meta-analysis, the data from each lab are summarized with an effect size estimate and standard error, and these statistics are then analyzed using a linear model. For hierarchical modeling, the linear model is extended to the participant level, and participants' data are analyzed as nested within labs. Despite these differences, both analyses should provide comparable results. Subsequently, we briefly outline the two modeling approaches.

Methods

Bayesian Model-averaged Meta-analysis

Both classical and Bayesian meta-analysis typically consider four different models: (1) fixed-effect null model, (2) fixed-effect alternative model, (3) random-effects null model, and (4) random-effects alternative model. In Bayesian model comparison, we may now compute Bayes factors to compare any two of these models. Bayesian model averaging (e.g., Hinne, Gronau, van den Bergh, & Wagenmakers, in press) allows for broader inference when considering several models simultaneously. Using model averaging one can calculate the evidence for the presence of an effect while taking into account uncertainty with respect to choosing a specific model. For the application here, this logic implies that we can assess evidence for the mortality salience effect without committing to the fixed-effect or random-effects models.

Specifically, the model-averaged Bayes factor in favor of the presence of an effect is obtained by comparing the models that allow for the presence of an effect (i.e., (2) and (4) above) to the models that state the effect is absent (i.e., (1) and (3) above). In a similar fashion one can calculate the model-averaged Bayes factor in favor of the presence of between-study heterogeneity by comparing the models that allow for the presence of between-study heterogeneity (i.e., (3) and (4) above) to the models that state between-study heterogeneity is absent (i.e., (1) and (2) above).

We follow Gronau, Heck, Berkhout, Haaf, and Wagenmakers (in preparation) for the specification of our Bayesian model-averaged meta-analysis. To conduct such an analysis, one needs to specify priors for the overall effect size across labs and the between-study standard deviation. For the between-study standard deviation we follow Gronau et al. (2017) and use an Inverse-Gamma(1, 0.15) prior. This prior is based on the empirical assessment of effect sizes from meta-analyses reported in *Psychological Bulletin* in the years 1990–2013 (van Erp, Verhagen, Grasman, & Wagenmakers, 2017). Van Erp et al. (2017) gathered all non-zero between-study standard deviation estimates for meta-analyses on standardized mean differences (e.g. Cohen's d), and the histogram approximately followed this distribution. For the overall effect size, we considered three different prior settings: (1) a zero-centered Cauchy distribution with scale $1\sqrt{2} \approx 0.707$ (*default* prior, Morey & Rouder, 2018), (2) a t-distribution with location 0.35, scale 0.102, and 3 degrees of freedom (*Oosterwijk* prior[3]), and (3)

[3]This *Oosterwijk* prior has been elicited for a reanalysis of a social psychology study (Gronau,

a normal distribution with mean 0.3 and standard deviation 0.15 (*Vohs* prior[4]). In line with the mortality salience hypothesis, all prior distributions on the overall effect size were truncated below at zero to allow only effect sizes in the expected direction. Readers interested in Bayesian model-averaging in meta-analysis may consult Gronau et al. (2017), Scheibehenne, Gronau, Jamil, and Wagenmakers (2017), and Landy et al. (in press).

Bayesian Hierarchical Modeling

For Bayesian hierarchical modeling we take advantage of the open availability of all collected data from the Many Labs 4 project. The dependent variable is the same across all studies, and participants are nested in studies resulting in a hierarchical data structure. We used the development by Rouder et al. (2019) with models similar to the ones used for the embodied cognition reanalysis by Rouder et al. (2019). There are four models under consideration: (1) The null model corresponds to the notion that none of the studies show an effect; this model is similar to the fixed-effect null model from the model-averaged meta-analysis. (2) The common-effect model corresponds to the notion that all studies show the same effect in the expected direction; this model is similar to the fixed-effect alternative model from the model-averaged meta-analysis. (3) The positive-effects model corresponds to the notion that all studies show an effect in the expected direction; and (4) the unconstrained model refers to the notion that the overall effect and study effects may vary freely; this model is similar to the random-effects alternative model from the model-averaged meta-analysis. We compute Bayes factors for models (2), (3), and (4) against model (1), the null model.

There are two critical prior settings to consider, the scale setting on the overall effect (μ_θ in Rouder et al., 2019) and the scale setting on the between-lab heterogeneity (σ^2_θ in Rouder et al., 2019). The scale on the overall effect corresponds to the expected size of the overall effect. As Rouder et al. (2019), we set this scale to 0.4 since we expect a small-to-medium effect size. The scale of the between-lab variance corresponds to the expected amount of variability in effect size across studies. Again, we kept the value of 0.24 as proposed by Rouder et al. (2019).

Ly, & Wagenmakers, in press), but we believe it is a reasonable prior for many psychological studies more generally.

[4]This *Vohs* prior has been specified by ego depletion experts to analyze ego depletion replication studies (Vohs et al., under review).

Preregistration and Approach

With these two approaches we are now ready to reanalyse the Many Labs 4 data. Subsequently, we report the results of the Bayesian reanalysis of the key findings reported by Klein et al. (2019), and the results of the Bayesian reanalysis of the key findings by Chatard et al. (2020). Finally, we provide the results of the multiverse analysis across all possible exclusion criteria.

The analyses, including prior settings, were preregistered on the Open Science Framework (osf.io/ae4wx, see also Appendix I). However, we decided to deviate from the preregistration by including more constellations of exclusion criteria. Specifically, we originally planned to only use participant-level exclusion criterion 1 and later decided to include all of them. We believe the changes help to provide a more complete analysis.

The Bayesian model-averaged meta-analyses are conducted using the R-package `metaBMA` (Heck & Gronau, 2017). The Bayesian hierarchical modeling is conducted using the R-package `BayesFactor` (Morey & Rouder, 2018). All R-code is provided at github.com/jstbcs/ml4-reanalysis.

Bayesian Reanalysis of Klein et al.'s Key Findings

Model-averaged Meta-analysis of Klein et al.

Figure 1A, shows the observed effect size estimates for the first participant-level exclusion criteria without applying any study-level exclusion criteria. The observed effect sizes from each study (grey points) are plotted in increasing order, and the grey bars show the 95% confidence intervals for the effect size estimates. A quick first assessment of Figure 1A shows that the confidence intervals of observed effect sizes from 18 of the 21 studies cover zero. The black points in the figure refers to estimated effect sizes from a meta-analytic random-effects alternative model with a two-sided default priors. This model takes the observed study-level variability of effect sizes into account, and therefore estimates less variability of true study effects than the observed effect sizes. For the individual studies, the credible intervals of all estimated effect sizes for all three analyses cover zero.

In order to estimate the overall effect size across studies (Hedges' g) we used the same model as was used to estimate the individual-study effects (i.e., a random-effects alternative model with the default prior). For the full sample (participant-level ex-

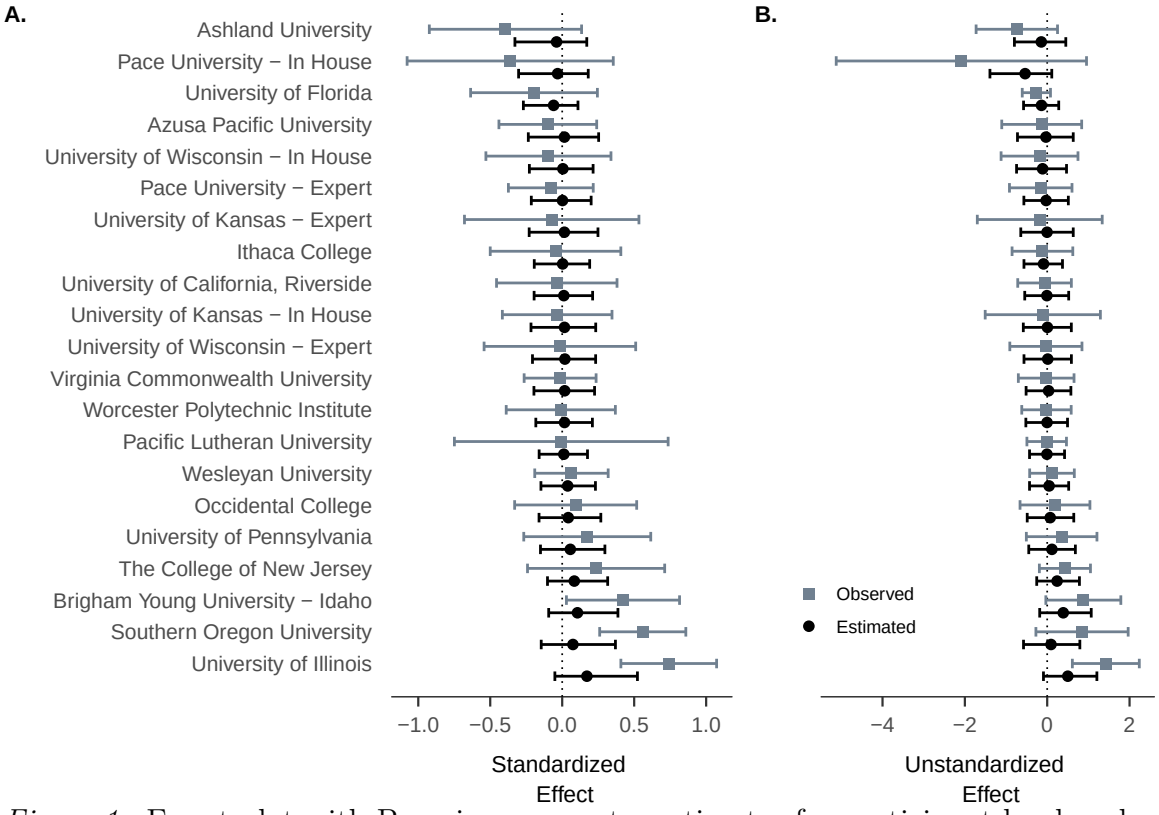

Figure 1. Forest plot with Bayesian parameter estimates for participant-level exclusion set 1 and no further study-level exclusions. A. Bayesian meta-analysis (with two-sided default prior). The grey points represent calculated effect sizes with 95% confidence intervals, the black points represent estimated effect sizes from the random-effects alternative model with 95% credible intervals. B. Bayesian hierarchical analysis. The grey points represent unstandardized observed effects for each study with 95% confidence intervals. The black points represent estimated unstandardized effects from the unconstrained model with 95% credible intervals.

clusion criterion 1) the overall effect size is estimated as 0.03, 95%CI = $[-0.07, 0.13]$; for participant-level exclusion criterion 2 the overall effect size is estimated as 0.03, 95%CI = $[-0.14, 0.21]$; and for participant-level exclusion criterion 3 the overall effect size is estimated as 0.07, 95%CI = $[-0.21, 0.33]$. The most consistent pattern is that the credible interval widens when the exclusion criterion becomes more restrictive. Overall, these estimates are more consistent with the absence of an effect rather than its presence.

To quantify the absence or presence of an effect we now turn to Bayes factor model comparison. The Bayes factors for the key analyses from Klein et al. (2019)

Table 2
Model-averaged Bayes factors for key analyses.

Inclusion Criteria				Effect BF_{01}			Heterogeneity BF_{01}
Participant-level	N-based	Proto-col	Labs	Default	Ooster-wijk	Vohs	Default
Klein et al. (2019)							
All	All	All	21	12.60	44.69	16.64	2.28
White & US-born	All	All	12	7.95	16.84	7.20	2.42
US-identity > 7	All	All	9	4.18	4.01	2.49	1.79
Chatard et al. (2020)							
All	N > 40	AA	7	3.82	5.84	2.75	2.54
White & US-born	N > 40	AA	7	1.42	0.90	0.66	2.08
US-identity > 7	N > 40	AA	7	1.45	0.73	0.62	1.89

Note. All Bayes factors are reported in favor of the null model. AA = Author-Advised.

are shown in the top three rows of Table 2. Note that not all studies are included for exclusion criterion 2 because data on ethnicity and country of birth were only collected for some of the labs. Likewise, the American identity was only assessed in the Author-Advised studies, and therefore exclusion criterion 3 leads to the inclusion of only nine studies. The first three Bayes factors in each row are model-averaged Bayes factors referring to evidence against an overall effect. All analyses across participant-level exclusions and prior choices provide evidence against an overall effect, with Bayes factors ranging from 44.69-to-1 to 2.49-to-1 in favor of the null model. Note that the Oosterwijk prior is the most optimistic prior with the least probability density close to zero. Therefore, the Bayes factors are somewhat larger for this prior—the optimistic predictions that follow from the Oosterwijk prior are least consistent with what the data show, which are effect sizes close to zero. The last Bayes factor in each row indicates evidence against heterogeneity of study effects averaged across models with and without an overall effect. These Bayes factors reflect that there is some evidence against study heterogeneity. In sum, the pattern of Bayes factors indicates evidence against an overall mortality salience effect across the three prior settings and the three data sets. These results are in line with the estimation results in Figure 1A, and with the overall effect size estimates from a two-sided model.

Hierarchical analysis of Klein et al.

Figure 1B shows the observed, unstandardized effects and the estimates from the unconstrained multilevel model for the first participant-level exclusion criteria. As can be seen, there is considerable hierarchical shrinkage reducing the variability of estimated effects as compared to observed effects. Effect size estimates from the unconstrained model (similar to Cohen's d) are 0.01, 95%CI = $[-0.11, 0.12]$ for participant-level exclusion criterion 1, 0.02, 95%CI = $[-0.17, 0.21]$ for exclusion criterion 2, and 0.05, 95%CI = $[-0.22, 0.32]$ for exclusion criterion 3. Note that posterior means are close to zero, and that all credible intervals cover zero. The estimates are therefore consistent with the absence of an overall effect.

Bayes factors are shown in the first three rows of Table 3. BF_{0f} refers to the Bayes factor between the null model and the unconstrained model; BF_{01} refers to the Bayes factor between the null model and the common-effect model where the overall effect is positive and there is no variability between study effects; and BF_{0+} refers to the Bayes factor between the null model and the positive-effects model where study effects may vary but all are consistently positive. All Bayes factors are in comparison to the preferred model, the null model, indicating evidence that none of the studies show an effect. The second best model is the common-effect model where all studies have the same, positive effect, and the Bayes factor between the null model and the common-effect model is between 10.34-to-1 to 2.11-to-1 in favor of the null model depending on the different participant-level exclusion criteria. In sum, this pattern indicates evidence against an overall mortality salience effect (null model), and even if there was an effect (common-effect model) there is no evidence for variability of study effects. These results are consistent across the three data sets, and they are in line with the estimation results shown in Figure 1B.

Summary of the Reanalysis for Klein et al.

Across both analyses, the meta-analytic approach using Bayesian model-averaging, and the hierarchical approach using participant-level data, we find no evidence for the mortality salience effect. The results are consistent across participant-level exclusion criteria and prior settings. Even though the evidence against an effect is more pronounced when all participants are included in the analysis, this pattern is easily explained by the resolution of the analysis with increasing numbers of observa-

Table 3
Bayes factors for key analyses.

Inclusion Criteria						
Participant-level	N-based	Protocol	Sample Size	BF_{0f}	BF_{01}	BF_{0+}
Klein et al. (2019)						
All	All	All	2,211	33.26	10.34	8,787.94
White & US-born	All	All	637	19.65	5.67	123.12
US-Identity > 7	All	All	277	9.24	2.11	13.43
Chatard et al. (2020)						
All	N > 40	AA	700	13.64	2.07	11.96
White & US-born	N > 40	AA	386	6.87	0.94	2.73
US-identity > 7	N > 40	AA	229	4.83	0.83	1.57

Note. All Bayes factors are reported in favor of the null model. AA = Author-Advised.

tions: The fewer observations, the less evidence in any direction, and the wider the estimated posterior distribution of the overall effect.

Bayesian Reanalysis of Chatard et al.'s Key Findings

Model-averaged Meta-analysis of Chatard et al.

For the reanalysis of the key findings of Chatard et al. (2020) we provide a forest plot of the most exclusive criteria—participant criterion 3, and only author-advised studies with more than 40 participants per cell inlcuded—in Figure 2. Together with Figure 1 Figure 2 illustrates the range of included study effects from the most liberal to the most restrictive criteria. Figure 2A again shows the results from a meta-analytic random effects model with unconstrained overall effect. Note that all confidence intervals (grey bars) and all credible intervals (black bars) include zero.

We estimated the overall effect size across studies (Hedges' g) using the settings from the default prior without constraining the direction of the overall effect. We did so for all data sets using the three participant-level exclusion criteria, only studies that had more than 40 participants per cell collected, and only Author-Advised studies. For participant-level exclusion criterion 1 the overall effect size is estimated as 0.08, 95%CI = $[-0.09, 0.25]$; for exclusion criterion 2 the overall effect size is estimated as 0.16, 95%CI = $[-0.07, 0.40]$; and for exclusion criterion 3 the overall effect size is estimated as 0.18, 95%CI = $[-0.10, 0.47]$. While the point estimates

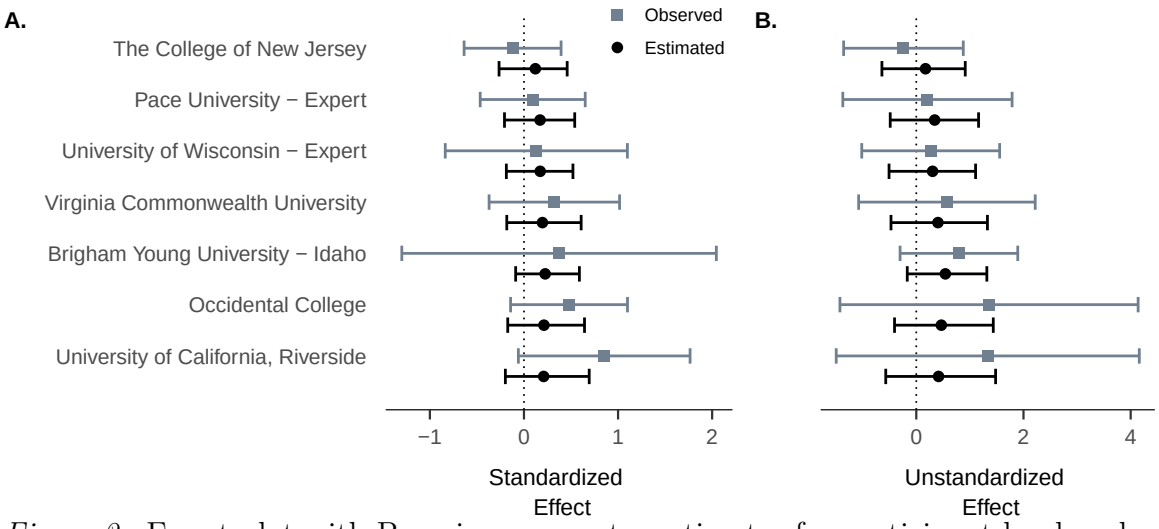

Figure 2. Forest plot with Bayesian parameter estimates for participant-level exclusion set 3 for studies with more than 40 particpants per cell and only author-advised studies included. A. Bayesian meta-analysis (with two-sided default prior). The grey points represent calculated effect sizes with 95% confidence intervals, the black points represent estimated effect sizes from the random-effects alternative model with 95% credible intervals. B. Bayesian hierarchical analysis. The grey points represent unstandardized observed effects for each study with 95% confidence intervals. The black points represent estimated unstandardized effects from the unconstrained model with 95% credible intervals.

are considerably larger than the ones when all studies are included, the posterior distributions and therefore also the credible intervals are considerably wider due to much smaller sample sizes. In this analysis, only seven studies were included, and only between 700 and 229 participants.

To quantify the absence or presence of an effect we again computed model-averaged Bayes factors. These are shown in the bottom three rows of Table 2. The first three Bayes factors in each row are model-averaged Bayes factors referring to evidence against an overall effect using different prior distributions. Here, the pattern is a bit more inconsistent than in the Klein et al. reanalysis, and the outcome depends on a combination of the prior settings and exclusion criteria: Bayes factors (weakly) favor the absence of an effect over its presence for all priors if participant-level exclusion criterion 1 is applied. For the smaller data sets using criteria 2 or 3, the Bayes factors are essentially inconclusive - for the default prior the Bayes factors are still in favor of the null hypothesis but close to 1. For the other two prior setting the Bayes factors are

in favor of the presence of an effect but, again, close to 1. The largest Bayes factor in favor of the presence of an effect is with the Vohs prior, and participant-level exclusion setting 3.

The last column in Table 2 shows the model-averaged Bayes factor quantifying evidence against heterogeneity of effect sizes across labs. Again, there is weak evidence against heterogeneity. In sum, this pattern is in line with the absence of evidence for or against an overall mortality salience effect.

Hierarchical Analysis of Chatard et al.

We also reanalyzed Chatard et al.'s findings with a hierarchical modeling approach. Figure 2B shows study estimates from the unconstrained model for the unstandardized effects. As with the standardized effects in panel A, all confidence intervals and credible intervals cover zero.

Effect size estimates from the unconstrained model (similar to Cohen's d) of 0.08, 95%CI = $[-0.12, 0.29]$ for participant-level exclusion criterion 1, 0.14, 95%CI = $[-0.11, 0.37]$ for participant-level exclusion criterion 2 and 0.18, 95%CI = $[-0.11, 0.48]$. Note that all credible intervals include zero, and even though the posterior mean increases with more conservative exclusion criteria the width credible interval increases as well implying increasing uncertainty about the effect size. The posterior distribution of the overall effect size is therefore again consistent with the absence of an overall effect.

The pattern of Bayes factors is somewhat less consistent than the estimation results across exclusions. Bayes factors are shown in the last three rows of Table 3. The pattern of Bayes factors is, as with the model-averaged analysis, dependent on the participant-level exclusion criterion. Under participant-level exclusion criterion 1 the preferred model is the null model, and it is weakly preferred over the second-best model, the common effect model, by a Bayes factor of $\text{BF}_{01} = 2.07$. For the other two exclusion criteria, the common-effect is preferred over the null model but the Bayes factors are even weaker (1.06 and 1.24 in over the null model). In sum, the pattern for the different data exclusions is in line with the conclusions from the model-averaged analysis: The Bayes factors show the absence of any consistent evidence for or against an effect.

Summary of the Reanalysis of Chatard et al.

Both the model-averaged meta-analysis and the hierarchical modeling approach show a similar pattern: Across the three participant-level exclusion criteria and different prior settings, there is only weak and inconsistent evidence for or against an overall mortality salience effect. Here, we advice readers not to overly interpret whether the Bayes factor is 1.5-to-1 for or against the overall effect—none of these Bayes factors are convincing. Instead, all of the analyses in this section point to the conclusion that more data are needed. The exclusion criteria applied here thinned out the data so much – in the final analytic data set only 10% of the initial data is retained – so that no firm conclusion is possible anymore.

Bayesian Multiverse Analysis Across All Exclusion Criteria

To assess the robustness of the previously reported results we conducted a multiverse analysis using the eleven data sets from Table 1 (i.e., all rows that are not grey). We conducted a model-averaged meta-analysis and report here the Bayes factors for the presence of an effect against its absence. The analysis is conducted using the three different prior distributions, the default prior, the Oosterwijk prior, and the Vohs prior. The Bayes factors are plotted in Figure 3 (y-axis). Bayes factors in favor of the mortality salience effect are above the horizontal line, and Bayes factors against the mortality salience effect are below the horizontal line. The x-axis refers to the number of participants whose data are included in the analysis. The size of the point reflects the number of studies included in the analysis. The majority of Bayes factors are in line with the absence of the mortality salience effect. Because the Bayes factor depends on the sample size, more evidence against morality salience comes from analyses that are based on more data (i.e., larger number of included participants and studies). Only two constellations of exclusion criteria provide evidence for the mortality salience effect.

To inspect the effect of prior settings one can view the points in Figure 3 that are in the same x-axis location. Remember that the default prior is the most vague prior and the Oosterwijk prior is more optimistic than the Vohs prior. For the three data sets with the largest numbers of participants Bayes factors are larger for more optimistic priors because evidence against optimistic and informed models accumulates faster when comparing to a null model. The same logic applies for situations where data are more ambiguous. The smallest data sets show a small positive overall

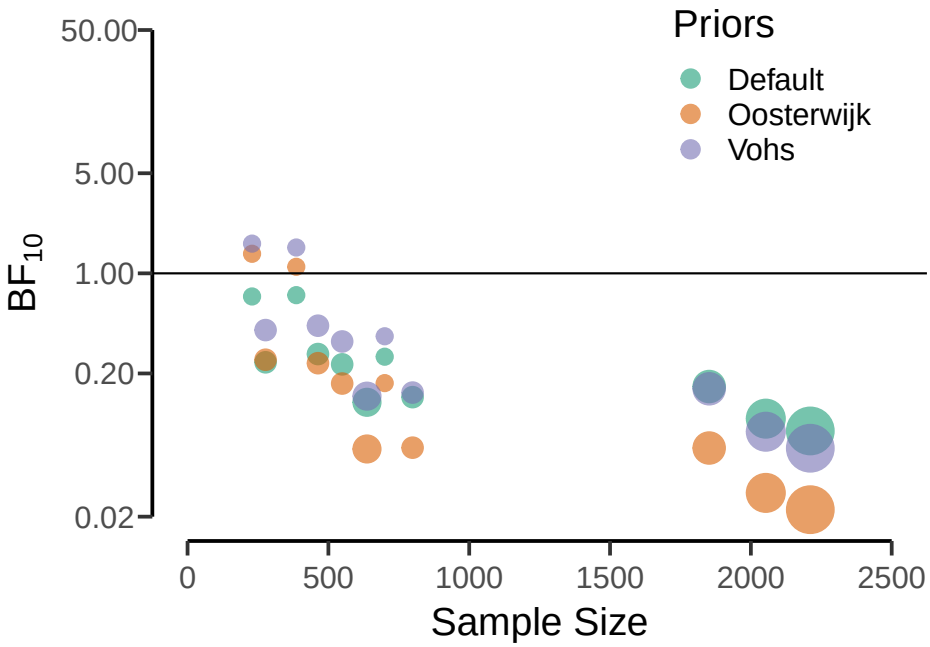

Figure 3. Results from the multiverse analysis: Bayes factors in favor of a mortality salience effect are above the horizontal line, Bayes factors against the mortality salience effect are below the horizontal line. The color of the points refers to the different priors on the overall effect, the size of the points refers to the number of studies included in the analysis, and the x-axis refers to the number of participants the analysis is based on. The majority of analyses provide evidence against the mortality salience effect.

effect, and evidence for this small effect accumulates faster with more optimistic priors than less optimistic ones. Therefore, the Bayes factors are only greater than one (i.e., in favor of an effect) for the Vohs and the Oosterwijk prior. Because the Vohs prior has more density for smaller effect sizes than the Oosterwijk prior, the Bayes factor favors an overall effect most for the Vohs prior.

Summary of the Multiverse Analysis

The evidence against the morality salience effect appears fairly robust against choices of exclusion criteria and priors. When conducting a large number of analyses on the same data some of these analyses will almost inevitably lead to some evidence in the opposite direction than the overall results. This is especially the case when the data provide relatively weak evidence (Bayes factors less than 5-to-1 against an effect). Bayes factors close to 1 may signal a lack of resolution of the data and therefore

the absence of evidence for or against an effect. When the number of participants is high and many studies are included there is convincing evidence against the mortality salience effect. The four Bayes factors that are weakly in favor of the mortality salience effect are based on two of the smallest data sets and the two more informative prior settings.

Conclusion

We conducted a Bayesian reanalysis of the Many Labs 4 project with varying exclusion criteria, priors, and model choices. In a Bayesian multiverse analysis we calculated a total of 33 model-averaged Bayes factors based on three different prior settings and 11 different data sets resulting from different data exclusion criteria derived from the Many Labs 4 preregistration (Klein et al., 2019). 29 of the 33 Bayes factors provide evidence against an overall mortality salience effect, ranging between 1.42-to-1 and 44.69-to-1 in favor of the absence of an effect. The remaining four Bayes factors provide only weak evidence for the presence of such an effect, ranging between 1.11-to-1 and 1.61-to-1 in favor of the presence of an effect. Additionally, we do not find evidence for heterogeneity of effects across studies. Even if we do not believe the evidence across 33 Bayesian model comparisons and assume there is an effect, this effect is so small (between $d = 0.03$ and $d = 0.18$) that it renders the entire field of mortality salience studies as uninformative: Most of the studies conducted in the past would have been vastly underpowered, and would require a very specific subgroup of participants.

Our analyses revealed that the evidence is relatively consistent across different exclusion criteria. For the current analysis, we assumed that all exclusion criteria are equally plausible. With this assumption we implicitly assigned an equal weight to all analyses. However, we admit that this may not be the case. Chatard et al. (2020) argue that their chosen criteria are superior when considering theoretical arguments and study planning. With their analysis, they implicitly introduced a weighing where all other exclusion options received a weight of zero. Readers can choose these weights themselves when they consider how to interpret the results reported here.

There are additional issues with selectively subsetting and reanalyzing data sets. A key danger is that for some subsets one always finds results opposite of the conclusions from the analysis of the full data set. On the study level, researchers should therefore first ensure that there is evidence for variability of studies that warrants such

subsetting. In the current analysis, we found evidence against study heterogeneity. When interpreting the results we therefore recommend to rely mainly on the estimates from the full data set. Additionally, subsetting the data inevitably reduces the resolution to detect an effect. The critics of the Many Labs 4 project (Chatard et al., 2020) based their main conclusions on analyses with smaller sample sizes. Ironically, while Chatard et al. (2020) argued that sample size should be considered when including studies their exclusion criteria actually reduced the power of the meta-analysis. To tackle this issue—and if there was evidence for study heterogeneity—one could include some of the subsetting criteria as predictor in the meta-analytic model (e.g. author-advised vs. in-house).

In summary, the multiverse analysis conducted here shows a certain convergence of results. Even though the degree of evidence varies, models with no effect of mortality salience are mostly preferred over models with an effect of mortality salience. This result highlights the robustness against choices of priors and exclusion criteria. The Bayesian multiverse approach provides rich results that go much beyond the original analyses by the Many Labs 4 team. Moreover, multiverse analyses can be executed easily, for example using JASP (JASP Team, 2019).[5] The current analyses were conducted in R, and the code is provided at github.com/jstbcs/ml4-reanalysis. The ease and informativeness of multiverse analyses show that this approach should be more generally used to analyze large-scale studies. The Many Labs idea is that the robustness of empirical phenomena becomes clear when data are collected across several labs. Similarly, the robustness of statistical conclusions becomes clear when data are analyzed using several thoughtfully selected models. A complete assessment of robustness and uncertainty therefore requires both many labs and many models.

[5]The Bayesian meta-analysis module is currently still in development. To still use it one can install the nightly build version of JASP.

References

Chatard, A., Hirschberger, G., & Pyszczynski, T. (2020). A word of caution about many labs 4: If you fail to follow your preregistered plan, you may fail to find a real effect.

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.

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (in preparation). A primer on bayesian model-averaged meta-analysis. *Manuscript in preparation*.

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (in press). Informed Bayesian *t*-tests. *The American Statistician*.

Gronau, Q. F., van Erp, S., Heck, D. W., Cesario, J., Jonas, K. J., & Wagenmakers, E.-J. (2017). A Bayesian model-averaged meta-analysis of the power pose effect with informed and default priors: The case of felt power. *Comprehensive Results in Social Psychology*, *2*, 123–138.

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*(4), 779-798.

Heck, D., & Gronau, Q. (2017). *metabma: Bayesian model averaging for random-and fixed-effects meta-analysis [r package]*.

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (in press). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*.

JASP Team. (2019). *JASP (Version 0.11)[Computer software]*. Retrieved from https://jasp-stats.org/

Jeffreys, H. (1939). *Theory of probability*. Oxford: Claredon.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773-795. Retrieved from http://amstat.tandfonline.com/doi/abs/10.1080/01621459.1995.10476572

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Chartier, C. R., ... others (2019). Many labs 4: Failure to replicate mortality salience effect with and without original author involvement.

Landy, J. F., Jia, M., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., ... Uhlmann, E. L. (in press). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*.

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor 0.9.12-4.2*. Comprehensive R Archive Network. Retrieved from http://cran.r-project.org/web/packages/

BayesFactor/index.html

Rouder, J. N., Haaf, J. M., & Aust, F. (2018). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, *85*, 41-56. Retrieved from https://doi.org/10.1080/03637751.2017.1394581

Rouder, J. N., Haaf, J. M., Davis-Stober, C. P., & Hilgard, J. (2019). Beyond overall effects: A Bayesian approach to finding constraints in meta-analysis. *Psychological Methods*.

Scheibehenne, B., Gronau, Q. F., Jamil, T., & Wagenmakers, E.-J. (2017). Fixed or random? A resolution through model-averaging. Reply to Carlsson, Schimmack, Williams, and Burkner. *Psychological Science*, *28*, 1698–1701.

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702-712. Retrieved from https://doi.org/10.1177/1745691616658637 (PMID: 27694465)

van Erp, S., Verhagen, A. J., Grasman, R. P. P. P., & Wagenmakers, E.-J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in *Psychological Bulletin* from 1990–2013. *Journal of Open Psychology Data*, *5*.

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., ..., & Wagenmakers, E.-J. (under review). *A multi-site, preregistered, paradigmatic test of the ego depletion effect.* (University of Minnesota, Minneapolis MN.)

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35–57.

Appendix: Preregistration

We will conduct a re-analysis of the Manylabs 4 project (Klein et al., 2019) using Bayesian meta-analytic techniques and multilevel modeling. There has been some debate about the preregistered exclusion criteria leading to a comment on the original manuscript by Chatard et al. (2020). The authors of the comment note that the mortality salience effect is present in the data, but can be statistically detected only when small studies (<40 participants per cell) are excluded and if only expert-advised studies are included. In the preregistration document the Manylabs-4 authors indeed state that power is deemed sufficient if 40 participants per cell (i.e. 80 participants in total) are collected, but the explicit exclusion criterion is 60 participants per study with no requirement on minimum sample size for the two cells. Only including expert-advised studies for the analysis was not preregistered.

In sum, there are now four different possible exclusion criteria under consideration. While we believe that the decision to exclude small studies from the meta-analysis is somewhat unusual—after all, the meta-analytic model is constructed to take sample size into account—, we plan to reanalyze the data using all four different proposed exclusion criteria, in increasing order of strictness:

1. All studies are included.

2. Only studies with data collected from $\geq$ 60 participants are included. This is the preregistered exclusion criterion.

3. Only studies with $\geq$ 40 participants per cell are included (i.e. 80 participants in total).

4. Only studies with $\geq$ 40 participants per cell and only expert-advised studies are included. This is the exclusion criterion used by Chatard et al. (2020).

Note that these are the study-level exclusion criteria. From the included studies we will analyze all participants that responded to all prompts.

We will conduct a model-averaged meta-analysis using JASP and the metaBMA package in R. This analysis will be modeled after Gronau et al. (2017). Specifically, we will use an informed prior distribution on heterogeneity across experiments (van Erp et al., 2017), and three different one-sided priors on group-level effect size: a default Cauchy with scale 0.707, the Oosterwijk prior (Gronau et al., in press), and

the Vohs prior (i.e., a normal distribution with mean 0.30 and standard deviation 0.15, as specified for a recent many-labs study on the ego-depletion effect).

Given that participant-level data are available we will also conduct a Bayesian multilevel analysis modeled after Rouder et al. (2019) where participants are nested in lab sites. We use a similar model to the one used for the embodied cognition reanalysis conducted by Rouder et al. (2019). There are two critical prior settings to consider, the scale settings on μ_θ and σ^2_θ. The scale on μ_θ corresponds to the expected size of the overall effect. As Rouder et al. (2019) we set this scale to 0.4. The scale of σ^2_θ corresponds to the expected amount of variability in effect size across studies. Again, we kept the value of 0.24 as proposed by Rouder et al. (2019).

Both analyses will be conducted using all four data exclusion rules. The interpretation of the results will center, firstly, on the Bayes factor for the presence or absence of an effect, and, secondly, on the size of the effect.