Improving Statistical Analysis in Team Science: The Case of a Bayesian Multiverse of Many Labs 4

Suzanne Hoogeveen
University of Amsterdam, The Netherlands

Sophie W. Berkhout
Utrecht University, The Netherlands

Quentin F. Gronau
University of Newcastle, Australia

Eric-Jan Wagenmakers
University of Amsterdam, The Netherlands

Julia M. Haaf
University of Amsterdam, The Netherlands

Abstract

Team science projects have become the gold standard for assessing the replicability and variability of key findings in psychological science. However, we believe the typical meta-analytic approach in these projects fails to match the wealth of collected data. Instead, we advocate the use of Bayesian hierarchical modeling for team science projects, potentially extended in a multiverse analysis. We illustrate this full-scale analysis by applying it to the recently published Many Labs 4 project. This project aimed to replicate the mortality salience effect – that being reminded of one's own death strengthens the own cultural identity. 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: the data provide evidence against a mortality salience effect across the majority of our analyses. We issue general recommendations to facilitate full-scale analyses in team science projects.

Keywords: Bayes factor, Bayesian hierarchical modeling, Replication, Team science

Introduction

A salient recent reform in psychological science is the trend towards 'team science'. In crowd-sourced collaborative projects, many different sites across the globe jointly collect data to answer questions about replicability and variability of effects

This research was supported in part by a Vici grant from the Netherlands Organisation for Scientific Research (NWO; 016.Vici.170.083) to EJW, by an NWO Veni grant (VI.Veni.201G.019) to JMH, by an NWO grant to QFG (406.16.528), and by a talent grant from the Amsterdam Brain and Cognition (ABC) research platform, University of Amsterdam (T0921) to SH.
Analysis code is provided at https://github.com/SuzanneHoogeveen/ml4-reanalysis.
Correspondence should be sent to Suzanne Hoogeveen, University of Amsterdam, Nieuwe Achtergracht 129B, 1018 WT Amsterdam, The Netherlands. E-mail may be sent to suzanne.j.hoogeveen@gmail.com.

(Chartier et al., 2018; Forscher et al., in press; Uhlmann et al., 2019). These team science efforts have become the gold standard for assessing the robustness of key findings in the psychological literature. Noteworthy examples of such large-scale endeavours are The Reproducibility Project: Psychology (Open Science Collaboration, 2015), Many Labs (Ebersole et al., 2016; Klein et al., 2022, 2014, 2018), ManyBabies (Frank et al., 2017; The ManyBabies Consortium, 2020), the Pipeline Project (Schweinsberg et al., 2016), and the Psychological Science Accelerator (Chen et al., 2018; Jones et al., 2021; Moshontz et al., 2018). These crowd-sourcing data collection efforts allow researchers to obtain larger samples and hence increase statistical power as well as to reach traditionally less-studied populations (i.e., non-Western participants; Henrich, Heine, & Norenzayan, 2010).

Given the wealth of data that is obtained in these collaborative projects, we believe it is important to fully make use of the available information in the statistical analysis. Unfortunately, the analytic strategies that are often taken in team science projects may not do justice to the collected data. While some projects, such as ManyBabies have conducted sophisticated hierarchical analyses, most of the Many Labs projects and other large-scale team science projects have used standard meta-analytic approaches. In these standard analyses, the data are summarized per lab or site and a frequentist meta-analysis is conducted, in which either a fixed or random effects structure is applied. We will refer to this type of analysis with compressed data as a 'minimal analysis'. We believe a minimal analysis constitutes a missed opportunity, as it both limits analytic possibilities and compromises the informativeness of the data. For instance, in a meta-analysis, one cannot investigate participant-level predictors and the data are reduced to mean effect size and its standard error per lab, thereby losing information about the primary data. A huge advantage of large-scale team science projects is that participant-level and sometimes even trial-level data within a person are available, so we believe one should use their full potential.

In the following, we will argue for what we will call a full-scale analysis instead of the minimal analysis in team science efforts. Specifically, we will demonstrate the usefulness of Bayesian hierarchical modeling (also known as multilevel modeling; see also Rouder, Haaf, Davis-Stober, & Hilgard, 2019). We first highlight general advantages of the Bayesian modeling approach and then illustrate our method by applying it to the recently published Many Labs 4 project (Klein et al., 2022).

Many Labs 4 is a large scale attempt to replicate the mortality salience effect from Terror Management Theory (Greenberg, Pyszczynski, Solomon, Simon, & 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. (2022) 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 meta-analytic effect size of $g = 0.07$, 95% CI = $[-0.03, 0.17]$. The authors concluded that they found "little evidence that priming mortality salience increased worldview defense compared to a control condition" and that "[t]he present evidence does [...] provide an important challenge for TMT to address" (Klein et al., 2022, p.10).

Bayesian Hierarchical Modeling

So what should such a full-scale analysis look like for Many Labs 4 or other team science projects? In the following we will describe four features that we believe a full-scale analysis for team science projects should include.

First, we believe a Bayesian analysis is preferred over a frequentist analysis, as the former allows one to obtain evidence for the null-hypothesis and to quantify (posterior) uncertainty (Wagenmakers et al., 2018). Especially in replication studies, the chances of obtaining null results are considerable. 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 large-scale replication efforts is that it allows a distinction to be made between evidence for the absence of the effect of interest and the absence of evidence for or against the effect (Keysers, Gazzola, & Wagenmakers, 2020). In other words, failure to successfully replicate a key effect could mean that the data are undiagnostic for determining whether or not the effect is present, or it could mean that the data provide substantial evidence against the presence of the effect. Obviously, this difference is highly consequential for interpreting the results of a study.

The second feature of a full-scale analysis relates to the hierarchical nature of data in team science projects. That is, instead of a meta-analysis, we advocate the use of a hierarchical model including all primary data, with participants nested within labs (Hoogeveen, Haaf, et al., 2022; Rouder et al., 2019). In a hierarchical model, the lowest-level data are nested within their higher-level groups, such as trials nested within participants, or participants nested within labs or countries. This structure makes it possible to assess general or overall effects as well as individual or lab-specific deviations from those overall effects. For a tutorial on Bayesian hierarchical modeling, we refer the reader to Veenman, Stefan, and Haaf (2022). Additional demonstrations of the Bayesian hierarchical modeling approach for team science efforts can be found in Hoogeveen, Haaf, et al. (2022), Gervais et al. (2017), Tierney et al. (2021) and Tierney et al. (in preparation). The hierarchical approach for team science efforts brings several benefits. First, by capitalizing on the full resolution of the data, no information is lost in the interim aggregation process. For instance, in a meta-analysis a relatively large standard error for a given lab or site might either reflect a heterogeneous sample or simply a small sample. In a hierarchical model, the source of the (im)precision of the estimate is retained and thus can be interpreted. Second and relatedly, hierarchical shrinkage reduces the influence of outlying labs with small samples, hence automatically weighing the contribution of the different labs towards the global estimate (Efron & Morris, 1977). Third, while study-level predictors may be included in a meta-analysis, the hierarchical model additionally allows for the inclusion of participant-level predictors and/or the assessment of interaction effects. Finally, in the hierarchical approach we can easily evaluate whether effects meaningfully differ per site/lab (e.g., in terms of WEIRDness or cross-cultural robustness; e.g., Hoogeveen, Haaf, et al., 2022).

The third feature of a full-scale analysis concerns the inclusion of theoretical constraint in the statistical analysis (Haaf, Klaassen, & Rouder, 2018; Haaf & Rouder, 2017; Rouder et al., 2019). Psychological theories typically constrain behavioral data in the sense that theories dictate ordinal predictions; observed effects are described in the form of "manipulation X causes *higher* scores on Y or *slower* responses" or "higher scores on X are associated with *lower* scores on Y or *faster* responses". Given the ordinal nature of the hypotheses, we believe statistical tests should reflect the theoretical predictions about the direction of effects. For instance, we expect participants who imagined their own death to identify *more* with American culture than participants who imagined watching TV, rather than just a difference between conditions.

The hierarchical nature of the data in team science projects allows for more informative testing of ordinal predictions beyond directional constraint at the aggregate level. Specifically, rather than testing whether on average, participants who imagined their own death identify more with American culture than participants who imagined watching TV, we can also test whether this pattern holds across every lab that is included in the analysis. This latter constitutes a much riskier prediction, as we now need the effect to be present in every single lab (e.g., 21 times, instead of once). This risky prediction is potentially rewarded in terms of evidence when the data reflect the predicted pattern, boosting the effect's credibility. Rouder et al. (2019) refer to this "does every study?" question as a test of qualitative differences, as it provides information on whether the effect of interest is qualitatively equal across studies (i.e., in the same direction).

Bayesian modeling methods are particularly well suited to test ordinal constraints at different levels, such as "is the overall mortality salience effect positive" or "does every lab show a positive mortality salience effect?". We would therefore advocate to include both versions of these theoretically-motivated ordinal constraints in the statistical analysis for team science projects (see also Rouder et al., 2019 and Haaf & Rouder, 2017 for application of these ordinal constraints in meta-analysis and individual cognitive performance, respectively).

Finally, the fourth feature of a full-scale analysis relates to assessing the robustness of the findings. That is, beyond using Bayesian hierarchical modeling, we believe team science projects can at least sometimes benefit from conducting a multiverse analysis (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016). In a multiverse analysis, the researcher can evaluate different potential constellations of the data (e.g., exclusions, theoretically relevant subgroups), priors and predictors without committing to one –perhaps arbitrarily– chosen analysis path. As will be demonstrated by the Many Labs 4 example below, there are often multiple defensible analytic choices that can be considered. A complete assessment of the robustness of a given effect might thus require many labs as well as many analyses (Wagenmakers, Sarafoglou, & Aczel, 2022). The multiverse approach not only presents a broader and more complete picture of the results, it also allows one to explore the consequences of analytic choices. For instance, does only including an ideal subgroup of participants indeed increase the evidence for the presence of the effect? Do exclusions based on manipulation-checks affect the evidence? Does the particular operationalization of a construct make a difference? Furthermore, the Bayes factor model comparison approach allows for a straightforward interpretation of the multiverse results; Bayes factors are continuous measures of relative evidence for one model over another (e.g., a common-effect model vs. a null model) so they can be compared across different multiverse paths. Moreover, Bayes factors automatically take into account the sample size and reflect the informativeness of the data.

Many Labs 4 Reanalysis

Given the outlined advantages of a full-scale analysis, in the following we will present a Bayesian multiverse reanalysis of the Many Labs 4 data using hierarchical models. Note that we also conducted a Bayesian model-averaged meta-analysis (Gronau, Heck, Berkhout, Haaf, & Wagenmakers, 2021), which is reported in the Appendix. The results of the model-averaged meta-analysis are qualitatively comparable to those of the hierarchical modeling reported below.

A Brief History

In December 2019, the Many Labs 4 authors posted a preprint of the project on PsyArXiv (Klein et al., 2019). Soon after, a critique of the analysis emerged in which Chatard, Hirschberger, and Pyszczynski (2020) pointed out that Klein and colleagues had not followed 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. Intrigued by these divergent reports, we then decided to conduct a Bayesian multiverse analysis. A preprint of this analysis was published in 2020 on PsyArXiv. Then in 2022, Klein and colleagues published their final results in *Collabra*, after which we also revisited the data, resulting in the current paper.

Include or Exclude?

Which of the different proposed analyses –Klein or Chatard– 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 to subset the full sample of 2,281 participants across 21 labs.

The full set of exclusion criteria employed by either Klein et al., Chatard et al., and ourselves consists of 5 layers of exclusion settings resulting in $3 \times 3 \times 2 \times 2 \times 2 = 72$ constellations of exclusion criteria. In the Appendix, we report the full set of these criteria as well as the rationale for choosing either of those. Table 1 shows all 45 unique constellations, the resulting number of studies and total number of included participants (see the Appendix for a table with all 72 constellations). In short, the *participant-level criteria* refer to theory-based arguments about among whom the mortality salience effect should occur: According to the original authors the effect may only be present among (2) those who self-identify as white and report to be born in the United States or even only (3) those who are white, American-born, and strongly identify with American culture (a score of 7 or higher on a 9-pt Likert scale). *N-based criteria* refer to the inclusion of labs based on the number of participants recruited per lab. *Protocol criteria* refer to the inclusion of both in-house labs and author-advised labs or only the latter, based on the suggestion that the effect may only emerge in author-advised studies as the mortality salience effect is highly sensitive to nuances in the study implementation (Greenberg et al., 1994). *Timing-based criteria* address Klein et al.'s decision to discard all observations collected by some in-house labs prior to the preregistration date (February 15th, 2017), resulting in the exclusion of 566 participants (25.4%). While we considered this exclusion wasteful and unnecessary, we added it as another layer in the multiverse analysis for the sake of completeness. The final layer of exclusion settings refers to the way in which the participant-level criteria are applied. That is, Klein et al. and Chatard et al. applied the participant-level exclusion criteria *only* to the author-advised protocols, which means that for exclusion criteria 2 and 3 all participants from the in-house labs where retained. However, since exclusion criteria 2 and 3 were specified by the original authors as a strict and genuine test of the theory, we believe that it is important to thoroughly apply these criteria to all participants, even if this means discarding participants where this information is unavailable.

Table 1

Exclusion constellations and resulting sample sizes

Participant-level	N-based	Protocol	Timing-based	Apply P-based	Sample Size	Labs
All	All	All	All	AA only	2225	21
White & US-born	All	All	All	AA only	1880	21
US-Identity > 7	All	All	All	AA only	1699	21
All	N > 60	All	All	AA only	2067	17
White & US-born	N > 60	All	All	AA only	1746	17
US-Identity > 7	N > 60	All	All	AA only	1593	17
All	N > 80	All	All	AA only	1866	14
White & US-born	N > 80	All	All	AA only	1545	14
US-Identity > 7	N > 80	All	All	AA only	1392	14
All	All	AA	All	AA only	798	9
White & US-born	All	AA	All	AA only	453	9
All	N > 80	AA	All	AA only	699	7
White & US-born	N > 80	AA	All	AA only	378	7
US-Identity > 7	N > 80	AA	All	AA only	225	7
All	All	All	After prereg	AA only	1659	20
White & US-born	All	All	After prereg	AA only	1314	20
US-Identity > 7	All	All	After prereg	AA only	1133	20
All	N > 60	All	After prereg	AA only	1544	17
White & US-born	N > 60	All	After prereg	AA only	1223	17
US-Identity > 7	N > 60	All	After prereg	AA only	1070	17
All	N > 80	All	After prereg	AA only	1343	14
White & US-born	N > 80	All	After prereg	AA only	1022	14
US-Identity > 7	N > 80	All	After prereg	AA only	869	14
All	All	AA	After prereg	AA only	797	9
White & US-born	All	AA	After prereg	AA only	452	9
US-Identity > 7	All	AA	After prereg	AA only	271	9
All	N > 60	AA	After prereg	AA only	698	7
White & US-born	N > 60	AA	After prereg	AA only	377	7
US-Identity > 7	N > 60	AA	After prereg	AA only	224	7
All	All	All	All	AA and IH	2211	21
White & US-born	All	All	All	AA and IH	983	16
US-Identity > 7	All	All	All	AA and IH	272	9
All	N > 60	All	All	AA and IH	2053	17
White & US-born	N > 60	All	All	AA and IH	897	13
All	N > 80	All	All	AA and IH	1852	14
White & US-born	N > 80	All	All	AA and IH	864	12
All	All	AA	All	AA and IH	799	9
All	N > 60	AA	All	AA and IH	700	7
All	All	All	After prereg	AA and IH	1650	20
White & US-born	All	All	After prereg	AA and IH	777	15
All	N > 60	All	After prereg	AA and IH	1535	17
White & US-born	N > 60	All	After prereg	AA and IH	702	13
All	N > 80	All	After prereg	AA and IH	1334	14
White & US-born	N > 80	All	After prereg	AA and IH	669	12

Note. Orange rows refer to Klein et al.'s key analyses; green rows refer to Chatard et al.'s key analyses; purple rows refer to our currently chosen analyses; AA = author-advised; IH = in-house. 'Apply P-based' indicates whether the participant-level exclusion criteria are applied to the author-advised labs only (retaining all in-house participants) or to both author-advised and in-house labs (missing data excluded).

In the following we will report a reanalysis for the three exclusion constellations of the key analyses from Klein et al. (2022, orange rows in Table 1), the three exclusion constellations from Chatard et al. (2020, green rows), and our own choice of exclusion criteria (purple rows). Subsequently, lacking compelling argumentation for or against any of the criteria, we decided to conduct an analysis based on the entire set of 45 unique constellations as a multiverse analysis (Steegen et al., 2016).

Disclosures

Preregistration

Our analyses, including prior settings, were preregistered on the Open Science Framework (osf.io/ae4wx, see also Appendix C). 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. Moreover, two additional exclusion layers only became apparent after the final version of the Many Labs 4 report was published, namely those related to the timing-based exclusion criteria and the application of the participant-level criteria to the author-advised only or author-advised and in-house labs. We believe that including these additional paths in the multiverse analysis helps to provide a more complete analysis. We also note that the preregistration includes both the hierarchical analysis and the model-averaged meta-analysis. The latter is reported in Appendix B.

Data and Materials

Readers can access the data and the R code to conduct all analyses (including all figures) at github.com/SuzanneHoogeveen/ml4-reanalysis.

Reporting

This study involved an analysis of existing data rather than new data collection.

Ethical approval

No ethical approval was required for this work as we did not collect any data.

Methods

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 (i.e., identification with American culture, operationalized through relative preference for American vs. non-American authors), and participants are nested in studies resulting in a hierarchical data structure. We employed a modeling approach similar to the one developed for the embodied cognition reanalysis by Rouder et al. (2019). That is, we used Bayes factor model comparison with hierarchical models reflecting different structures of the data, varying in the extent to which they constrain their predictions. We believe this approach satisfies the analytic desiderata for team science projects outlined before, namely: appropriately accounting for the nested structure of the data without compromising on informativeness, directly testing both the presence of an overall mortality salience effect, as well as the presence of between-study heterogeneity, and reflecting theoretical constraints on the direction of the effect.

Concretely, there are four models under consideration: (1) The null model corresponds to the notion that none of the studies show an effect; this model assumes no overall experimental effect nor heterogeneity between studies (2) The common-effect model corresponds to the notion that all studies show the same effect in the expected direction; this model assumes no heterogeneity between studies (3) The positive-effects model corresponds to the notion that all studies show an effect in the expected direction, yet to varying degrees; and (4) the unconstrained model refers to the notion that the overall effect and study effects may vary freely (in direction and size). We compute Bayes factors for models (2), (3), and (4) against model (1), the null model. Evidence for model (1) would indicate the absence of a mortality salience effect across all labs; evidence for (2) would indicate that on average, people who contemplate their own death identify more strongly with their culture than people who contemplate watching TV, to a similar degree across labs; evidence for (3) would indicate that in all of the labs, people who contemplate their own death identify more strongly with their culture than people who contemplate watching TV, but to varying degrees across labs; evidence for (4) would indicate that in some labs, people who contemplate their own death identify more strongly with their culture than people who contemplate watching TV whereas in other labs, people who contemplate watching TV identify more strongly with their culture than people who contemplate their own death.

Box 1. Model specifications

The base model for the mortality salience effect is a mixed linear model. Let Y_{ijk} be the rating for the ith lab, the jth participant, and the kth condition. Then

$$Y_{ijk} \sim N(\alpha_i + x_k\theta_i, \sigma^2),$$

where α_i is the ith lab's specific overall culture identification rating effect, and θ_i is the ith lab's mortality salience effect. The variable $x_k = -0.5, 0.5$ if $k = 1, 2$ respectively, with $k = 1$ when condition is 'watching TV' and $k = 2$ when condition is 'contemplate death'. Here, θ_i is the parameter of interest, that is varied across models to reflect the different constraints. Specifically, for the null model we specify $\theta_i = 0$. For the common-effect model $\theta_i = v$ where v represents the true value for the mortality salience effect that is constrained to be positive ($v > 0$). For the positive-effects model θ_i comes from a distribution with a mean mortality salience effect (μ_θ) and between-study variability in the size of this effect (σ^2_θ): $\theta_i \sim N_+(\mu_\theta, \sigma^2_\theta)$ where the N_+ represents a normal distribution truncated at below zero to reflect the prediction that $\theta_i > 0$. Finally, for the unconstrained model, we let θ_i free to vary in size and direction: $\theta_i \sim N(\mu_\theta, \sigma^2_\theta)$.

There are two critical prior settings to consider: the scale setting on the overall effect (μ_θ) and the scale setting on the between-lab heterogeneity (σ^2_θ). These scales can be roughly interpreted as standardized effect size such as Cohen's d. 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) – i.e., 60% of the expected overall effect.

The Bayesian hierarchical modeling is conducted using the R-package `BayesFactor` (Morey & Rouder, 2018).

Results

In the following, we will first reanalyze the data from the key findings reported by Klein et al. (2022) using our proposed full-scale analysis, then those from Chatard et al. (2020), and finally report the analysis of the data based on our own choice of exclusion criteria constellations.

Bayesian Reanalysis of Klein et al.'s Key Findings

Figure 1A shows the observed, unstandardized effects and the estimates from the unconstrained hierarchical model for the first participant-level exclusion criterion. This is the main analysis that is the basis for the key claims of the Many Labs 4 project, as reported in the published paper (Klein et al., 2022). The authors included participants whose data was collected after the lead team posted their preregistration, and only studies that featured more than 60 observations (before participant-level exclusions). The participant-level exclusion criteria were only applied to author-advised studies, while all participants from the in-house studies were retained.

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 0.02, 95% CI $[-0.12, 0.16]$ for participant-level exclusion criterion 1, 0.04, 95% CI $[-0.11, 0.19]$ 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 2. 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 5.35-to-1 to 4.21-to-1 in favor of the null model depending on the different participant-level exclusion criteria. If we allowed for variability across effects but maintained that the effect should be present across *all* studies, we would obtain strong evidence against this hypothesis, with Bayes factors ranging between 629-to-1 and 158-to-1 in favor of the null model over the positive-effects model. In sum, this pattern indicates evidence against an overall mortality

Table 2

Bayes factors for key analyses.

			Evidence			
Participant-level	N	Labs	BF_{0f}	BF_{01}	BF_{0+}	Effect [95% CI]
Klein et al. (2022)						
All	1544	17	12.44	5.35	628.62	0.02 [-0.12, 0.16]
White & US-born	1223	17	9.35	4.21	204.55	0.04 [-0.11, 0.19]
US-Identity > 7	1070	17	6.95	4.50	157.52	0.05 [-0.11, 0.21]
Chatard et al. (2020)						
All	699	7	14.61	2.16	13.32	0.08 [-0.12, 0.28]
White & US-born	378	7	6.76	0.95	2.61	0.14 [-0.11, 0.39]
US-Identity > 7	225	7	4.47	0.79	1.42	0.18 [-0.12, 0.49]
Current choice						
All	2211	21	35.21	10.33	10,490.12	0.01 [-0.11, 0.12]
White & US-born	983	16	21.46	13.99	2,538.61	-0.04 [-0.20, 0.12]
US-Identity > 7	272	9	8.44	2.77	11.88	0.05 [-0.22, 0.32]

Note. All Bayes factors are reported in favor of the null model.

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 1.

In summary, the null results are consistent across participant-level exclusion criteria. 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 observations: The smaller the number of observations, the less evidence there is in any direction, and the wider the estimated posterior distribution of the overall effect.

Bayesian Reanalysis of Chatard et al.'s Key Findings

We also reanalyzed Chatard et al.'s findings with a hierarchical modeling approach. Figure 2 shows study estimates from the unconstrained model for the unstandardized effects. All confidence intervals and credible intervals cover zero.

Effect size estimates from the unconstrained model (similar to Cohen's d) are 0.08, 95% CI $[-0.12, 0.28]$ for participant-level exclusion criterion 1, 0.14, 95% CI

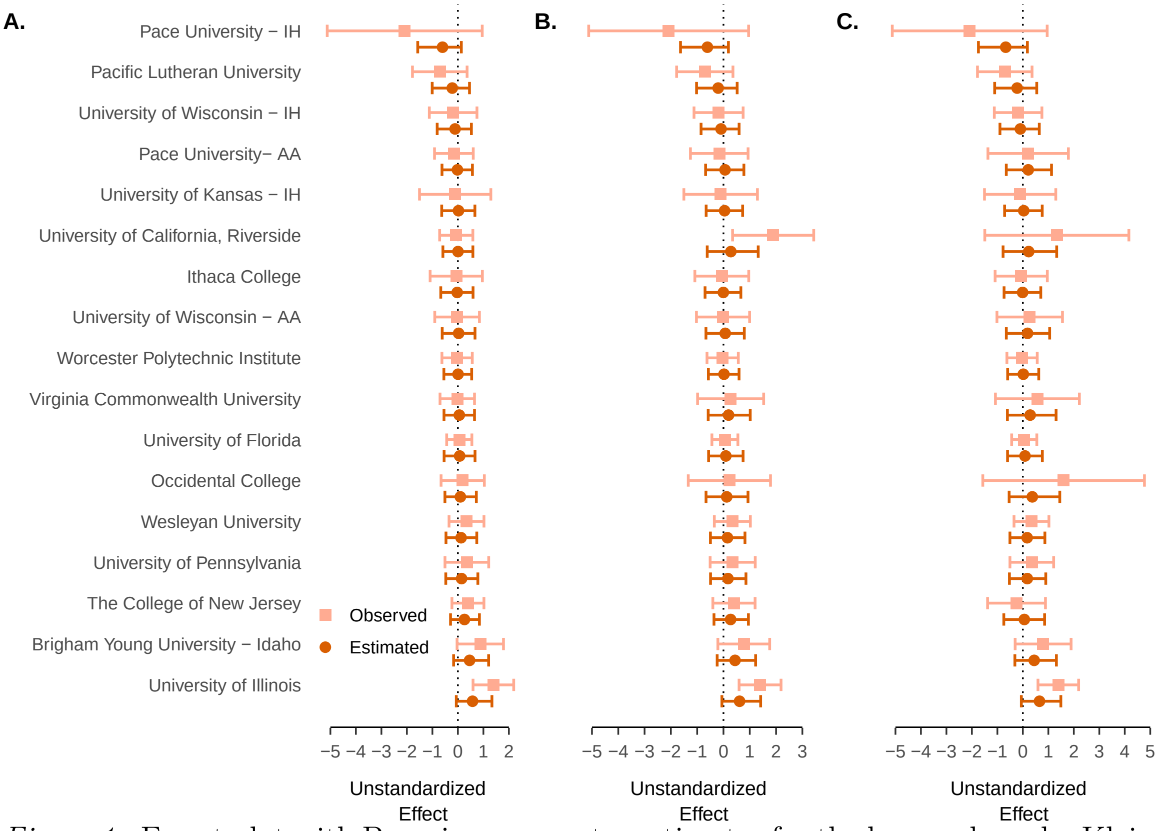

Figure 1. Forest plot with Bayesian parameter estimates for the key analyses by Klein et al. for the three participant-level exclusion sets (applied to author-advised protocol participants only) with data collected after the lead team posted their preregistration, and only studies that featured more than 60 observations. **A.** Participant-level exclusion set 1. The light orange squares represent unstandardized observed effects for each study with 95% confidence intervals. The dark orange points represent estimated unstandardized effects from the unconstrained model with 95% credible intervals. **B.** Participant-level exclusion set 2. **C.** Participant-level exclusion set 3. The estimates are sorted by the size of the observed effects for participant-level exclusion set 1 (i.e., panel A).

$[-0.11, 0.39]$ for participant-level exclusion criterion 2 and 0.18, 95% CI $[-0.12, 0.49]$ for participant-level exclusion criterion 3. Note that all credible intervals include zero, and even though the posterior mean increases with more conservative exclusion criteria, the width of the 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 across exclusions than the estimation results. Bayes factors are shown in the middle three rows of Table 2. The pattern of Bayes factors is 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 $\text{BF}_{01} = 2.16$. 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 1.29 in favor of the common-effect model over the null model). In sum, the Bayes factors results are in line with the absence of any (consistent) evidence for or against an effect. Across the three participant-level exclusion criteria, 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 Analysis of our Current Choice

We also included an analysis of the Many Labs 4 data using our own choice of the exclusion criteria. Following Klein et al. (2022) and Chatard et al. (2020), we looked at all three participant-level exclusion criteria, while settling on one particular choice for the other factors that seemed most sensible to us. The goal for this choice was to include the maximum number of participants but still adhering to the recommendations by the original authors to give the effect the best chance. Specifically, we included all complete data, from all labs and protocols and applied the participant-level exclusions to both author-advised labs and in-house labs, discarding missing values. For exclusion criterion 1 – completeness of the measures –,

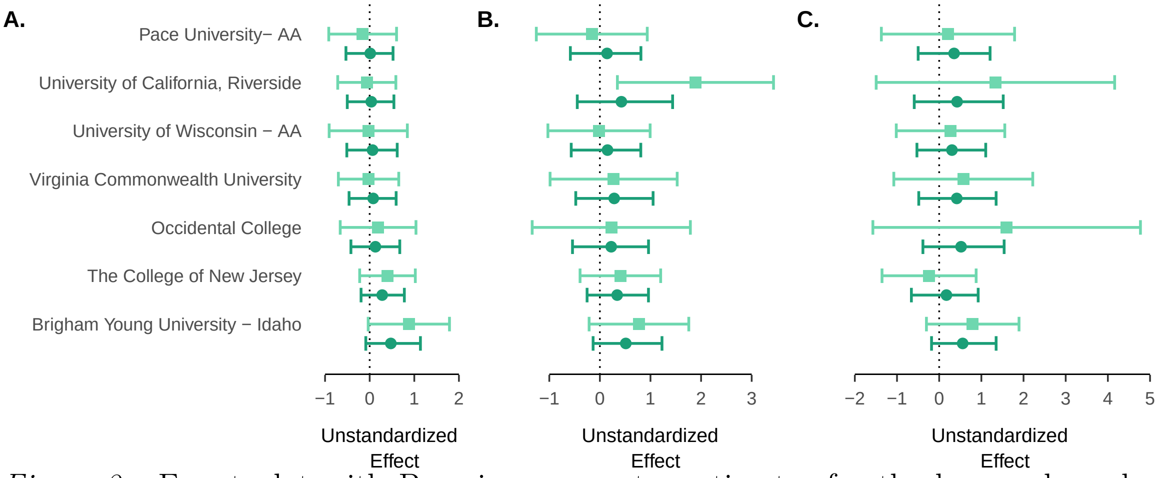

Figure 2. Forest plot with Bayesian parameter estimates for the key analyses by Chatard et al. for the three participant-level exclusion sets, only studies that featured more than 80 observations, and for author-advised labs only. **A.** Participant-level exclusion set 1. The light green squares represent unstandardized observed effects for each study with 95% confidence intervals. The dark green points represent estimated unstandardized effects from the unconstrained model with 95% credible intervals. **B.** Participant-level exclusion set 2. **C.** Participant-level exclusion set 3. The estimates are sorted by the size of the observed effects for participant-level exclusion set 1 (i.e., panel A).

we did retain participants for labs where no explicit information on missingness was available, as long as they were assigned to an experimental condition and answered both items of the dependent variable.[1] Note that our choice of analysis paths leads to quite variable numbers of participants (between $N = 2{,}211$ and $N = 272$).

Figure 3 shows the study estimates from the unconstrained model for the unstandardized effects. Again, all confidence and credible intervals include zero. Effect size estimates from the unconstrained model are 0.01, 95% CI $[-0.11, 0.12]$ for participant-level exclusion criterion 1, -0.04, 95% CI $[-0.20, 0.12]$ for participant-level exclusion criterion 2, and 0.05, 95% CI $[-0.22, 0.32]$ for participant-level exclusion criterion 3. Again, all credible intervals overlap with zero, and the width of the credible interval increases with fewer observations included in the analysis, as less data implies more uncertainty. Note that the absence of estimates for certain labs in panels B and C of Figure 3 is due to the participant-level exclusion criteria leaving no participants in these particular labs (rather than excluding any labs per se).

[1] Klein et al. (2022) also retained participants from one author-advised lab where information on completeness of the data was unavailable, so we consider this a reasonable decision.

The Bayes factors paint a similar picture; the evidence against the presence of the mortality salience effect is stronger with a larger sample size. For the most inclusive sample with exclusion criterion 1 (N = 2,211), the null model outperforms the common-effect model by a factor of 10.33. For exclusion criterion 2, the estimated effect goes slightly in the direction opposite to the hypothesis, hence the Bayes factor more strongly favors the null-model over the common-effect model, $\text{BF}_{01} = 13.99$. Finally, for the most strict exclusion criterion, only 272 observations are retained. As a result, we get a much weaker Bayes factor of 2.77 in favor of the null-model over the common-effect model. In sum, with our chosen set of exclusion criteria, we obtained strong to weak evidence against the mortality salience effect.

Bayesian Multiverse Analysis Across All Exclusion Criteria

To assess the robustness of the previously reported results we conducted a multiverse analysis using the 45 unique data sets from Table 1. We used the same hierarchical model construction as reported above and report here the Bayes factors for the presence of an effect against its absence. The Bayes factors are plotted in Figure 4 (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 $\text{BF}_{\text{effect0}}$ is the weighted average of the evidence for the common-effect versus the null model and the unconstrained model (varying effect) versus the null model. The x-axis refers to the evidence for between-study heterogeneity in the data. The $\text{BF}_{\text{heterogeneity0}}$ is calculated by taking the evidence for the unconstrained model versus the common-effect model. The size of the points reflects the number of participants whose data are included in the analysis and the color of the points highlights the key analyses described before.

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. Additionally, none of the analysis paths provide evidence for heterogeneity (all $\text{BF}_{\text{heterogeneity0}} < 0$).

In sum, the evidence against the mortality salience effect appears relatively robust against choices of exclusion criteria. When conducting a large number of analyses on the same data some of these analyses will almost inevitably lead to

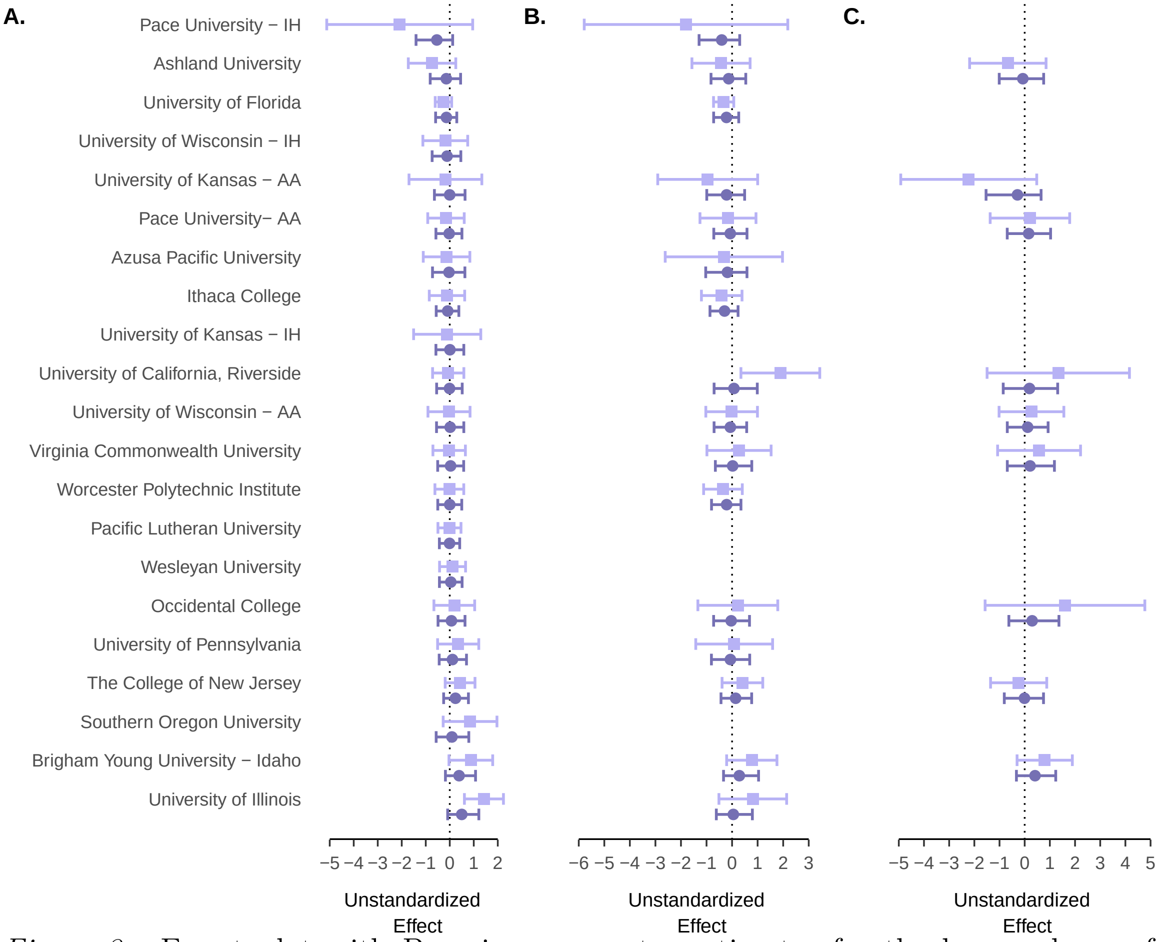

Figure 3. Forest plot with Bayesian parameter estimates for the key analyses of our choice for the three participant-level exclusion sets (applied to both author-advised and in-house protocol participants), including all participants and all labs. **A.** Participant-level exclusion set 1. The light purple squares represent unstandardized observed effects for each study with 95% confidence intervals. The dark purple points represent estimated unstandardized effects from the unconstrained model with 95% credible intervals. **B.** Participant-level exclusion set 2. **C.** Participant-level exclusion set 3. The estimates are sorted by the size of the observed effects for participant-level exclusion set 1 (i.e., panel A).

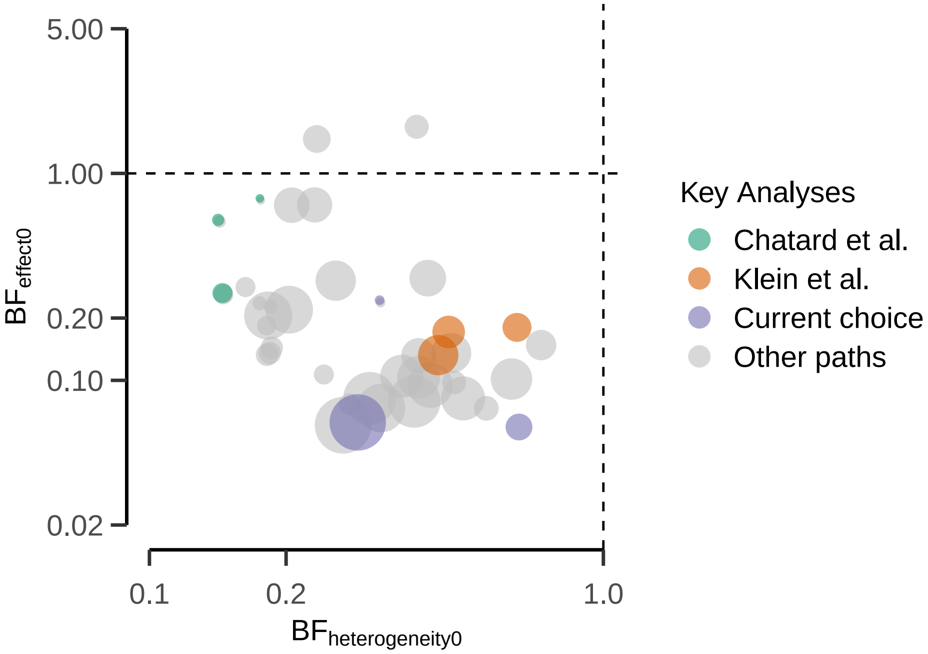

Figure 4. Results from the Bayesian multiverse analysis: Bayes factors in favor of a mortality salience effect are above the horizontal line, Bayes factors against the mortality salience effect are below the horizontal line. All analyses provide evidence against between-study heterogeneity as shown by all heterogeneity Bayes factors are smaller than 1 on the x-axis. The color of the points refers to the different key analyses sets, and the size of the points refers to the number of participants the analysis is based on. All but 2 of analyses provide evidence against the mortality salience effect.

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 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 two Bayes factors that are weakly in favor of the mortality salience effect are based on less than half of the original data and prefer the presence of the effect only by a factor of 1.5 and 1.7.

Prior Sensitivity

In addition to assessing the effects of various data exclusion decisions, we might also investigate the role of prior choices on inference. Specifically, we looked at the dependence of the Bayes factors on the prior settings for the overall effect and for the between-study variability. While some researchers have argued that the influence of the prior on the results should be minimized (e.g., by using uninformative default settings; Aitkin, 1991; Gelman, Carlin, Stern, & Rubin, 2004; Kruschke, 2013), we believe the influence of the prior is a meaningful and inherently informative element of Bayesian inference (Rouder et al., 2018; Vanpaemel, 2010). Nevertheless, the extent to which reasonable prior choices affect the results clearly speaks to the robustness of the conclusions.

For the main analysis, we used a scale of 0.4 on the overall effect (μ_θ)and 0.24 on the between-study variability (σ^2_θ), indicating an expected effect size of Cohen's *d* around 0.4 and 60% of that effect size for the variability between labs (see top rows for each analysis set in Table 3). To examine the Bayes factors under different prior settings, we systematically both doubled and halved the scales on μ_θ and σ^2_θ, reflecting expectations of a small effect, a medium-to-large effect, very little between-study variability and medium between-study variability.

Table 3 shows the Bayes factors resulting from crossing these combinations for the key analyses (for participant-level exclusion set 1) and Figure 5 shows the evidence across all 45 unique data exclusion paths for each of the 4 different prior setting combinations. Most support for the effect is obtained under the expectation of a small effect and most support for between-study heterogeneity is obtained under the expectation of little between-study variability. Nevertheless, across all settings, evidence somewhat in favor of a mortality salience effect only occurs in 12/180 (6.7%) paths. Evidence in favor of heterogeneity across studies is obtained under 2/180 (1.1%) paths. Another observation from these plots is that while the prior setting for the overall effect changes the global strength of the evidence, it does not appear to affect the multiverse paths differentially, as the dots seem to move upwards or downward uniformly. In contrast, the prior setting for the between-study variability not only affects the overall evidence for heterogeneity, it also influences the range of the Bayes factors between multiverse paths. Specifically, the prior expectation of little variability reduces the evidence against heterogeneity and makes the Bayes factors more similar across paths, whereas the expectation of much variability not only leads to more evidence against heterogeneity but also enhances the differences between paths. In sum, choices of prior scales can slightly boost or reduce the evidence in favor of the effect. Yet, in this case, the effects of reasonable prior choices are rather contained; the null-model is still consistently preferred over models with a mortality

Table 3
Bayes factors for key analyses (participant-level exclusion set 1) under different prior settings.

scale on μ_θ	scale on σ_θ^2	BF_{01}	BF_{0f}	BF_{0+}
Klein et al. (2022)				
0.40	0.24	5.41	12.50	791.66
0.20	0.12	2.86	3.92	24.79
0.20	0.48	2.85	27.14	90,214.79
0.60	0.12	7.93	11.18	19.35
0.60	0.48	7.79	76.50	246,167.53
Chatard et al. (2020)				
0.40	0.24	2.17	14.97	13.93
0.20	0.12	1.32	4.41	2.27
0.20	0.48	1.30	35.54	180.27
0.60	0.12	3.14	9.86	1.82
0.60	0.48	3.14	83.13	191.19
Current choice				
0.40	0.24	10.28	35.75	12,127.13
0.20	0.12	5.28	8.70	166.42
0.20	0.48	5.27	119.11	∞
0.60	0.12	15.37	25.91	174.74
0.60	0.48	15.08	324.46	∞

Note. All Bayes factors are reported in favor of the null model.

salience effect.

Conclusion

We conducted a Bayesian reanalysis of the Many Labs 4 project with varying exclusion criteria and prior settings. In a Bayesian multiverse analysis using hierarchical models we calculated a total of 45 sets of Bayes factors based on different combinations of 5 layers of data exclusion criteria derived from the Many Labs 4 preregistration, the comment by Chatard et al. (2020), the published article by Klein et al. (2022) and our own judgments. 43 out of 45 Bayes factors provide evidence against an overall mortality salience effect, ranging between 1.32-to-1 and 16.94-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 1.45-to-1 and 1.68-to-1 in

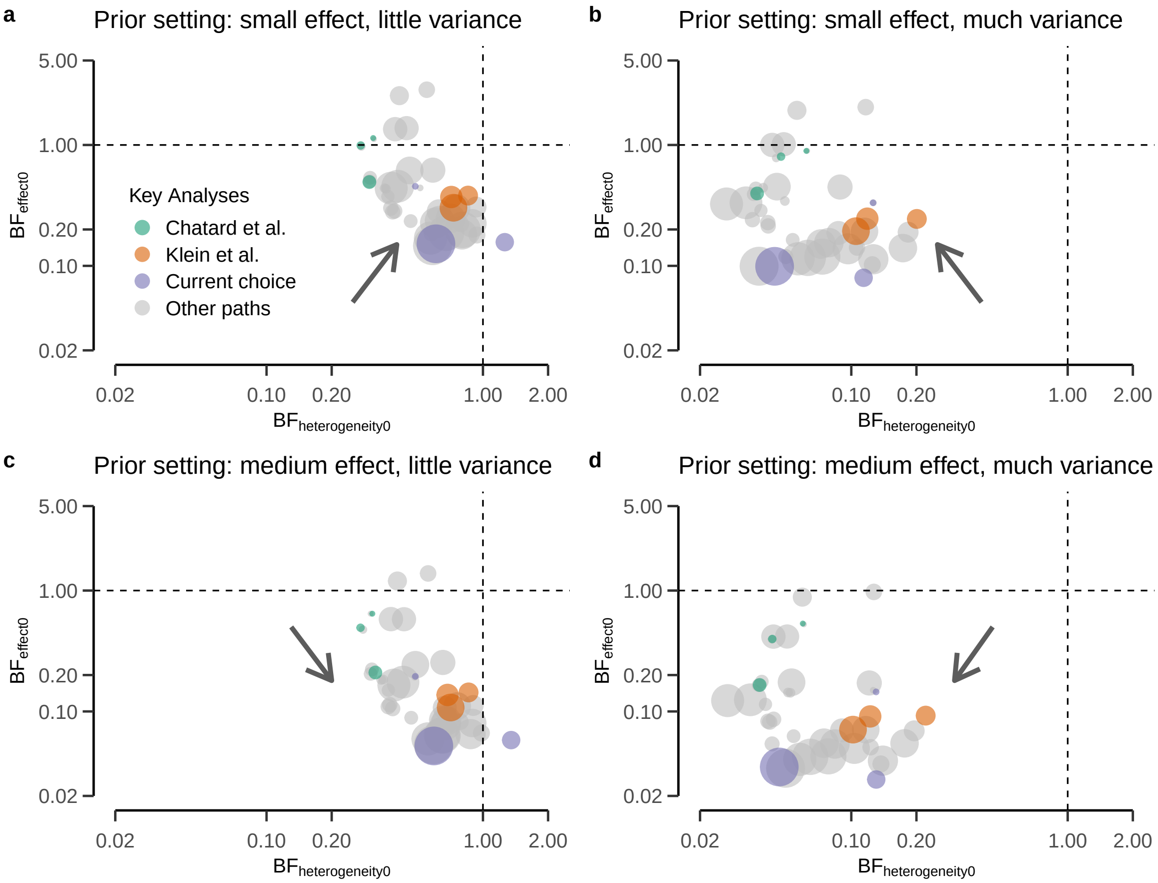

Figure 5. Results from the Bayesian multiverse analysis under different prior settings for the overall effect and the between-study variance in the effect. The arrows show the overall trend relative to the main analysis with the primary prior settings.

favor of the presence of an effect. Additionally, we find some evidence against heterogeneity of effects across studies. Finally, the pattern of results remains qualitatively equal under different reasonable prior settings for the overall effect and the between-study variability. In combination, we would argue we conducted a full-scale analysis of the data provided by the Many Labs 4 project, an inspection from various different angles. Even if we do not believe the evidence from this full-scale analysis and assume there is an effect, this effect is so small (between $d = -0.04$ 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, would require a very specific subgroup of participants, and would therefore also not be generalizable across a broader range of the population.

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.[2]

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 (Harrer, Cuijpers, Furukawa, & Ebert, 2021, Chapter 5). 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 overall. To tackle this issue – and if there was evidence for study heterogeneity – one could include some of the subsetting criteria as dummy-coded predictor in the hierarchical model instead of disregarding the data all together (e.g., author-advised vs. in-house).

Furthermore, we believe the Many Labs 4 case and its development from preprint to published article highlights an important potential drawback of preregistration. In the final article, the Many Labs 4 lead team decided to discard all observations collected prior to the preregistration date, resulting in the removal of more than a quarter of the data. As mentioned, we consider this removal of data wasteful and unnecessary. In this case, we believe the fact that data collection was crowd-sourced and the added value of retaining 556 perfectly valid observations justify "breaking the strict rules of preregistation" that data collection should only start after the analysis plan has been preregistered. As noted by DeHaven (2017), preregistration is "a plan, not a prison". So rather than discarding a large portion of the

[2] Ideally, as with the data exclusion criteria themselves, their weights in a multiverse analysis should be chosen before knowing – or at least without consideration of – the results.

data for the main analyses, we believe a transparent statement on the timing issue would have sufficed in this case. In general, preregistration by definition should not trump common sense and researchers' judgment.

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 exclusion criteria. The Bayesian multiverse approach using hierarchical models provides rich results that go much beyond the original analyses by the Many Labs 4 team. In particular, we believe the current approach satisfies the desiderata of a full-scale analysis in team science projects: (1) providing evidence on a continuous scale from evidence against the crucial effect through inconclusive evidence to evidence in favor of the crucial effect, (2) applying hierarchical modeling to appropriately account for the nested structure of the data, (3) evaluating both the evidence for the experimental effect and the evidence for between-lab heterogeneity, (4) reflecting theoretical constraints on the effect of interest (i.e., ordinal constraints), and (5) evaluating the robustness of the findings by exploring a multitude of relevant analysis paths.

Both Bayes factor model comparison and Bayesian hierarchical modeling are gaining popularity in psychological science. Recent tutorial papers make these approaches more accessible; for instance, see Wagenmakers et al. (2018) and Rouder et al. (2018) for an introduction to Bayes factor model comparison and Veenman et al. (2022) and Rouder and Province (2019) for tutorials on Bayesian hierarchical modeling. Finally, the ease and informativeness of Bayesian multiverse analyses show that this approach should be more generally used to analyze team science projects. The current analyses were conducted in R, and the code is provided at github.com/SuzanneHoogeveen/ml4-reanalysis.

General Recommendations

In sum, we believe the amount of time and effort spent on team science projects and the resulting wealth of data, deserve a full-scale analysis. We believe a Bayesian hierarchical modeling approach is ideally suited for such an analysis as it allows evidence to be quantified both for and against an effect of interest, and it facilitates the consideration of theoretical constraint in the data. In the following, we will highlight four additional general recommendations for team science that facilitate a full-scale analysis.

Our first recommendation is to use all data that are available. Most directly, this means using a hierarchical model with all primary data nested in studies rather than a meta-analysis based on compressed and aggregated data. Furthermore, while participant-level exclusions may be explored (see point 2), we would advice never to apply study-level exclusions based on sample size. In particular, more data always means more statistical power and more resolution. Additionally, hierarchical shrinkage will automatically reduce the influence of outlying labs with relatively few observations by more strongly pulling these observations towards the global estimate.

Our second recommendation is to conduct a multiverse analysis (Steegen et al., 2016) to investigate the evidence across different reasonable exclusion criteria, model choices, or prior settings. As illustrated by the Many Labs 4 project, team science efforts often involve a range of reasonable options for data exclusion criteria, prior settings, and perhaps other analytic choices. In order to get a full picture of the robustness and potential relevant dimensions of the data affecting the outcomes, analysts could explore multiple analytic paths (see also: Tierney et al., in preparation, 2021). In some cases it might make sense to apply different weights to different paths of the multiverse, for instance based on theoretical or methodological grounds.

Our third recommendation is to preregister but remain open to justifiable deviations. Especially in highly complex projects with crowd-sourced data collection and many involved parties, unexpected events and deviations are the norm rather than the exception. At least in our personal experience, none of the team science projects went exactly as planned, and many required reconsideration of preregistered choices (e.g., Hoogeveen, Haaf, et al., 2022; Hoogeveen, Sarafoglou, et al., 2022; Tierney et al., in preparation, 2021). While full transparency is clearly key in these situations, we believe the quality of the eventual analysis and hence the validity of the conclusions should outweigh the strict adherence to the preregistration. Another option to ensure uncontaminated data analysis would be to use *blinded analysis* (MacCoun & Perlmutter, 2015, 2018), in which analysts perform their analysis on an altered version of the data (e.g., shuffling the dependent variable, adding noise to the data, or switching labels of categorical variables). Only after the analysts are fully satisfied with the analysis, the blind is lifted and the real data are revealed (see Dutilh, Sarafoglou, & Wagenmakers, 2019; Sarafoglou, Hoogeveen, & Wagenmakers, 2022 for more information on analysis blinding).

Our fourth recommendation is to consider collaborating with methodologists on the statistical analysis. Typically, team science efforts involve relatively extensive and complex data (e.g., hierarchically structured). We believe the time and effort put into data collection and study design also justify spending some additional time, effort, and resources on data analysis expertise. For the sake of illustration: if we imagine that each participating lab in the Many Labs 4 project invested 15 minutes per participant, this comes down to 21 labs spending about 1,589 minutes on data collection, for a total of 556 hours.[3] Given this huge investment of time and effort, the overall project quality might benefit from also matching the investment into the analysis, potentially by outsourcing the analysis to methodological and statistical experts. At least in our personal experience, experts are often eager to help out (and get their hands on "real data" for a change). For example, we have been involved in the data analysis for a couple of team science projects (e.g., Camerer et al., 2018; Tierney et al., in preparation, 2021). Having an independent analysis team may also make it easier to justify deviations from the preregistration and to apply differential weights to paths in the multiverse analysis, as either of these decisions can be made independently from the analysts.

The idea of team science efforts such as the Many Labs projects 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 in a full-scale analysis (Wagenmakers et al., 2022). A complete assessment of robustness and uncertainty therefore requires many labs, many models, perhaps many analysis paths, and ideally many collaborating experts.

Contributions

Contributorship was documented with CRediT taxonomy using tenzing (Holcombe, Kovacs, Aust, & Aczel, 2020).

Conceptualization: S.H., E.-J.W., and J.M.H.
Formal analysis: S.H., S.W.B., and J.M.H.
Funding acquisition: S.H., Q.F.G., E.-J.W., and J.M.H.

[3] This is probably an underestimation, given that Many Labs projects typically spend a great deal of time and effort on quality control and accountability standards for the data collection procedures. For instance, in Many Labs 4, participating labs were required to videotape a mock session with the experimenter and a mock participant.

Methodology: S.H., S.W.B., Q.F.G., and J.M.H.
Supervision: E.-J.W. and J.M.H.
Visualization: S.H., S.W.B., and J.M.H.
Writing - original draft: S.H. and J.M.H.
Writing - review & editing: S.W.B., Q.F.G., and E.-J.W.

Conflicts of Interest

The authors declare that there were no conflicts of interest with respect to the authorship or the publication of this article.

Acknowledgements

This research was supported in part by a Vici grant from the Netherlands Organisation for Scientific Research (NWO; 016.Vici.170.083) to E.-J.W., by an NWO Veni grant (VI.Veni.201G.019) to J.M.H., by an NWO grant to Q.F.G. (406.16.528), and by a talent grant from the Amsterdam Brain and Cognition (ABC) research platform, University of Amsterdam (T0921) to S.H.

For all analyses we used `R` (Version 4.1.2; R Core Team, 2021) and the R-packages `BayesFactor` (Morey & Rouder, 2021), `coda` (Plummer, Best, Cowles, & Vines, 2006), `cowplot` (Wilke, 2020), `dplyr` (Wickham, François, Henry, & Müller, 2022), `ggplot2` (Wickham, 2016), `kableExtra` (Zhu, 2021), `knitr` (Xie, 2015), `lemon` (Edwards, 2020), `MASS` (Venables & Ripley, 2002), `Matrix` (Bates, Maechler, & Jagan, 2022), `MCMCpack` (Martin, Quinn, & Park, 2011), `metaBMA` (Heck et al., 2019), `metafor` (Viechtbauer, 2010), `papaja` (Aust & Barth, 2022), `Rcpp` (Eddelbuettel & Balamuta, 2018; Eddelbuettel & François, 2011) and `tinylabels` (Barth, 2021).

References

Aitkin, M. (1991). Posterior Bayes Factors. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(1), 111–142.

Aust, F., & Barth, M. (2022). papaja: Prepare reproducible APA journal articles with R Markdown [Computer software manual]. Retrieved from https://github.com/crsh/papaja (R package version 0.1.1)

Barth, M. (2021). tinylabels: Lightweight variable labels [Computer software manual]. Retrieved from https://cran.r-project.org/package=tinylabels (R package version 0.2.2)

Bates, D., Maechler, M., & Jagan, M. (2022). Matrix: Sparse and dense matrix classes and methods [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=Matrix (R package version 1.4-1)

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating replicability of social science experiments in *Nature* and *Science*. *Nature Human Behaviour*, *2*, 637–644.

Chartier, C., Kline, M., McCarthy, R., Nuijten, M., Dunleavy, D. J., & Ledgerwood, A. (2018, November). The cooperative revolution is making psychological science better. *APS Observer*, *31*(10).

Chatard, A., Hirschberger, G., & Pyszczynski, T. (2020, February). *A word of caution about Many Labs 4: If you fail to follow your preregistered plan, you may fail to find a real effect* [Preprint]. PsyArXiv. doi: 10.31234/osf.io/ejubn

Chen, S.-C., Szabelska, A., Chartier, C. R., Kekecs, Z., Lynott, D., Bernabeu, P., … Schmidt, K. (2018, November). Investigating object orientation effects across 14 languages.
doi: 10.31234/osf.io/t2pjv

DeHaven, A. C. (2017). *Preregistration: A plan, not a prison.*

Dutilh, G., Sarafoglou, A., & Wagenmakers, E.-J. (2019). Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*, 1–28. doi: 10.1007/s11229-019-02456-7

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

Eddelbuettel, D., & Balamuta, J. J. (2018). Extending extitR with extitC++: A Brief Introduction to extitRcpp. *The American Statistician*, *72*(1), 28-36. doi: 10.1080/00031305.2017.1375990

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal* *of Statistical Software*, *40*(8), 1–18. doi: 10.18637/jss.v040.i08

Edwards, S. M. (2020). lemon: Freshing up your 'ggplot2' plots [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=lemon (R package version 0.4.5)

Efron, B., & Morris, C. (1977). Stein's Paradox in Statistics. *Scientific American*, *236*, 119–127. doi: 10.1038/scientificamerican0577-119

Forscher, P. S., Wagenmakers, E.-J., Coles, N. A., Silan, M. A., Dutra, N. B., Basnight-Brown, D., & IJzerman, H. (in press). The Benefits, Barriers, and Risks of Big Team Science. *Perspectives on Psychological Science*. doi: 10.31234/osf.io/2mdxh

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435. doi: 10.1111/infa.12182

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian Data Analysis (2nd ed.)*. Boca Raton (FL): Chapman & Hall/CRC.

Gervais, W. M., Xygalatas, D., McKay, R. T., van Elk, M., Buchtel, E. E., Aveyard, M., ... Bulbulia, J. (2017). Global evidence of extreme intuitive moral prejudice against atheists. *Nature Human Behaviour*, *1*(8), 0151. doi: 10.1038/s41562-017-0151

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. doi: 10.1037/0022-3514.67.4.627

Gronau, Q. F., Heck, D. W., Berkhout, S. W., Haaf, J. M., & Wagenmakers, E.-J. (2021, July). A primer on Bayesian model-averaged meta-analysis. *Advances in Methods and Practices in Psychological Science*, *4*(3), 25152459211031256. doi: 10.1177/25152459211031256

Gronau, Q. F., Ly, A., & Wagenmakers, E.-J. (2019). Informed Bayesian t-tests. *The American Statistician*, 1–13. doi: 10.1080/00031305.2018.1562983

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. doi: 10.31222/osf.io/heamz

Haaf, J. M., Klaassen, F., & Rouder, J. N. (2018). Capturing Ordinal Theoretical Constraint in Psychological Science. *PsyArXiv*. doi: 10.31234/osf.io/a4xu9

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models. *Psychological Methods*, *22*(4), 779–798. doi: 10.31234/osf.io/ktjnq

Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. (2021). *Doing Meta-Analysis with R: A Hands-On Guide*. New York: Chapman and Hall/CRC. doi: 10.1201/9781003107347

Heck, W., D., Gronau, F., Q., Wagenmakers, ., & E.-J. (2019). metabma: Bayesian model averaging for random and fixed effects meta-analysis [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=metaBMA

Heck, D. W., & Gronau, Q. F. (2017). metaBMA: Bayesian model averaging for random- and fixed-effects meta-analysis [R Package].

Henrich, J., Heine, S. J., & Norenzayan, A. (2010, June). The weirdest people in the world? *Behavioral and Brain Sciences*, *33*(2-3), 61–83. doi: 10.1017/S0140525X0999152X

Hinne, M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020, June). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science*, *3*(2), 200–215. doi: 10.1177/2515245919898657

Holcombe, A. O., Kovacs, M., Aust, F., & Aczel, B. (2020). Documenting contributions to scholarly articles using CRediT and tenzing. *PLoS One*, *15*(12), e0244611.

Hoogeveen, S., Haaf, J. M., Bulbulia, J. A., Ross, R. M., McKay, R., Altay, S., . . . van Elk, M. (2022). The Einstein effect provides global evidence for scientific source credibility effects and the influence of religiosity. *Nature Human Behaviour*, *6*(4), 523–535. doi: 10.1038/s41562-021-01273-8

Hoogeveen, S., Sarafoglou, A., Aczel, B., Aditya, Y., Alayan, A., Allen, P., . . . Wagenmakers, E.-J. (2022). A many-analysts approach to the relation between religiosity and well-being. *Religion, Brain & Behavior*. doi: 10.31234/osf.io/pbfye

Jeffreys, H. (1939). *Theory of Probability* (First ed.). Oxford, UK: Oxford University Press.

Jones, B. C., DeBruine, L. M., Flake, J. K., Liuzza, M. T., Antfolk, J., Arinze, N. C., . . . Coles, N. A. (2021, January). To which world regions does the valence–dominance model of social perception apply? *Nature Human Behaviour*, *5*(1), 159–169. doi: 10.1038/s41562-020-01007-2

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Journal of the American Statistical Association*, *90*, 773–795. doi: 10.2307/2291091

Keysers, C., Gazzola, V., & Wagenmakers, E.-J. (2020, July). Using Bayes factor hypothesis testing in neuroscience to establish evidence of absence. *Nature Neuroscience*, *23*(7), 788–799. doi: 10.1038/s41593-020-0660-4

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Hilgard, J., . . . Ratliff, K. A. (2022). Many labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra: Psychology*, *8*(1), 35271. doi: 10.1525/collabra.35271

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C. A., Nosek, B. A., Chartier, C. R.,

… Ratliff, K. A. (2019, December). *Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement* [Preprint]. PsyArXiv. doi: 10.31234/osf.io/vef2c

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahńik, S., Bernstein, M., … Nosek, B. (2014). Investigating variation in replicability: A “Many Labs” replication project. *Social Psychology*, *45*, 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. (2018). Many Labs 2: Investigating variation in replicability across sample and setting. *Advances in Methods and Practices in Psychological Science*, *1*, 443–490.

Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. doi: 10.1037/a0029146

Landy, J. F., Jia, M. L., Ding, I. L., Viganola, D., Tierney, W., Dreber, A., … Uhlmann, E. L. (2020). Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychological Bulletin*, *146*(5), 451–479. doi: 10.1037/bul0000220

MacCoun, R., & Perlmutter, S. (2015). Hide results to seek the truth: More fields should, like particle physics, adopt blind analysis to thwart bias. *Nature*, *526*, 187–189.

MacCoun, R., & Perlmutter, S. (2018). Blind analysis as a correction for confirmatory bias in physics and in psychology. In S. O. Lilienfeld & I. Waldman (Eds.), *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions.* John Wiley and Sons.

Martin, A. D., Quinn, K. M., & Park, J. H. (2011). MCMCpack: Markov chain monte carlo in R. *Journal of Statistical Software*, *42*(9), 22. doi: 10.18637/jss.v042.i09

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes factors for common designs.*

Morey, R. D., & Rouder, J. N. (2021). Bayesfactor: Computation of bayes factors for common designs [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=BayesFactor (R package version 0.9.12-4.3)

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., … Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, *1*, 501–515. doi: 10.1177/2515245918797607

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

Plummer, M., Best, N., Cowles, K., & Vines, K. (2006). Coda: Convergence diagnosis and output analysis for mcmc. *R News*, *6*(1), 7–11. Retrieved from https://journal.r

-project.org/archive/

R Core Team. (2021). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project.org/

Rouder, J. N., Haaf, J. M., & Aust, F. (2018, January). From theories to models to predictions: A Bayesian model comparison approach. *Communication Monographs*, *85*(1), 41–56. doi: 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*, *24*, 606–621. doi: 10.1037/met0000216

Rouder, J. N., & Province, J. M. (2019). Bayesian Hierarchical Models in Psychological Science: A Tutorial. In *New Methods in Cognitive Psychology.* Routledge.

Sarafoglou, A., Hoogeveen, S., & Wagenmakers, E.-J. (2022). Comparing analysis blinding with preregistration in the many-analysts religion project. *Advances in Methods and Practices in Psychological Science.* doi: 10.31234/osf.io/6dn8f

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.

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., ... Uhlmann, E. L. (2016, September). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55–67. doi: 10.1016/j.jesp.2015.10.001

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

The ManyBabies Consortium. (2020, March). Quantifying sources of variability in infancy research using the infant-directed-speech preference:. *Advances in Methods and Practices in Psychological Science.* doi: 10.1177/2515245919900809

Tierney, W., Cyrus-Lai, W., Hoogeveen, S., Haaf, J. M., Landy, J. F., Hardy, J., III, ... Uhlmann, E. L. (in preparation). *Who respects an angry woman? A pre-registered re-examination of the relationships between gender, emotion expression, and status conferral.* [Unpublished Manuscript].

Tierney, W., Hardy, J., III, Ebersole, C., Viganola, D., Clemente, E., Gordon, E., ... Uhlmann, E. L. (2021). A creative destruction approach to replication: Implicit work and sex morality across cultures. *Journal of Experimental Social Psychology*, *93*, 104060.

Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., ... Nosek, B. A. (2019, September). Scientific Utopia III: Crowdsourcing science. *Perspectives on Psychological Science*, *14*(5), 711–733. doi: 10.1177/1745691619850561

van Erp, S., Verhagen, 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*.

Vanpaemel, W. (2010, December). Prior sensitivity in theory testing: An apologia for the Bayes factor. *Journal of Mathematical Psychology*, *54*(6), 491–498. doi: 10.1016/j.jmp.2010.07.003

Veenman, M., Stefan, A., & Haaf, J. M. (2022). *Bayesian hierarchical modeling: An introduction and reassessment.* PsyArXiv.

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with s* (Fourth ed.). New York: Springer. Retrieved from https://www.stats.ox.ac.uk/pub/MASS4/ (ISBN 0-387-95457-0)

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. Retrieved from https://doi.org/10.18637/jss.v036.i03

Vohs, K. D., Schmeichel, B. J., Lohmann, S., Gronau, Q. F., Finley, A. J., Ainsworth, S. E., ... Albarracín, D. (2021, October). A multisite preregistered paradigmatic test of the ego-depletion effect. *Psychological Science*, *32*(10), 1566–1581. doi: 10.1177/0956797621989733

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57. doi: 10.3758/s13423-017-1343-3

Wagenmakers, E.-J., Sarafoglou, A., & Aczel, B. (2022). One statistical analysis must not rule them all. *Nature*, *605*(7910), 423–425. doi: 10.1038/d41586-022-01332-8

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis.* Springer-Verlag New York. Retrieved from https://ggplot2.tidyverse.org

Wickham, H., François, R., Henry, L., & Müller, K. (2022). dplyr: A grammar of data manipulation [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=dplyr (R package version 1.0.10)

Wilke, C. O. (2020). cowplot: Streamlined plot theme and plot annotations for 'ggplot2' [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=cowplot (R package version 1.1.1)

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida:

Chapman and Hall/CRC. Retrieved from https://yihui.org/knitr/ (ISBN 978-1498716963)

Zhu, H. (2021). kableextra: Construct complex table with 'kable' and pipe syntax [Computer software manual]. Retrieved from https://CRAN.R-project.org/package=kableExtra (R package version 1.3.4)

Appendix A: Full Sets of Exclusion Criteria

Both Klein et al. (2022) 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 = 2,225).
2. In addition to exclusion criterion 1, exclude participants who do not self-identify as white and/or who report not to be born in the United States (leaving N = 1,880).[4]
3. In addition to exclusion criteria 1 and 2, exclude participants who responded below 7 on the 9-point American Identity item (leaving N = 1,699).

Note that both Klein et al. (2022) and Chatard et al. (2020) applied the participant-level exclusion criteria *only* to the author-advised protocols, which means that for exclusion criteria 2 and 3 all participants from the in-house labs where retained. The reason for this approach is that for many in-house labs, the necessary information to apply these criteria was often unavailable. However, this choice indicates that the authors implicitly assumed that all in-house participants were white, born in the US, and strongly identified with American culture. Data from the author-advised labs nonetheless showed that only 56.5% of participants self-identified as white and were born in the US (i.e., exclusion criterion 2) and only 33.8% were white, born in the US, and strongly identified with American culture (i.e., exclusion criterion 3). Since exclusion criteria 2 and 3 were specified by the original authors as a strict and genuine test of the theory, we believe that it is important to thoroughly apply these criteria to all participants, even if this means discarding participants where this information is unavailable. In addition, for some of the in-house labs, information on ethnicity and country of birth was in fact available, so retaining participants who do not meet the criterion seems hard to defend. Based on these considerations, we included another layer to the exclusion constellations related to the application of the participant-level exclusion criteria:

[4] 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.

1. Apply participant-level criteria to author-advised labs only (retaining all in-house participants)
2. Apply participant-level criteria to both author-advised labs and in-house labs (discarding all missing values on the relevant variables).

In addition to the 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. (2022): "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). While the authors had not applied any N-based exclusion criteria in the preprint, they did so in the published version (Klein et al., 2022). 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.

Moreover, 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 $K = 9$).

Lastly, in the published version, Klein et al. (2022) added an another layer of exclusion settings related to the timing of the data collection. As some In-House labs had started data collection before the lead team's analysis plan was finalized, Klein et al. (2022) decided to discard all observations collected prior to the preregistration date (February 15th, 2017). This resulted in the exclusion of 566 participants (25.4%). We consider this exclusion wasteful and unnecessary, since it concerned only In-House studies that were free to design their own protocols. Moreover, the lead team had not accessed or inspected the data collected by the different labs prior to their preregistration, so the analysis plan could not have been contaminated by the already-collected data.[5] For the sake of completeness, we included the timing-based exclusion setting as another layer in the multiverse analysis:

1. Include data collected anytime.
2. Exclude data collected before the lead team's analysis plan was preregistered (i.e., February 15th, 2017).

These 5 layers of exclusion settings result in $3 \times 3 \times 2 \times 2 \times 2 = 72$ constellations of exclusion criteria. Note that some of the 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 45 instead of 72 unique constellations. Table 1 shows all 45 unique constellations, the resulting number of studies and total number of included participants (see Appendix C for a table with all 72 constellations).

In the published article, Klein et al. (2022) based their main conclusions on three of these constellations (orange rows in Table 1): Including studies from labs with more than 60 participants, excluding data that was collected prior to the preregistration date, including both author-advised and in-house protocols, varying the participant-level exclusion criteria while applying these only to participants from author-advised labs.[6] 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 (green rows): Excluding studies with fewer than 40 participants per condition ($N > 80$), excluding In-House studies, including data collected at any time, with varying participant-level exclusion criteria, applied to only the author-advised

[5] Note that the timing-based criteria and the application of the participant-level criteria are actually irrelevant for the key analyses by Chatard et al. (2020), as they only affect in-house protocols which the authors excluded anyways.

[6] We note that the preprinted version by Klein et al. (2019) adopted different exclusion criteria; no timing-based exclusions were applied nor any study-level exclusions.

protocols. The purple rows correspond to our own choice of analysis paths. Specifically, we included all complete data, from all labs and protocols and applied the participant-level exclusions to both author-advised labs and in-house labs, discarding missing values.

Appendix B: Bayesian Model-averaged Meta-analysis

Here, we report the results of an alternative Bayesian analysis for team science projects data: a Bayesian model-averaged meta-analysis (Gronau et al., 2021, 2017). For the 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.

Methods

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, 2020) 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 et al. (2021) 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[7]), and (3) a normal distribution with mean 0.3 and standard deviation 0.15 (*Vohs* prior[8]). 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. (2020). The Bayesian model-averaged meta-analyses are conducted using the R-package `metaBMA` (Heck & Gronau, 2017).

Results

Model-averaged Meta-analysis of Klein et al.

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 sample under participant-level exclusion criterion 1 the overall effect size is estimated as $0.06,\ 95\%\text{CI} = [-0.06, 0.18]$; for participant-level exclusion criterion 2 the overall effect size is estimated as $0.09,\ 95\%\text{CI} = [-0.05, 0.22]$; and for participant-level exclusion criterion 3 the overall effect size is estimated as $0.08,\ 95\%\text{CI} = [-0.06, 0.23]$. The most consistent pattern is that the credible interval slightly 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

[7] This *Oosterwijk* prior has been elicited for a reanalysis of a social psychology study (Gronau et al., 2017), but we believe it is a reasonable prior for many psychological studies more generally.

[8] This *Vohs* prior has been specified by ego depletion experts to analyze ego depletion replication studies (Vohs et al., 2021).

Table 4

Model-averaged Bayes factors for key analyses.

			Effect BF_{01}			Heterogeneity BF_{01}
Participant-level	N	Labs	Default	Oosterwijk	Vohs	Default
Klein et al. (2022)						
All	1544	17	4.45	10.71	4.18	1.89
White & US-born	1223	17	2.79	5.02	2.14	1.45
US-Identity > 7	1070	17	3.34	5.90	2.57	1.33
Chatard et al. (2020)						
All	699	7	4.04	6.36	2.97	2.63
White & US-born	378	7	1.43	0.90	0.66	2.06
US-Identity > 7	225	7	1.44	0.72	0.62	1.88
Current choice						
All	2211	21	12.60	44.69	16.64	2.28
White & US-born	983	16	19.42	67.73	25.90	2.03
US-Identity > 7	272	9	4.13	3.90	2.44	1.79

Note. All Bayes factors are reported in favor of the null model. The different column names for the effect BF_{01} refer to the different priors used.

model comparison. The Bayes factors for the key analyses from Klein et al. (2022) are shown in the top three rows of Table 4. 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 10.71-to-1 to 2.14-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 overall effect size estimates from a two-sided model.

Model-averaged Meta-analysis of Chatard et al.

For the reanalysis of the key findings of Chatard et al. (2020) 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 $0.08,\ 95\%\text{CI} = [-0.10, 0.25]$; for exclusion criterion 2 the overall effect size is estimated as $0.16,\ 95\%\text{CI} = [-0.08, 0.41]$; and for exclusion criterion 3 the overall effect size is estimated as $0.18,\ 95\%\text{CI} = [-0.10, 0.47]$. While the point estimates are considerably larger than the ones when all protocols 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 699 and 225 participants.

To quantify the absence or presence of an effect we again computed model-averaged Bayes factors. These are shown in the middle three rows of Table 4. 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 4 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.

Model-averaged Meta-analysis of our own choice

For our own choice of exclusion criteria settings, we also conducted a model-averaged meta-analysis. That is, we estimated the overall effect size across studies (Hedges' g) for the three participant-level exclusion criteria applied to both author-advised labs and in-house labs –discarding missing values–, while including all complete data, from all labs and protocols.

Effect size estimates from the random effects model with the default prior are 0.03, 95% CI $[-0.07, 0.13]$ for participant-level exclusion criterion 1, -0.03, 95% CI $[-0.17, 0.13]$ for participant-level exclusion criterion 2, and 0.07, 95% CI $[-0.20, 0.34]$ for participant-level exclusion criterion 3. Again, all credible intervals overlap with zero, and the width of the credible interval increases with fewer observations included in the analysis, as less data implies more uncertainty.

Similar to the analysis of Klein et al.'s key findings, the Bayes factors indicate evidence against the mortality salience effect. All analyses across participant-level exclusions and prior choices provide evidence against an overall effect, with Bayes factors ranging from 67.73-to-1 to 2.44-to-1 in favor of the null model. For exclusion criteria 1 and 2, the data strongly prefer the null model across all prior setting. Mirroring the Klein et al. analysis, the strongest evidence against the effect is obtained under the most optimistic prior, namely the Oosterwijk prior. These Bayes factors for the heterogeneity between studies again reflect that there is some evidence against study heterogeneity.

Model-averaged Meta-analysis Multiverse

As with the hierarchical analysis, we also conducted a multiverse analysis across all unique data exclusion constellations using the Bayesian model-averaged meta-analytic approach. The analysis is conducted using the three different prior distributions, the default prior, the Oosterwijk prior, and the Vohs prior. The results of this analysis are shown in Figure 6. The Bayes factors for the mortality salience effect are plotted on the y-axis and the Bayes factors for between-study heterogeneity are plotted on the x-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 size of the point reflects the number of participants included in the analysis. The majority of Bayes factors are in line with the absence of the mortality salience effect and all Bayes factors indicate evidence against heterogeneity.

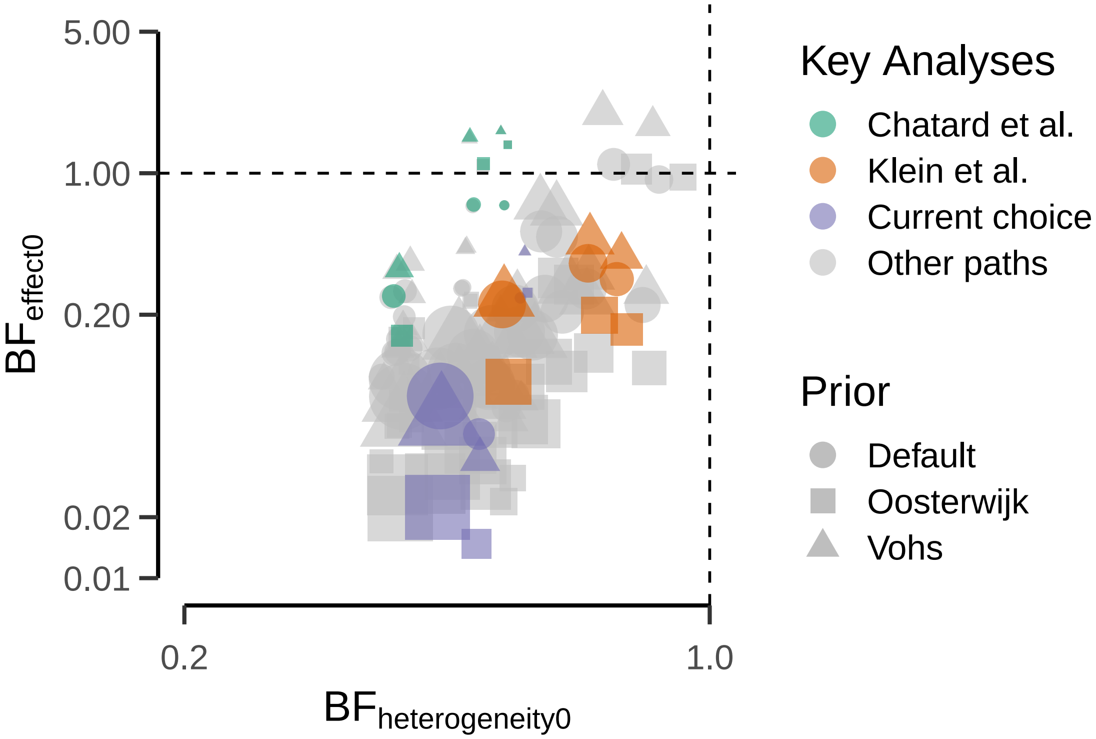

Figure 6. Results from the multiverse analysis for the model-averaged meta-analysis: Bayes factors in favor of a mortality salience effect are above the horizontal line, Bayes factors against the mortality salience effect are below the horizontal line. All analyses provide evidence against between-study heterogeneity as shown by all heterogeneity Bayes factors are smaller than 1 on the x-axis. The color of the points refers to the different key analyses sets, the shape of the points refers the different prior settings in the meta-analysis, and the size of the points refers to the number of participants the analysis is based on. The majority of analyses provide evidence against the mortality salience effect.

Because the Bayes factor depends on the sample size, as a general trend, more evidence against morality salience comes from analyses that are based on more data (i.e., larger number of included participants). To inspect the effect of prior settings one can view the pattern of the different shapes in Figure 6. Remember that the default prior is the most vague prior and the Oosterwijk prior is more optimistic than the Vohs prior. For most data sets Bayes factors are larger for more optimistic priors because evidence against optimistic and informed models accumulates faster when comparing to a null model. A final pattern that emerges from Figure 6 that was not observed in the hierarchical analysis, is the positive correlation between evidence for the effect and for heterogeneity; analyses that provide stronger evidence against the mortality salience effect also provide stronger evidence against between-study heterogeneity, and vice versa.

Appendix C: Preregistration

We will conduct a re-analysis of the Many Labs 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 Many Labs 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 ≥ 60 participants are included. This is the preregistered exclusion criterion.

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

4. Only studies with ≥ 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, Ly, & Wagenmakers, 2019), 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.

Appendix D: Full table of data exclusion constellations

Table 5
Exclusion constellations and resulting sample sizes

Participant-level	N-based	Protocol	Timing-based	Apply P-based	Sample Size	Labs
All	All	All	All	AA only	2225	21
White & US-born	All	All	All	AA only	1880	21
US-Identity > 7	All	All	All	AA only	1699	21
All	N > 60	All	All	AA only	2067	17
White & US-born	N > 60	All	All	AA only	1746	17
US-Identity > 7	N > 60	All	All	AA only	1593	17
All	N > 80	All	All	AA only	1866	14
White & US-born	N > 80	All	All	AA only	1545	14
US-Identity > 7	N > 80	All	All	AA only	1392	14
All	All	AA	All	AA only	798	9
White & US-born	All	AA	All	AA only	453	9
US-Identity > 7	All	AA	All	AA only	272	9
All	N > 60	AA	All	AA only	699	7
White & US-born	N > 60	AA	All	AA only	378	7

Continued on next page

Table 5 continued

Participant-level	N-based	Protocol	Timing-based	Apply P-based	Sample Size	Labs
US-Identity > 7	N > 60	AA	All	AA only	225	7
All	N > 80	AA	All	AA only	699	7
White & US-born	N > 80	AA	All	AA only	378	7
US-Identity > 7	N > 80	AA	All	AA only	225	7
All	All	All	After prereg	AA only	1659	20
White & US-born	All	All	After prereg	AA only	1314	20
US-Identity > 7	All	All	After prereg	AA only	1133	20
All	N > 60	All	After prereg	AA only	1544	17
White & US-born	N > 60	All	After prereg	AA only	1223	17
US-Identity > 7	N > 60	All	After prereg	AA only	1070	17
All	N > 80	All	After prereg	AA only	1343	14
White & US-born	N > 80	All	After prereg	AA only	1022	14
US-Identity > 7	N > 80	All	After prereg	AA only	869	14
All	All	AA	After prereg	AA only	797	9
White & US-born	All	AA	After prereg	AA only	452	9
US-Identity > 7	All	AA	After prereg	AA only	271	9
All	N > 60	AA	After prereg	AA only	698	7
White & US-born	N > 60	AA	After prereg	AA only	377	7
US-Identity > 7	N > 60	AA	After prereg	AA only	224	7
All	N > 80	AA	After prereg	AA only	698	7
White & US-born	N > 80	AA	After prereg	AA only	377	7
US-Identity > 7	N > 80	AA	After prereg	AA only	224	7
All	All	All	All	AA and IH	2211	21
White & US-born	All	All	All	AA and IH	983	16
US-Identity > 7	All	All	All	AA and IH	272	9
All	N > 60	All	All	AA and IH	2053	17
White & US-born	N > 60	All	All	AA and IH	897	13
US-Identity > 7	N > 60	All	All	AA and IH	225	7
All	N > 80	All	All	AA and IH	1852	14
White & US-born	N > 80	All	All	AA and IH	864	12
US-Identity > 7	N > 80	All	All	AA and IH	225	7
All	All	AA	All	AA and IH	799	9
White & US-born	All	AA	All	AA and IH	453	9
US-Identity > 7	All	AA	All	AA and IH	272	9
All	N > 60	AA	All	AA and IH	700	7
White & US-born	N > 60	AA	All	AA and IH	378	7
US-Identity > 7	N > 60	AA	All	AA and IH	225	7

Continued on next page

Table 5 continued

Participant-level	N-based	Protocol	Timing-based	Apply P-based	Sample Size	Labs
All	N > 80	AA	All	AA and IH	700	7
White & US-born	N > 80	AA	All	AA and IH	378	7
US-Identity > 7	N > 80	AA	All	AA and IH	225	7
All	All	All	After prereg	AA and IH	1650	20
White & US-born	All	All	After prereg	AA and IH	777	15
US-Identity > 7	All	All	After prereg	AA and IH	271	9
All	N > 60	All	After prereg	AA and IH	1535	17
White & US-born	N > 60	All	After prereg	AA and IH	702	13
US-Identity > 7	N > 60	All	After prereg	AA and IH	224	7
All	N > 80	All	After prereg	AA and IH	1334	14
White & US-born	N > 80	All	After prereg	AA and IH	669	12
US-Identity > 7	N > 80	All	After prereg	AA and IH	224	7
All	All	AA	After prereg	AA and IH	798	9
White & US-born	All	AA	After prereg	AA and IH	452	9
US-Identity > 7	All	AA	After prereg	AA and IH	271	9
All	N > 60	AA	After prereg	AA and IH	699	7
White & US-born	N > 60	AA	After prereg	AA and IH	377	7
US-Identity > 7	N > 60	AA	After prereg	AA and IH	224	7
All	N > 80	AA	After prereg	AA and IH	699	7
White & US-born	N > 80	AA	After prereg	AA and IH	377	7
US-Identity > 7	N > 80	AA	After prereg	AA and IH	224	7

Note. Orange rows refer to Klein et al.'s key analyses; green rows refer to Chatard et al.'s key analyses; purple rows refer to our chosen analyses; grey rows are repeated data sets and not included in the multiverse analysis; AA = author-advised. 'Application P-based' indicates whether the participant-level exclusion criteria are applied to the author-advised labs only (retaining all in-house participants) or to both author-advised and in-house labs (missing data excluded).