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

Suzanne Hoogeveen[1], Sophie W. Berkhout[2], Quentin F. Gronau[1], Eric-Jan Wagenmakers[1], & Julia M. Haaf[1]

[1] University of Amsterdam

[2] Utrecht University

Author Note

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 concerning this article should be addressed to Suzanne Hoogeveen, Nieuwe Achtergracht 129B, 1018 WT Amsterdam, The Netherlands. E-mail: suzanne.j.hoogeveen@gmail.como

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

Word count: X

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

Model

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

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

where μ is the grand mean, α_i is the ith lab's specific overall 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: $\theta_i \sim N(\mu_\theta, \sigma^2_\theta)$.

Reanalysis

We just run the analyses for all *unique* dataset based on the 72 exclusion criteria constellations. Let's look at the table that shows all possibilities.

```
## [1] 72
```

Table 1
Exclusion constellations and resulting sample sizes

Participant-level	N-based	Protocol	Timing-based	Application P-based	Sample Size	N Studies
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
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

Table 1 continued

Participant-level	N-based	Protocol	Timing-based	Application P-based	Sample Size	N Studies
White & US-born	N > 60	AA	All	AA and IH	378	7
US-Identity > 7	N > 60	AA	All	AA and IH	225	7
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 AA-labs only (retaining all IH-participants) or to both AA- and IH-labs (missing data excluded).

Table 2

Unique Exclusion constellations and resulting sample sizes

Participant-level	N-based	Protocol	Timing-based	Application P-based	Sample Size	N Studies
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

Table 2 continued

Participant-level	N-based	Protocol	Timing-based	Application P-based	Sample Size	N Studies
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. 'Application P-based' indicates whether the participant-level exclusion criteria are applied to the AA-labs only (retaining all IH-participants) or to both AA- and IH-labs (missing data excluded).

Reanalysis with Exclusion Criterion .1.1.1.2

This is our all-inclusive analysis. It includes data from all participants that have completed the relevant measures, all studies, and applied the participant-level exclusion criteria to both Author-Advised (AA) protocols and In-House (IH) protocols. Note that for many IH protocols, the relevant information for participant-level exclusion criteria 3 (and 2 to a lesser degree) is missing. We decided to exclude participants for whom nationality and country of origin (exclusion criterion 2) or identification with American culture (exclusion criterion 3) is unknown as we cannot assume that people met this requirement. For exclusion criterion 1 – completeness of the measures –, we did retain participants for labs where no information was available, as long as they were assigned to an experimental condition and answered both items of the dependent variable.

Reanalysis with Exclusion Criterion .2.1.2.1

This is the original 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, only studies that featured more than 60 observations (before participant-level exclusions). The participant-level exclusion criteria were only applied to AA-studies, which means that for exclusion criteria 2 and 3 all participants from the IH-studies were retained, indicating that the authors implicitly assumed they were all American, born in the US, and strongly identified with American culture. Exclusion of the observations collected by labs prior to the lead team's preregistration was uploaded caused the authors to discard 566 observations (25.44%). Note that the timing-based and the study-level the N-based exclusions are applied in the published article but not in the preprint that appeared in 2019.

```
##          F0          10          P0    post.pos   prior.pos
## 0.080371832 0.186834210 0.001590784 0.001694858 0.085630000
```

```
## [1] 0.02258143
```

```
##       2.5%      97.5%
## -0.1157235  0.1574719
```

```
##          F0          I0          P0    post.pos   prior.pos
## 0.106959244 0.237408530 0.004888851 0.003898173 0.085285000
```

```
## [1] 0.04001041
```

```
##       2.5%      97.5%
## -0.1105811  0.1888211
```

```
##          F0          I0          P0    post.pos   prior.pos
## 0.143968205 0.222373555 0.006348483 0.003796481 0.086095000
```

```
## [1] 0.04620923
```

```
##       2.5%      97.5%
## -0.1132737  0.2086956
```

Reanalysis with Exclusion Criterion .3.2.1.1

(1,3,2,1,1), (2,3,2,1,1), and (3,3,2,1,1) from the comment by Chatard, Hirschberger, and Pyszczynski (2020). The authors argued that a valid test of the theory as formulated by the original authors would only include the AA-studies. They additionally read the preregistration as stating that only labs that collected data from at least 80 participants would be included in the analysis. Following Klein et al. (2022), they applied the participant-level exclusion criteria only the AA-studies, although in this case that does not matter as all IH-data is excluded anyway. No timing-based exclusion criteria were applied.

```
##          F0          I0          P0   post.pos  prior.pos
```

```
## 0.06845900 0.46340624 0.07507517 0.13467340 0.12280500
```

```
## [1] 0.07816388
```

```
##       2.5%      97.5%
## -0.1228947  0.2803502
```

```
##        F0        10        P0  post.pos prior.pos
## 0.1479266 1.0560503 0.3825679 0.3154130 0.1219600
```

```
## [1] 0.1398324
```

```
##       2.5%      97.5%
## -0.1095187  0.3941900
```

```
##        F0        10        P0  post.pos prior.pos
## 0.2206676 1.2942052 0.6968107 0.3832073 0.1213550
```

```
## [1] 0.1791353
```

```
##       2.5%      97.5%
## -0.1149615  0.4867350
```

Summary

We want to create a figure that shows the evidence against heterogeneity on the x-axis and evidence against the effect on the y-axis. The size of the points will reflect N. The effect-evidence will be reflected by a weighted average (model average) of the common effect and unconstrained effect model. The heterogeneity-evidence will be simply the evidence for the fixed model vs. the unconstrained model.

Add arrows to the figures with different prior settings to reflect the general trend relative to the main results.

Table 3

Bayes factors for key analyses.

			Evidence			
Participant-level	Sample size	Labs	BF_{0f}	BF_{01}	BF_{0+}	Effect [95% CI]
Klein et al. (2022)						
All	1544	17	12.50	5.41	791.66	0.02 [-0.11, 0.15]
White & US-born	1223	17	9.32	4.25	241.65	0.04 [-0.11, 0.19]
US-Identity > 7	1070	17	7.07	4.56	159.02	0.05 [-0.11, 0.21]
Chatard et al. (2020)						
All	699	7	14.97	2.17	13.93	0.08 [-0.13, 0.28]
White & US-born	378	7	6.76	0.96	2.75	0.14 [-0.11, 0.39]
US-Identity > 7	225	7	4.43	0.78	1.40	0.18 [-0.11, 0.49]
Current choice						
All	2211	21	35.75	10.28	12,127.13	0.01 [-0.11, 0.12]
White & US-born	983	16	21.30	13.88	3,633.28	-0.04 [-0.20, 0.12]
US-Identity > 7	272	9	8.43	2.71	11.33	0.05 [-0.22, 0.32]

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

```
## [1] 2

## [1] 12

## [1] 1.111111

## [1] 6.666667

## [1] 7

## [1] 15.55556

## [1] 0

## [1] -0.03781871  0.18203340
```

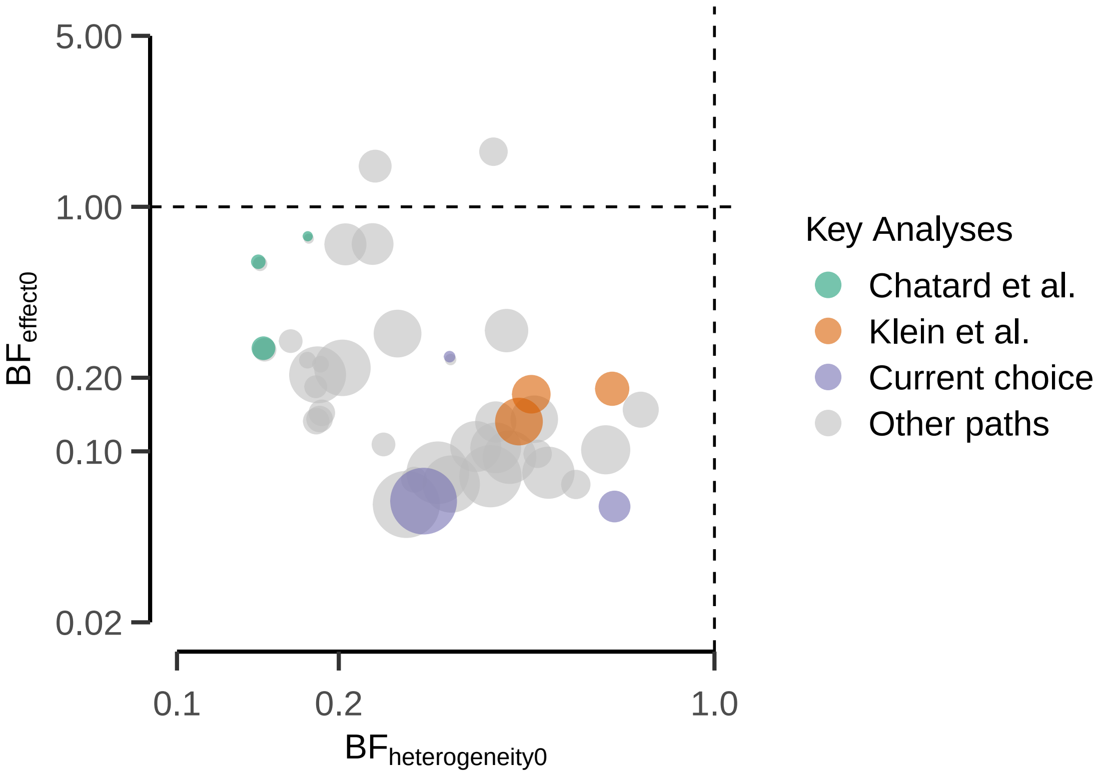

Figure 1. Results from the Bayesian multiverse analysis: Bayes factors in favor of a mortality salience effect are above the horizontal line, Bayes factors against the mortality salience effect are below the horizontal line. All analyses provide evidence against between-study heterogeneity as shown by 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. The majority of analyses provide evidence against the mortality salience effect.

Bayesian Model-average Meta-analysis

```
## [1] 0.078601859 0.004986638 0.182665408
## [1] 0.099819623 0.008485666 0.225519961
## [1] 0.100912387 0.006691452 0.234720789
## [1] 0.107026501 0.006121433 0.263727820
## [1] 0.18280208 0.01880661 0.41231070
## [1] 0.21381888 0.01837633 0.48534839
## [1] 0.051917612 0.002747982 0.138746209
## [1] 0.051835787 0.001392296 0.160473027
```

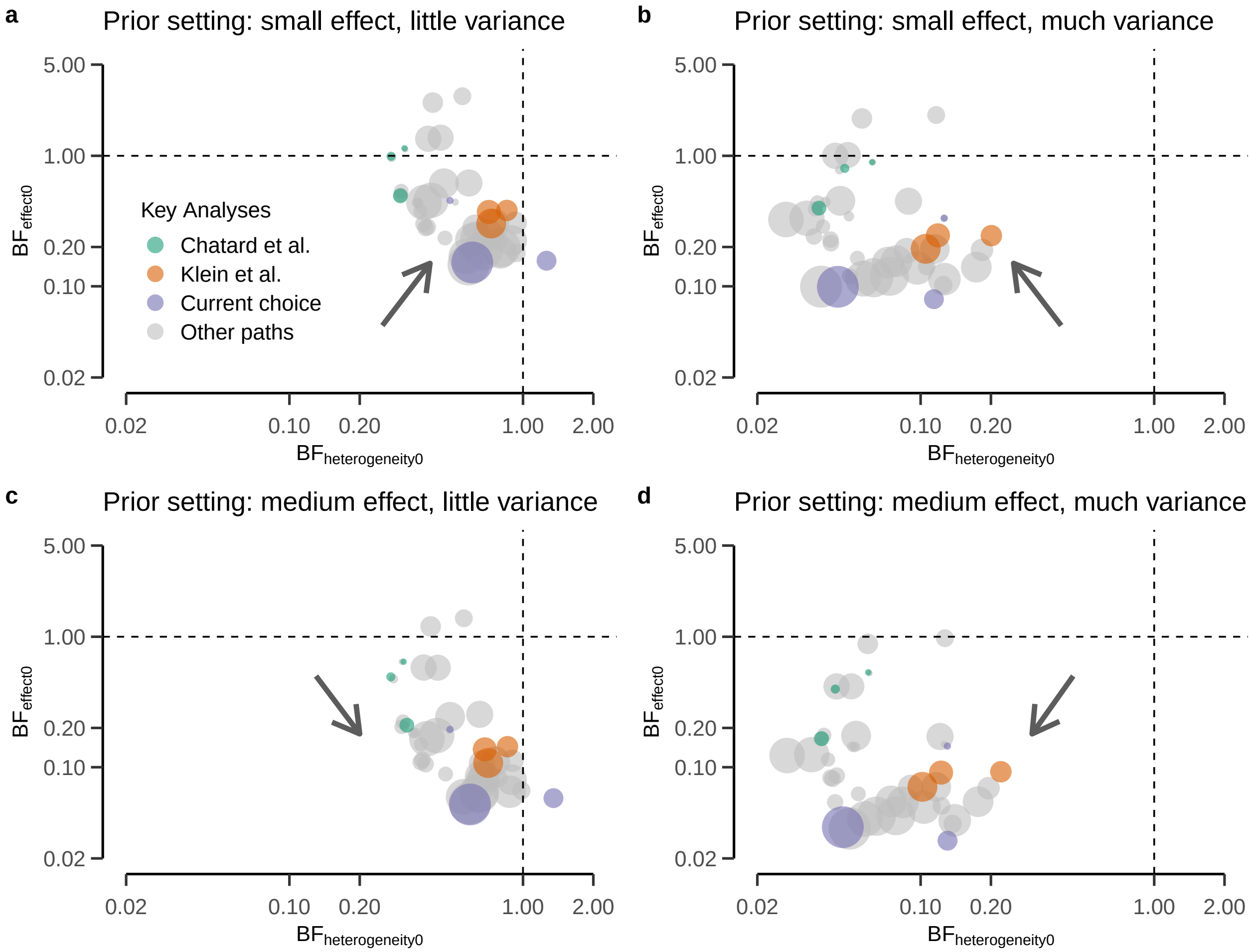

Figure 2. Results from the Bayesian multiverse analysis under different prior settings for the overall effect and the between-study variance in the effect. The arrows show the overall trend relative to the main analysis with the primary prior settings.

```
## [1] 0.137500645 0.006155378 0.359535026
```

For the estimation, we want to rerun the models for the key sets, but without truncated priors (to allow an effect size below zero).

```
##          [,1]  [,2] [,3]
## 12121  0.06 -0.06 0.18
## 22121  0.09 -0.05 0.22
## 32121  0.08 -0.06 0.23
## 12211  0.08 -0.10 0.25
```

Table 4
Bayes factors for key analyses (participant-level exclusion set 1) under different prior settings.

scale on μ_θ	scale on σ^2_θ	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.

```
## 22211  0.16 -0.08 0.41
## 32211  0.18 -0.10 0.47
## 11112  0.03 -0.07 0.13
## 21112 -0.03 -0.17 0.13
## 31211  0.07 -0.20 0.34
```

Forest Plots

Robustness Checks

Table 5

Bayes factors for key analyses.

			Effect BF_{01}			Heterogeneity BF_{01}
Participant-level	Sample size	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.

Table 6

Bayes factors for key analyses.

			Evidence			
Participant-level	Sample size	Labs	BF_{0f}	BF_{01}	BF_{0+}	Effect [95% CI]
Klein et al. (2022)						
All	1544	17	12.50	5.41	791.66	0.02 [-0.11, 0.15]
White & US-born	1223	17	9.32	4.25	241.65	0.04 [-0.11, 0.19]
US-Identity > 7	1070	17	7.07	4.56	159.02	0.05 [-0.11, 0.21]
Chatard et al. (2020)						
All	699	7	14.97	2.17	13.93	0.08 [-0.13, 0.28]
White & US-born	378	7	6.76	0.96	2.75	0.14 [-0.11, 0.39]
US-Identity > 7	225	7	4.43	0.78	1.40	0.18 [-0.11, 0.49]
Current choice						
All	2211	21	35.75	10.28	12,127.13	0.01 [-0.11, 0.12]
White & US-born	983	16	21.30	13.88	3,633.28	-0.04 [-0.20, 0.12]
US-Identity > 7	272	9	8.43	2.71	11.33	0.05 [-0.22, 0.32]

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

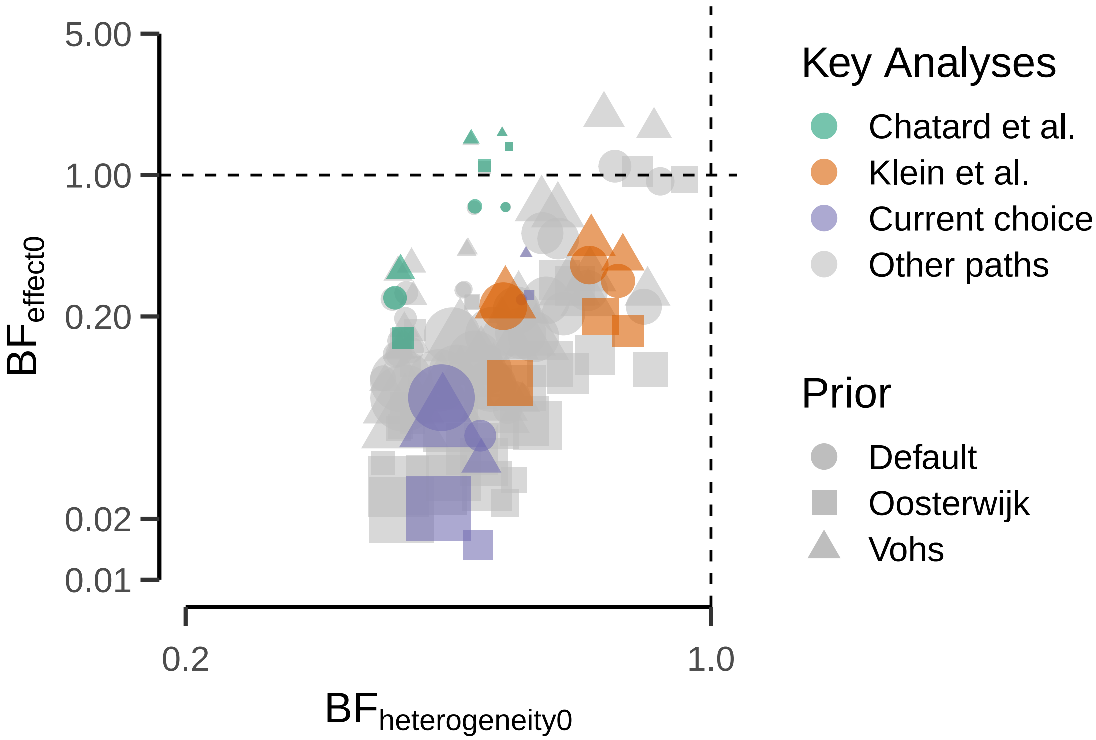

Figure 3. Results from the multiverse analysis for the model-averaged meta-analysis: Bayes factors in favor of a mortality salience effect are above the horizontal line, Bayes factors against the mortality salience effect are below the horizontal line. All analyses provide evidence against between-study 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 setting 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.

References

Chatard, A., Hirschberger, G., & Pyszczynski, T. (2020). *A Word of Caution about Many Labs 4: If You Fail to Follow Your Preregistered Plan, You May Fail to Find a Real Effect* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/ejubn

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. https://doi.org/10.1525/collabra.35271

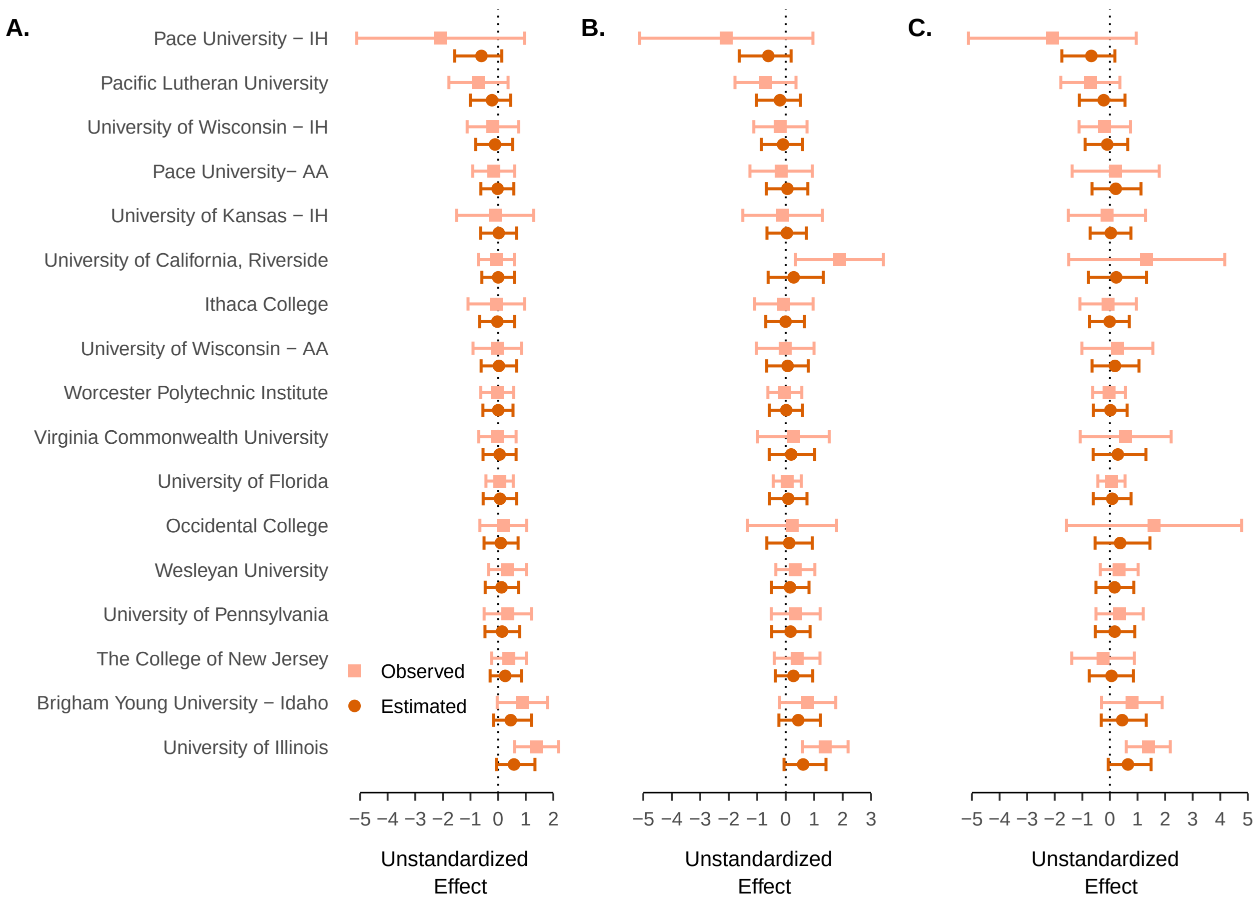

Figure 4. Forest plot with Bayesian parameter estimates for the key analyses by Klein et al. for the three participant-level exclusion sets (applied to AA-participants only) with data collected after the lead team posted their preregistration, and only studies that featured more than 60 observations. **A.** Participant-level exclusion set 1. The grey points represent unstandardized observed effects for each study with 95% confidence intervals. The black points represent estimated unstandardized effects from the unconstrained model with 95% credible intervals. **B.** Participant-level exclusion set 2. **C.** Participant-level exclusion set 3.

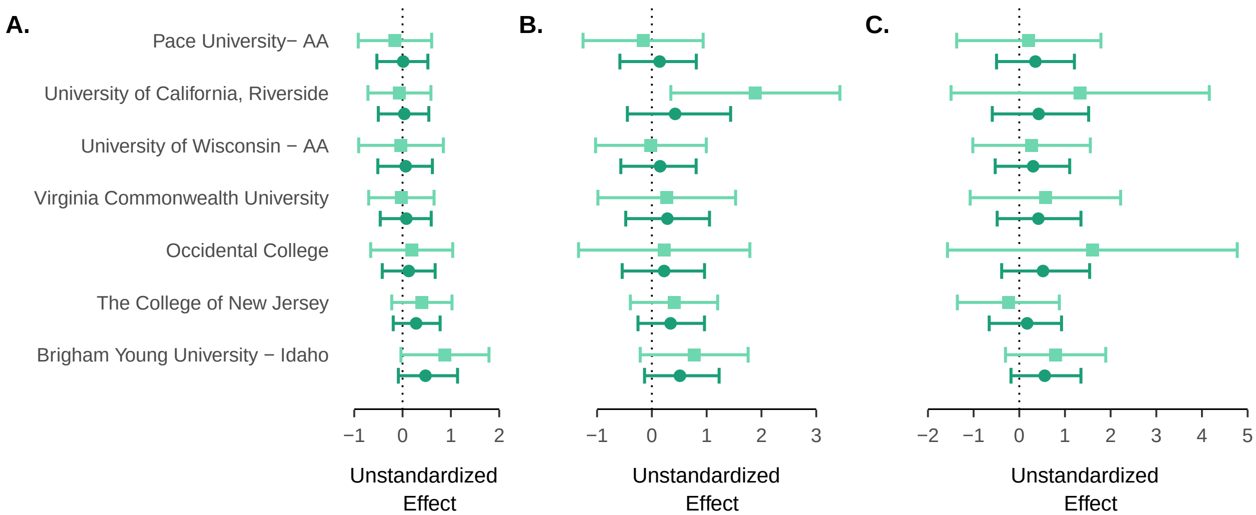

Figure 5. Forest plot with Bayesian parameter estimates for the key analyses by Chatard et al. for the three participant-level exclusion sets, only studies that featured more than 80 observations, and for AA-labs only. **A.** Participant-level exclusion set 1. The grey points represent unstandardized observed effects for each study with 95% confidence intervals. The black points represent estimated unstandardized effects from the unconstrained model with 95% credible intervals. **B.** Participant-level exclusion set 2. **C.** Participant-level exclusion set 3.

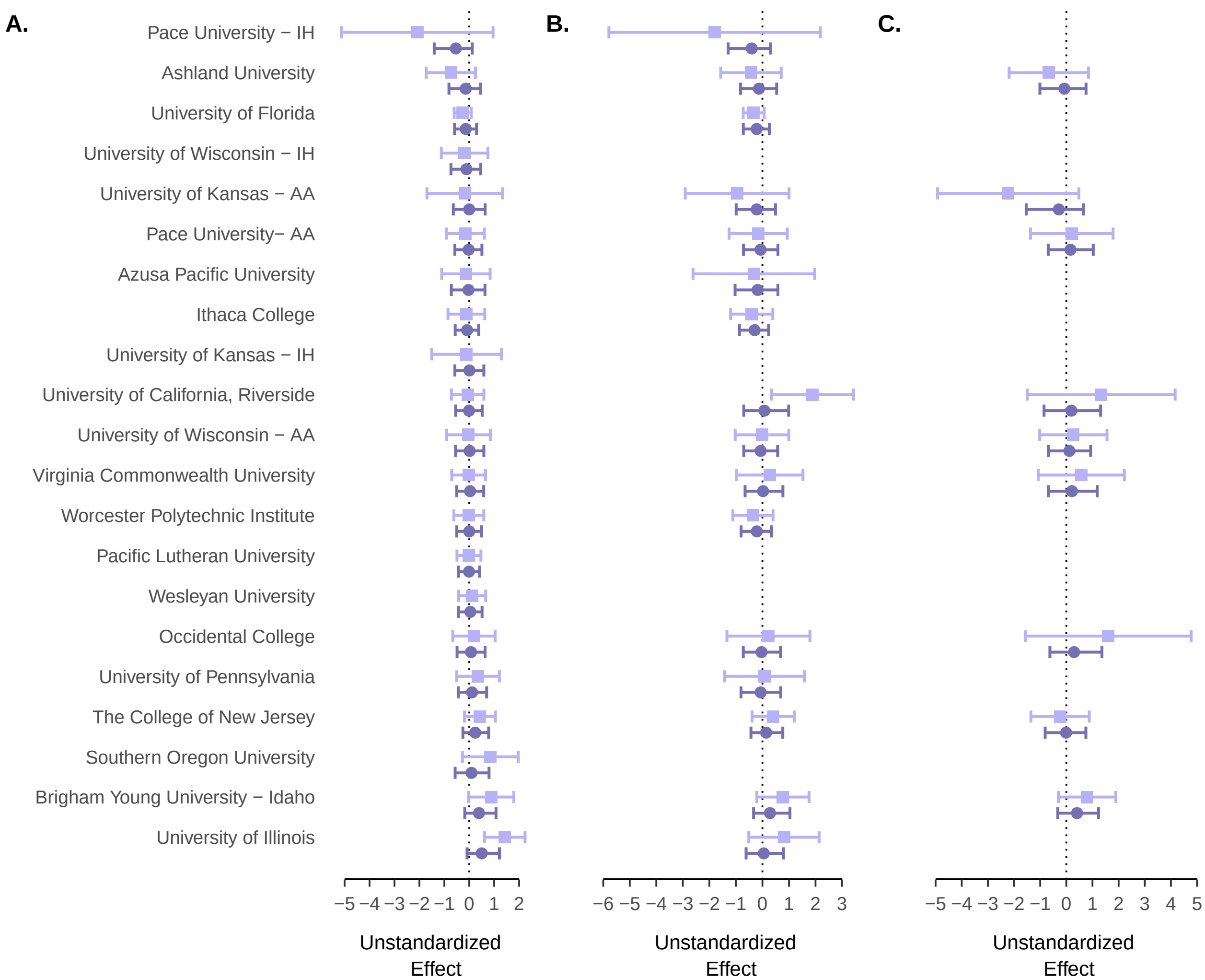

Figure 6. Forest plot with Bayesian parameter estimates for the key analyses of our choice for the three participant-level exclusion sets (applied to both AA- and IH-participants), including all participants and all labs. **A.** Participant-level exclusion set 1. The grey points represent unstandardized observed effects for each study with 95% confidence intervals. The black points represent estimated unstandardized effects from the unconstrained model with 95% credible intervals. **B.** Participant-level exclusion set 2. **C.** Participant-level exclusion set 3.