Managing the terror of publication bias: A comprehensive p-curve analysis of the Terror Management Theory literature

Lihan Chen  
McGill University

Rachele Benjamin  
University of British Columbia

Yingchi Guo  
University of British Columbia

Addison Lai  
University of British Columbia

Steven J. Heine (✉️ heine@psych.ubc.ca)  
University of British Columbia

Research Article

Keywords: Terror management theory, p-curve, z-curve, meta-analysis, replication crisis

Posted Date: May 15th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-1254756/v1

License: ☝️ This work is licensed under a Creative Commons Attribution 4.0 International License. Read Full License
Abstract

We assessed the evidential value of the large literature (k=826 studies) investigating the mortality salience (MS) hypothesis from terror management theory. We employed a multitool assessment approach and reviewed past efforts to replicate experiments testing the MS hypothesis, and conducted a p-curve, a z-curve, and a random effects meta-analysis including bias corrections of the selection model, PET-PEESE, and WAAP-WLS on the studies. Overall, the different meta-analytic tools pointed to conflicting conclusions, reflecting differences in the methodology and philosophy of these tools. Our synthesis of these findings suggests there are true effects underlying some studies of the MS hypothesis, although the effects are highly heterogeneous, and the majority of studies are underpowered. We recommend future replications to assume a smaller effect size (r = .10 ~ .15) and to follow expert guidance in the experimental protocol. Given the conflicting findings that emerged, we suggest that future attempts to evaluate other literatures would benefit from a multitool assessment approach.

Full Text

One of the core principles underlying scientific progress is that scientific findings should be replicable (e.g., Popper, 1959; Schmidt, 2009). In recent years, scientists have become uncomfortably aware of just how poor the evidence is for the replicability of many findings, and many have argued that the scientific method is in the throes of a replication crisis (e.g., Ioannidis, 2005; Pashler & Wagenmakers, 2012). While serious doubts about the replicability of key findings extend across scientific disciplines (e.g., Begley & Ellis, 2012; Prinz, Schlange, & Asadullah, 2011), by a variety of measures the replicability of many key psychological findings has been especially concerning (Open Science Collaboration, 2015; Wagenmakers et al., 2016). Moreover, within psychology, the replicability of studies in the discipline of social psychology has been particularly poor (e.g., Youyou, Yang, & Uzzi, 2023). This has resulted in much interest in evaluating the evidence for many key social psychological phenomena.

The replicability of psychology studies has been undermined by researchers engaging in a number of questionable research practices that have inflated effect size estimates (John, Loewenstein, & Prelec, 2012; Simmons, Nelson, & Simonsohn, 2011), such as hypothesizing after the results are known (also known as HARKing; Kerr, 1998), leaving moderators, dependent, and independent variables undisclosed if they did not produce an effect (Simmons et al., 2011), adding participants in increments until an effect was significant, and selectively excluding outliers in order to report $p < .05$ (John et al., 2012). These strategies are collectively known as p-hacking (Simonsohn, Nelson, & Simmons, 2014a), and though researchers may not always be aware when they are p-hacking (Vazire, 2015), these problematic practices overestimate the true size of a given effect. Moreover, historically the field has suffered from publication biases, in that studies that find significant evidence for an effect have tended to be far more likely to be published in comparison to those that fail to find significant evidence (Greenwald, 1975; Rosenthal, 1979). Given how common these questionable research practices seem to be (John et al., 2012; Stefan & Schönbrodt, 2023), they not only raise questions about the evidential value of a single
study, but about the evidential value of an entire literature. But how can researchers best assess the evidential value of an expansive psychological literature?

We direct this question to a consideration of terror management theory (TMT). TMT has explored how people confront the existential anxieties that arise from their awareness that they inevitably will die (Greenberg, Pyszczynski, & Solomon, 1986). Building on the work of Ernest Becker (1973) the theory suggests that people seek to assuage their existential anxieties by defending their cultural worldviews to achieve feelings of symbolic immortality. The theory posits that because an individual’s culture can exist long after one’s death, people can become imbued with the potential immortality that is inherent in their cultures by either bolstering their cultural worldview or by viewing themselves as living up to the standards of their cultures. People can achieve feelings of symbolic immortality by engaging in a variety of symbolic acts, such as investing in one’s national identity (Greenberg, Pyszczynski, Solomon, Simon, & Breus, 1994), derogating members of other religious groups (Vail, Courtney, & Arndt, 2019), or punishing people who are breaking the laws (Rosenblatt, Greenberg, Solomon, Pyszczynski, & Lyon, 1989).

TMT has been a highly influential theory that has been explored in more than 1500 studies conducted in more than 20 countries (Vail et al., 2019), and several of the foundational studies have been cited over 1000 times each (see Greenberg et al., 1994). The theory has had much reach into the popular culture, being the subject of films (e.g., Flight from Death, 2003) and popular books (e.g., Solomon, Greenberg, & Pyszczynski, 2015). On the other hand, the theory has remained rather controversial. Some critics have maintained that the theory is too imprecise, allowing for heterogeneous and conflicting results to be framed as supportive evidence (e.g., Martin & van den Bos, 2014), alternative theoretical accounts have been proposed (e.g., Kirkpatrick & Navarrete, 2006; Proulx & Heine, 2006; but see rebuttals by Pyszczynski et al., 2015), and the theory’s replicability has been called into question (e.g, Klein et al., 2022; Yen & Cheng, 2013).

TMT posits three core hypotheses regarding how people engage with thoughts of their own mortality: 1) an anxiety-buffer hypothesis, which suggests that people possess a buffer that serves to reduce the anxiety caused by thoughts of one’s own death (e.g., Greenberg et al., 1992); 2) a death thought accessibility hypothesis, which suggests that death thoughts will become more accessible when the anxiety-buffer has been undermined (Hayes et al., 2010), and 3) a mortality salience (MS) hypothesis, which predicts that people will engage in more defense of their cultural worldview whenever they encounter thoughts about their mortality (e.g., Rosenblatt et al., 1989). The vast majority of research has targeted the MS hypothesis, and we focus our assessment of the literature surrounding it. Most of the research on the MS hypothesis has explored how people engage in distal defenses against death reminders, which emerge most strongly after a delay (Pyszczynski, Greenberg, & Solomon, 1999), such that the mortality thoughts are no longer a focus of conscious attention (Solomon, Greenberg, & Pyszczynski, 1991).

Given the enormous database of studies that have tested the MS hypothesis, and the controversy that surrounds it, our exploration of TMT’s replicability begins with the MS hypothesis. Below, we consider a
number of different methods by which we can use to evaluate the evidential value of a literature, and how those speak to the case of TMT.

**Evaluating the Replicability of a Single Study**

Replication is a hallmark of science (Popper, 1959), and often this is assessed by determining how well the findings from a single study replicate (see Nosek et al., 2022). Prior to the replication crisis, this question was typically addressed in a rather haphazard way: individual labs would sometimes conduct studies that included a replication of an original finding, often in an attempt to extend those findings. However, there were limited guidelines for how to conduct such replications, and there was the widespread sense that any failed replication efforts would be unlikely to be accepted for publication by journals. In this context, it was difficult to gain a clear sense from the literature of how likely any individual studies were to be replicated.

In the aftermath of the replication crisis, researchers began to make more transparent efforts to document the replicability of individual studies. One approach to evaluate the replicability of a single study is to conduct a pre-registered high-powered replication of a selected study (Simmons, Nelson, & Simonsohn, 2021). If an effect is real, and a study is well-powered enough to be able to detect this effect at a high threshold of likelihood, then a pre-registered study would be expected to replicate. There have been many efforts to conduct pre-registered replications for a variety of psychological phenomena, and these have had varying degrees of success (e.g., Garrison, Tang, & Schmeichel, 2016; Persson, Andersson, Koppel, Vastfjall, & Tinghog, 2021).

We have identified a total of 6 pre-registered studies of the MS hypothesis that have been conducted. Three of these studies failed to replicate the original findings (see Rodríguez-Ferreiro, Barberia, González-Guerra, & Vadillo, 2019; Sætrevik & Sjåstad, 2019; Schindler, Reinhardt, & Reinhard, 2021), whereas 3 of these studies successfully replicated the original studies (Dunn, White, & Dahl, 2020; Schindler, Pfattheicher, Reinhard, & Greenberg, 2019; Vail, Courtney, & Arndt, 2019). Overall, efforts to provide a preregistered high power replication of a single study of the MS hypothesis have provided mixed evidence.

Another popular approach for replicating a single study that has emerged is the multisite replication project (e.g., Open Science Collaboration, 2015). This approach recruits a number of labs which each attempt to replicate the same study, and then the results of the overall effort can be analyzed. The key benefits of this approach are that these efforts result in very large sample sizes and the multiple labs reduce the likelihood that the results are the product of any author effects. There have been several attempts to evaluate various social psychological phenomena with this approach. One investigation identified 36 multisite replication efforts that were designed to test a single social psychology study (Baumeister, Tice, & Bushman, in press). Of these 36 attempts, only 11% succeeded, 14% offered mixed results, and a full 75% failed. This dismal rate of replication from the multisite replication approach suggests matters do not look very optimistic for social psychology. However, critics of the multisite
replication approach argue that the magnitude of effects may be underestimated because of concerns of a lack of engagement among participants, and a low fidelity mirroring of the original study’s procedures, resulting in weaker tests of the hypotheses (Baumeister et al., in press; Ellefson & Oppenheimer, in press; Schimmelpfennig et al., 2023).

The multisite replication approach has been used in an attempt to replicate one of the early experiments on the MS hypothesis (Greenberg et al., 1994, Study 1), across 21 labs and 2,220 participants (Klein et al., 2022). The study was conducted in two separate ways: one half of the labs were asked to recreate the study based on their own (in-house) interpretation of the methods section of the paper, whereas the other group of labs were provided with more detailed instructions from the advice of the original authors. The results of this attempt did not provide much support for the MS hypothesis. The meta-analytic effect size was negligible (Hedges’ $g = 0.03$), even in those studies where the replicators strictly adhered to the original authors’ guidelines. However, Chatard, Hirschberger, and Pyszczynski (2020) noted that some of the studies included in the multisite replication failed to meet the preregistered criteria regarding the minimum sample size. When these authors re-analyzed the data of only those studies that met the minimum preregistered sample size they found that the key effect weakly replicated in the studies that followed the advice of the original authors (Hedges’ $g = 0.13, p = .045$), at least when one particular set of exclusion criteria was applied. In sum, the multisite replication of the MS hypothesis is rather discouraging. The effect does not appear replicable when researchers try an in-house replication of the studies, whereas the efforts that were guided by the original authors’ advice reveal conflicting evidence depending on the exclusion criteria that are followed.

We have thus found several efforts to address the question of the replicability of individual studies testing the MS hypothesis. These various attempts to replicate individual studies testing the MS hypothesis has thus resulted in somewhat conflicting evidence that makes it difficult to draw a firm conclusion on the replicability of these studies.

It would seem that the question of whether a single study replicates would be rather concrete and easy to answer: does the study provide evidence for the same significant effect that the original study produced? Nonetheless, there remains considerable debate regarding the most appropriate standards by which to evaluate replication attempts (e.g., Nelson, Simmons, & Simonsohn, 2018; Nosek et al., 2022), and thus it is difficult to provide a definitive statement about the replicability of many studies. However, the complications regarding the assessment of whether a single study replicates would seem to get far more complex when we turn to evaluations of the replicability of an entire literature.

**Methods for Evaluating the Evidential Basis of a Literature**

It is a far more challenging enterprise to evaluate the evidential basis of an entire literature than of any given single study. Literatures for psychological phenomena tend to report many different kinds of tests of a given hypothesis, as well as questions regarding boundary conditions, sampling issues, mechanisms, and conceptual replications. The substantial heterogeneity that exists across a literature...
means that it is rarely clear which of the many effects that have been identified in a mature literature could be seen to best speak to the evidential basis of that literature (Stanley, Carter, & Doucouliagos, 2018). For example, it seems quite plausible that any given large literature would contain both effects that reliably replicate as well as other effects that reliably fail to replicate, and calculating a single overall average effect may be of limited utility (Simonsohn, Simmons, & Nelson, 2022). How can we best evaluate the evidential value from an expansive literature that includes so many different extensions of a particular theory?

One common solution that has often been relied upon to evaluate a literature has been the traditional meta-analysis. A meta-analysis summarizes the effects for a given phenomenon across a specified universe of studies that have investigated it. Typically, authors of meta-analyses specify a set of inclusion and exclusion criteria by which they decide which studies, and which effects from those studies, will be incorporated into the analysis. Traditional meta-analysis takes one of two approaches: the fixed effects approach, which estimates a single effect size that is presumed to underlie all studies in the universe, or the random effects approach, which captures variations in effect sizes among studies that use different operationalizations and populations. Psychologists typically use the random effects approach, treating the population effect size as a distribution with mean \( \delta \) and standard deviation \( \tau \). The mean effect size from meta-analyses is often cited to indicate the magnitude of the effects found in the literature. The standard deviation reflects the heterogeneity of those effects, which tend to be quite high in meta-analyses of psychological literatures— the median estimate of heterogeneity in meta-analyses published in *Psychological Bulletin* over the past 30 years was over 70% (Van Erp et al., 2017).

Though traditional meta-analyses have frequently been turned to for assessing the evidential value of literatures, they have some key shortcomings. First, the aforementioned concerns of publication biases mean that studies that find significant evidence for an effect have been far more likely to be published in comparison to those that fail to find significant evidence. Meta-analyses tend to be largely based on published studies which make the problem of publication bias especially concerning. Even when the true population effect is null, traditional meta-analysis may yield a substantial effect size estimate by aggregating the false positives in the literature. In cases where the true population effect does exist, the meta-analytic effect size estimate is still likely to be inflated due to the same publication bias. Second, the many kinds of \( p \)-hacking efforts that researchers have relied upon will artificially inflate the estimate of an effect size for any individual study. By summarizing the effects from the published literature meta-analyses tend to aggregate the cumulative effects of all of those \( p \)-hacking efforts. As a result, the estimates of the average effect from virtually any traditional meta-analysis are likely to be greatly exaggerated (Nelson et al., 2018). In response to the inherent biases of meta-analyses, a number of meta-statistical tools and bias correction techniques have been introduced to address these limitations. These methods include the weighted-least squares meta-regression with the weighted average of adequately powered studies (WAAP-WLS; Stanley & Doucouliagos, 2017; Ioannidis, Stanley, & Doucouliagos, 2017), the selection model (Hedges, 1984; Iyengar & Greenhouse, 1988; Vevea & Hedges, 1995), and the precision-effect test and precision-effect estimate with standard errors (PET-PEESE; Stanley & Doucouliagos, 2014).
There have been a number of meta-analyses conducted on the MS hypothesis (see Burke, Martens, & Faucher, 2010; Burke, Kosloff, & Landau, 2013; Martens, Burke, Schimel, & Faucher, 2011). The most extensive one performed a random effects analysis on 277 individual studies, and identified that the average effect size of that literature was $r = .35$, corresponding to a moderate effect (Burke et al., 2010). This provides an optimistic assessment of the evidential value of TMT; however, because this analysis did not employ any of the various bias reduction techniques, it would surely seem to be an overestimation of the magnitude of the actual underlying effect. Moreover, a re-analysis revealed that experimenter effects accounted for a lot of variance (Yen and Cheng, 2013) which undermines any conclusions that may be drawn from the meta-analysis.

The Present Research

Given the many critiques of traditional meta-analysis methods, as well as the questions regarding the replicability of individual TMT studies that have employed the multisite and the preregistered high powered single study approach, at present we do not have a particularly convincing evaluation of the evidential basis of the MS hypothesis. To address this, we sought to provide a comprehensive evaluation of the literature testing the MS hypothesis using a multitool assessment approach, which employed the most current state-of-the-art analytic tools. As we describe in detail below, we evaluated the evidence of the literature examining the MS hypothesis by conducting 1) a $p$-curve analysis, 2) a $z$-curve analysis, and 3) traditional meta-analyses using three bias-correction approaches (selection model, PET-PEESE, and WAAP-WLS). Each of these analyses were applied to a universe of more than 800 studies investigating the MS hypothesis, resulting in what we believe to be the largest $p$-curve and $z$-curve analyses to date. With such a large literature, and a reliance on the field's most current tools, this should provide a thorough assessment of the evidential value of the theory. See Table 1 for a brief summary of our findings.

Evaluation of Evidential Value 1: A $P$-curve Analysis

The $p$-curve, introduced by Simonsohn et al. (2014a), is a technique to estimate the file drawer of a literature by focusing on the distribution of published $p$-values that fall below the standard convention of $p < .05$. The $p$-curve is, in essence, a histogram of all significant statistical tests of the focal hypothesis in a literature. In the case where there is no population effect present, all $p$-values are equally likely. In contrast, when there is a true population effect, smaller $p$-values are more likely than larger $p$-values, forming a right skewed distribution (see a simulated distributions in Figure 1) – the higher the power of the test, the more right skewed this distribution becomes. Therefore, Simonsohn et al. (2014a) suggest that when the histogram of the $p$-values in a literature reveals a flat distribution of $p$-values of a similar height, it suggests a lack of evidential value in that literature. By the same logic, the presence of right skewness indicates the existence of a population effect, and the amount of the skewness measures the power of these studies. When the $p$-curve is tested against a flat distribution, it can reveal whether $p$-
values were most likely to be generated by a null effect. In contrast, when it is tested against a
distribution with a right skew corresponding to 33% power, it can reveal whether the studies in the
literature have been conducted with sufficient power.

*p*-curve assumes that the effects in the meta-analysis were only published if they met the condition of \( p < .05 \), which restricts the kind of effects that can be included in the \( p \)-curve, namely, only effects that are
associated with the focal hypotheses of the studies. The interaction terms must be used when there is an
attenuating interaction, and simple main effects must be used when there are reversing interactions; the
logic is that these effects are more likely to be selected only based on \( p < .05 \) (see Simonssohn et al.,
2014b). Consequently, a \( p \)-curve analysis will often include a mix of main effects, simple main effects,
and interactions, depending on what was hypothesized by each study.

Simulation studies find that the \( p \)-curve performs generally well (Carter, Schönbrodt, Gervais, & Hilgard,
2019), however, as the \( p \)-curve does not incorporate a random effects model, its significance test yields a
somewhat narrow interpretation under heterogeneity. Specifically: 1) If a set of studies contain both null
effects and nonzero effects, the \( p \)-curve test detects whether some of those studies contain nonzero
effects; 2) When the \( p \)-curve is used to estimate the average power of the studies, it only applies to the
studies included in the meta-analysis, rather than the population of potential studies of the phenomenon
(Simmons, Nelson, & Simonssohn, 2014a). The goal and philosophy of the \( p \)-curve therefore differ from
those of meta-analyses based on random effects models, which aim to estimate the average size of a
population of heterogeneous effects that varies due to differences in study designs and environmental
factors (e.g., see Carter et al., 2019; also see Simonssohn, Simmons, & Nelson, 2022, for \( p \)-curve’s
perspective on average effect).

Under low heterogeneity, the differences between the philosophies of \( p \)-curve and random-effects-based
meta-analyses is not of notable concern. In the presence of high heterogeneity, however, interpreting the
\( p \)-curve tests as tests for overall effects would result in a high false positive rate. This false positive rate
grows as heterogeneity increases, and as the number of included studies increases (see McShane,
Böckenholt, & Hansen, 2016; Carter et al., 2019). When interpreted as intended, however, the authors of
the \( p \)-curve maintain that \( p \)-curve is robust to heterogeneity, and is only compromised by the inclusion of
highly powered outlier studies (see Simmons, Nelson, & Simonssohn, 2018). Nevertheless, because the
number of included studies in our analysis is so large, we preplanned subgroup analyses to ameliorate
the impact of heterogeneity, with an accompanying *Shiny* application (Chang et al., 2020) to explore
whether the results of the main \( p \)-curve hold for the less heterogeneous subgroups, which are presented in
the Supplementary Online Material (SOM). Lastly, we explored the robustness of our \( p \)-curve results by
conducting sensitivity analyses, examining how outlier removal would affect our results.

**Results of P-Curve Analysis**

The main analysis using the \( p \)-curve includes \( K = 826 \) studies (see Figure 2) that satisfied all the inclusion
criteria. Because studies in the full sample includes outliers and involve idiosyncratic methods and
Diverse populations, the resulting heterogeneity of effect sizes means that we should not interpret $p$-curve's evidential value tests as tests for an overall average effect of all similar studies. Instead, our main analysis focuses on the narrower interpretation of whether some studies have underlying true effects, and what is the expected average rate of replication for the specific studies included in our sample.

**Inferential tests and estimated power.** The test for right skew determines that there is evidential value if the inferential test on the half curve (these are $p$-values smaller than .025) yields $pp < .05$, or both the full and half $p$-curves yield $pp < .10$. Using the half $p$-curve (all values below .025), we find significant right skew, $Z = -14.08, p < .0012$. Using the full $p$-curve, we also find significant right skew, $Z = -14.52, p < .001$. Thus, we conclude there are nonzero effects underlying some studies in our sample.

Since the right skew test is significant, we do not interpret the results for the flatness test. However, we note that Simonsohn et al. (2014a), in designing the flatness test, suggested using 33% power as the cutoff to decide whether the power of a body of literature is sufficient. Based on the shape of the $p$-curve of the literature, the average power of all 826 studies is estimated to be 26%, $CI_{95} = [21.6\%, 29.7\%]$, significantly below 33% cutoff. This result suggests that tests of the MS hypothesis are generally underpowered in the literature (See Figure 2 and Table S2). We suggest it is more likely that this relatively low level of power in the literature is the result of some studies having very low power, whereas others have reasonable power, rather than inferring than most studies in the literature have around 26% power.

In sum, on the basis of the inferential tests and power estimates, we conclude that the entire universe of studies testing the MS hypothesis contains studies that investigate real effects. However, the modal study is underpowered and unlikely to reliably detect or estimate the effect, and direct replication attempts are unlikely to succeed at the same sample size. The average $N$ in the universe of studies was 114, which corresponds to 57 or fewer participants per cell in between subject designs; roughly enough to detect an effect size of at least $r = 0.18$ with 27% power.

**$p$-curve sensitivity analyses.** We examined the robustness of our main finding using the sensitivity analysis provided by the $p$-curve app. By removing the highest or the lowest $p$-values one by one and recomputing the results, we can see how many $p$-values from the extremes of the distribution would need to be excluded before the conclusions of our $p$-curve analyses would be changed. This helps us determine if any of our findings are driven by a few extreme data points. We found that for the test of right skewness to be nonsignificant ($p > .05$), 311 of the lowest $p$-values in the full $p$-curve and 147 on the half $p$-curve must be dropped. Because these numbers are so high, we conclude that our results for the evidential value test is robust against outliers.

**Evaluation of Evidential Value 2: A $Z$-curve Analysis**

$Z$-curve 2.0 (which we will simply refer to as “$z$-curve”) is an alternative to $p$-curve that has been developed more recently (Bartoš & Schimmack, 2021). Similar to $p$-curve, $z$-curve fits the observed distribution of the $z$-values to the theoretical distribution of $z$-values given an effect size, using the best fit
to estimate the power of the studies. As power increases, the best fit distribution above the significance threshold of \( z = 1.96 \) will resemble a larger proportion of the full \( z \)-curve (see simulated distributions in Figure 3). While a one-to-one correspondence can be drawn from \( p \)-values to \( z \)-values under some assumptions, and both methods are used to estimate the power of the studies, there are several important differences between the two approaches. First, \( z \)-curve explicitly incorporates a random effects model, which allows it to directly handle heterogeneity in effect sizes using a mixture of \( z \)-distributions. Further, the \( z \)-curve provides two power estimates (Bartoš & Schimmack, 2021): 1) The conditional average power of the studies that yielded significant effects, called the expected replication rate (ERR), which is the probability that an exact replication of a published effect will succeed, and is an equivalent to the \( p \)-curve power estimate, and 2) the unconditional average power of the studies in the literature, called the expected discovery rate (EDR), which is the overall probability of obtaining significant effects in a literature with a mix of significant and nonsignificant results. When modeling a heterogeneous literature the mixture model used for power estimation may incorporate a mix of effect sizes, including null effects. A comparison between the ERR and EDR allows for a more fine-grained evaluation of the state of the literature than what is offered by the \( p \)-curve.

### Results of Z-Curve Analysis

Re-analyzing the 826 significant effects from the \( p \)-curve using \( z \)-curve (see Figure 5), we found an estimated ERR of .19 (CI\( _{95} = [.13, .26] \)), suggesting that a significant result can be replicated at a success rate of 19%. This estimate is slightly more pessimistic than the \( p \)-curve estimate of 26% power, though not nearly as pessimistic as the EDR, estimated at .08 (CI\( _{95} = [.05, .16] \)). The EDR suggests that the unconditional mean power of the entire literature, including the tests on both true and null effects, is a mere 8%. This is very close to the minimum of 5% unconditional power under the significance threshold of \( p < .05 \) that would be yielded from a literature containing only null effects. The difference between the EDR and ERR can be thought of as the difference between the predicted success rate of the exact replication of a published effect (19%), and the estimated rate that an average MS effect will turn out to be significant in general after taking the file drawers into consideration (8%). Therefore, while the \( z \)-curve agrees with the \( p \)-curve that the published literature may contain some true findings, it further suggests that the literature overall has an abundance of studies that are either showing false positives due to a combination of publication bias and \( p \)-hacking, or were conducted with such low power that their results are indistinguishable from false positives.

The \( z \)-curve uses a mixture of \( z \)-distributions to directly model the heterogeneity in effect sizes, and therefore some maintain that it outperforms \( p \)-curve when heterogeneity is high (see Brunner & Schimmack, 2020; Bartos & Schimmack, 2020; but see Simmons et al., 2018). For this reason, it is possible that the \( z \)-curve's more pessimistic result is more reflective of the average study of the MS hypothesis. However, it is worth nothing that this latest version of the \( z \)-curve has only recently been published, and has not been extensively studied in independent investigations in the same way other approaches we included have.
Evaluation of Evidential Value 3: Traditional Meta-Analytic Approaches

As we noted above, traditional meta-analyses are susceptible to publication bias and an aggregation of \( p \)-hacking efforts, and some techniques that work within this framework have been developed to address the bias. We employed three of these methods, the selection model (Iyengar & Greenhouse, 1988), PET-PEESE (precision-effect test and precision-effect estimate with standard errors), and weighted-least squares meta-regression with the weighted average of adequately powered studies (WAAP-WLS; Stanley & Doucouliagos, 2017; Ioannidis, Stanley, & Doucouliagos, 2017). Due to methodological differences in these more traditional methods (e.g., \( p \)-curve requires the selection of significant effects related to the focal hypotheses, while traditional meta-analysis tends to select similar types of effects regardless of their relationships to the hypotheses or significance; see Methods for more details), we decided it was unsuitable to directly apply them on the coded effects for the \( p \)-curve and \( z \)-curve analyses. We have therefore re-coded the same studies for the analyses in this section, following a more traditional selection and coding procedure similar to that of Burke et al. (2010), which selects simple main effects from the studies included, rather than a mix of main and interaction effects associated with the focal hypotheses.

Selection model

Selection models build on the traditional random effects model by incorporating thresholds to model for the fact that studies with larger \( p \)-values are less likely to be published. The average effect size (\( \delta \)), the heterogeneity (\( \tau \), the standard deviation of the effect size distribution), and the probabilities that larger \( p \)-values are published, are estimated simultaneously using maximum likelihood. A simple case of a selection model is the three-parameter selection model, where the parameter \( p_1 \) is the average probability that a nonsignificant effect gets published, under the two-tailed \( p \)-value threshold of .05. This approach assumes that an effect with a two-tailed \( p < .05 \) is always published, whereas an effect with a two-tailed \( p \geq .05 \) is published \( p_1 \) of the time. Additional thresholds can be added to the selection model. For example, “marginal” effects \( (.05 \leq p < .10) \) can be given their own probability of publication that is independent of the probability that effects with \( p \geq .10 \) gets published, and “surprising” effects \( (p \geq .95) \) can be given another independent probability of publication.

Similar to the \( p \)-curve, the selection model relies on the simplistic assumption that publication probability of an effect is fully determined by its \( p \)-value. Whereas \( p \)-curve attempts to ensure the tenability of this assumption by requiring only focal hypotheses be included in the analysis, selection model instead addresses more complex patterns of publication probability by incorporating multiple \( p \)-value thresholds. The effect size estimates from selection model can therefore be sensitive to threshold specification. Even with multi-thresholds, selection model may not be able to adequately capture important nuances in the publication bias in a literature. This could lead selection model to underestimate effects in the presence of publication bias and \( p \)-hacking, as shown in simulation studies (Carter et al., 2019). It has been
suggested that selection model analysis is conducted as a sensitivity analysis accompanying other meta-analytic approaches, rather than as a stand-alone analysis (McShane et al., 2016).

**PET-PEESE**

Another technique that operates within the traditional meta-analytic framework is PET-PEESE (Stanley & Doucouliagos, 2014). The issue targeted by PET-PEESE is that studies with smaller sample sizes are likely to produce less precise estimates, which are more likely to deviate from the population true effect. Normally this would not bias a meta-analytic estimate; we would observe equal random deviations in both directions. However, publication bias means that random deviations are more likely to land in the published literature if they produce a large positive effect. The logic of PET-PEESE is that if we fit this relationship to a regression, we should in theory obtain the true population effect when we examine the intercept, the predicted effect size when the standard error is 0.

The PET analysis is a weighted meta-regression predicting the true effect size from standard error. While PET has good Type I error control, it tends to underestimate the effect size. It has therefore been recommended that if PET rejects the null, the PEESE analysis should be used to estimate the effect size. PEESE is the same meta-regression as PET, except it uses the squared standard error as the predictor, which leads to a less biased and more precise effect size estimate. The PET-PEESE procedure thus combines the Type I error control of PET and better effect size estimate of PEESE.

Overall, PET-PEESE has a low false positive rate, but its estimation of the effect size tends to be imprecise with a sample size of small, heterogeneous studies, resulting in large standard errors, such that “if all studies are small, PET-PEESE has little power to identify a genuine empirical effect” (p. 590, Stanley, 2017). This issue is typically addressed by using a more liberal significance criterion, e.g., Stanley and Doucouliagos (2014) recommended using \( \alpha = .10 \). Although the number of studies included in the analysis can serve to increase power, our broad inclusion criteria also increase the heterogeneity of the effects we analyze, which can cause the PET-PEESE estimate to be less efficient, potentially offsetting the benefit. Lastly, PET-PEESE can underestimate the size of small effects in the presence of \( p \)-hacking and publication bias (Carter et al., 2019). Overall, PET-PEESE may be an especially conservative means by which to analyze a literature characterized by small sample sizes, high heterogeneity, and publication bias.

**WAAP-WLS**

Stanley & Doucouliagos (2017) showed that an unrestricted weighted least square (WLS) meta-regression approach would yield effect size estimates that, while biased under selective publication, are generally less so compared the traditional random effects estimates. Additional bias correction can be achieved by applying the unrestricted WLS approach to the subset of effects associated with 80% power or higher (while discarding the rest), in a procedure known as the weighted average of the adequate power (WAAP; Iaonnidis, Stanley, & Doucouliagos, 2017). Following the WAAP-WLS procedure (see Carter
et al., 2019), when a set of studies contains no adequately powered effects, or exactly one such effect, WAAP is not applicable, and only the uncorrected WLS estimate is available; when there is more than one adequately powered effect, the WAAP estimate is produced using only those effects.

Simulation studies by Carter et al. (2019) showed that WAAP-WLS can provide a reasonably unbiased estimate of the effect size with a high power to detect the effect given that a true underlying effect exists. However, the same studies also revealed that, in the presence of publication bias, WAAP-WLS has unacceptable Type I error rates in virtually all conditions the authors investigated (small vs. large sample sizes, high vs. low heterogeneity, medium vs. high publication bias, low vs. high p-hacking), with a 40% chance to falsely conclude that an effect exists even in its best performing conditions. This liberal approach can thus serve as a robustness check to the more conservative PET-PEESE.

Results of Meta-Analysis Including Bias Corrections

Traditional Meta-Analysis. We first conducted a traditional meta-analysis test on the 687 re-coded effects. The advantage of meta-analysis over the p-curve is that it can produce an effect size estimate on the simple main effects of the MS hypothesis. And, because we followed a procedure similar to Burke et al. (2010), our findings can be compared directly except with roughly a decade more data points included. Performing traditional random effects meta-analysis, without applying any bias corrections, we found an estimated effect size of Hedge’s $g = .66$ ($\text{CI}_{95} = [0.62, 0.69]$), with heterogeneity $\tau = .40$ ($\text{CI}_{95} = [.36, .43]$). This effect roughly corresponds to $r = .31$, and is a comparably large effect to the Burke et al. (2010) test ($N = 277$), which yielded $r = .35$. The funnel plot (Figure 6) for the effects in this meta-analysis shows a high degree of asymmetry, indicating the presence of publication bias, which we next address through bias correction methods.

Selection Model. Applying the selection model, we estimate the probability that “marginally” significant effects (one tailed $0.025 \leq p < 0.05$, corresponding to two-tailed $0.05 \leq p < 10$), nonsignificant effects (one tailed $0.05 \leq p \leq 0.95$), and “surprising” significant effects ($p > 0.95$) are published. The selection model estimates these probabilities along with the average effect size (in Hedge’s $g$) and the variance of the effect size simultaneously, which corrects for potential publication bias in the meta-analytic estimates.

Overall, our universe of studies yielded 444 significant effects, 95 “marginally” significant effects, 142 nonsignificant effects, and 6 “surprising” effects. The selection model assumes significant effects are published under a probability of 1, and further estimates that marginally significant effects are published at $p_1 = 1.14$ ($\text{CI}_{95} = [.86, 1.43]$), nonsignificant effects at $p_2 = .16$ ($\text{CI}_{95} = [.11, .21]$), and “surprising” effects at $p_3 = .03$ ($\text{CI}_{95} = [.00, .06]$). With these estimates of publication bias, the average effect was estimated at Hedge’s $g = 0.21$ ($\text{SE} = 0.07, \text{CI}_{95} = [0.06, 0.35]$), with heterogeneity $\tau = .53$ ($\text{CI}_{95} = [.45, .60]$); a significance test against the null ($z = 2.84, p = .005$) suggests that the population average effect is nonzero, consistent with what we find in the p-curve analysis. This effect size estimate via the selection
model roughly corresponds to a correlation of $r = .10$, a much smaller effect size compared to the uncorrected estimate.

**PET-PEESE.** Applying the PET model surprisingly yielded a negative estimate for the fixed effect of the MS hypothesis, Hedge's $g = -0.08$ (SE = 0.04, CI₉₅ = [-0.15, -0.01]), which was significant, $t(685) = -2.19, p = .029$. Interpreted directly, this would suggest the true MS effect is opposite to the researchers’ hypothesized direction. However, this counterintuitive finding may be explained by the fact that PET is prone to producing imprecise and biased estimates under high heterogeneity, small sample sizes, and a large proportion of $p$-hacking (Stanley & Doucouliagos, 2014; Stanley 2017; Carter et al., 2019). A negative estimation from PET may emerge, when researchers who are more confident in their effects opt for smaller sample sizes while engaging in $p$-hacking, resulting in larger effects with large standard errors. This may lead to a stronger positive association between standard error and the effect size than it would otherwise naturally occur, biasing the intercept (and hence the effect size estimate) downward. Regardless of the reason for this observation, we follow the recommended practice (Stanley & Doucouliagos, 2017): Since we do not find a significant positive effect under PET, we do not interpret the PEESE results, and conclude that the PET-PEESE procedure provided no evidence in support of an MS effect.

**WAAP-WLS.** Under WAAP-WLS, 89 out of the 687 effects were considered adequately powered (>80%), which means our analysis will only include 13% of the effects collected. Nevertheless, since there is more than one adequately powered effect, we proceed to report the WAAP estimate: Based on 89 studies, we found an estimate of $g = .30$ (SE = 0.02, CI₉₅ = [.25, .35]), $t(88) = 12.76, p < .001$. This estimate, roughly corresponding to $r = .15$, is slightly larger the selection model estimate, as we would expect from the most liberal correction we have applied. Nevertheless, this is notably smaller than the uncorrected estimate of $r = .31$.

**General Discussion**

We have examined the evidential value of the MS hypothesis from a variety of perspectives, using many current and championed meta-analytic tools. We reviewed past efforts to replicate a single test of the MS hypothesis using high-powered and pre-registered replication efforts as well as a multisite replication effort. We also conducted a $p$-curve analysis, a $z$-curve analysis, and a classic meta-analysis with bias-corrected selection models, PET-PEESE, and WAAP-WLS of the entire published literature of studies (between 687 and 826 effects, depending on the analysis). We also conducted these same analyses for several subgroups of studies which are reported in the SOM. We now draw some conclusions from our efforts.

We went into this analysis with the hope that the various tools for assessing the evidential value of the MS hypothesis would converge and point to some straightforward conclusions. The most striking finding from our analyses is how little convergence there was across the different analytic tools. This divergence emerged both for analyses of the entire universe of studies as well as for analyses of smaller subgroups.
of studies which are reported in the SOM. This divergence likely emerges because each analytic tool speaks to a different aspect of the state of the literature, and taken in isolation, may point to a conclusion that is at odds with ones that would be drawn from other tools. The advantage of the multitool approach that we have taken here is that we can consider the results from all of the tools at once. We can evaluate the different findings knowing the respective tradeoffs for each of these approaches, and see how the results compare across the various tools. This multitool approach can help to paint a larger, and more comprehensive picture, where we can assess the evidential value in different ways. We describe the results from each analytic tool below and then attempt to integrate these results into some take-home conclusions.

We begin our discussion by considering the most extreme estimates. The traditional meta-analysis was the most optimistic of the MS hypothesis, pointing to a moderate to large effect size concurring with the previous findings by Burke et al. (2010). However, we note that traditional meta-analyses provide inflated estimates of effect sizes in the presence of publication bias and p-hacking. This is likely the case here as well. We do not recommend that readers accept the estimated effect size from the traditional meta-analysis at face value, and we do not draw any conclusions from it below.

On the other extreme, the PET-PEESE yielded the most pessimistic findings, even providing evidence for a significant negative effect. Taken at face value, this suggests people engage in significantly less worldview defense when their mortality is made salient than when it is not. However, when effects are highly heterogeneous and when sample sizes are small, the PET-PEESE analysis is known to yield imprecise estimates (Stanley & Doucouliagos, 2014; Stanley, 2017; Carter et al., 2019). Further, PET-PEESE can underestimate the effect size when there is publication bias and when p-hacking is present (Carter et al., 2019), which are two conditions likely affecting the literature. We suggest that the PET-PEESE estimate of a significant negative effect is likely spurious, but that the absence of a significant positive effect under PET suggests this procedure finds no evidence supporting the MS hypothesis.

Z-curve's unconditional power estimate is also high pessimistic, as it is practically at the lowest theoretically possible value, suggesting that the overall average MS effect, when we consider deviations (such as conceptual replication) from the most successful designs, may be trivial in magnitude.

The rest of the analyses pointed to conclusions that are somewhat more optimistic. While the z-curve analysis also yields a very low conditional power estimate, it was notably higher than the unconditional power estimate. Conditional power differs from unconditional power in that it applies only to the specific study designs and environmental factors (such as participant demographics) of studies included in our universe of studies. The higher conditional power estimate therefore suggests some evidential value in published studies that yielded significant findings. This finding is echoed by the p-curve analysis, which yielded a similar but high conditional power estimate.

The selection model provides evidence for a small effect consistent with the MS hypothesis. This estimate for the effect size is slightly lower than that provided by WAAP-WLS, as expected from an approach that tends to be more liberal. We suggest that the average effect of the literature may be within
the range estimated by the selection model and WAAP-WLS (i.e., \(r\) is between .10 and .15), although this average may have resulted from a mix of effects, many of which higher than .15, and many of which lower than .10.

**On Replication Efforts**

Past efforts to conduct pre-registered and high-powered replications of selected individual studies testing the MS hypothesis also point to mixed conclusions. Of the 6 relevant studies that we found, 3 of them successfully replicated the original study (Dunn et al., 2020; Schindler et al., 2019; Vail et al., 2019), whereas 3 of them did not (Rodríguez-Ferreiro et al., 2019; Sætrevik & Sjåstad, 2019; Schindler et al., 2021). Among the 3 that did not replicate, the findings were in the same direction as predicted in one of the studies (Rodrigues-Ferreiro et al., 2019) whereas the other two yielded mixed results across various hypotheses and moderators (Sætrevik & Sjåstad, 2019; Schindler et al., 2021). Given that the power estimates of these replication attempts were based on the effect sizes that were reported in the original studies (which our various meta-analyses of the literature indicate were likely inflated), this raises the possibility that the pre-registered studies would have been more successful had they relied upon more conservative estimates of the effect size. We encourage future preregistered replications of the MS hypothesis to use smaller estimates of effect size (i.e., \(r = .10 \sim .15\)).

A multisite replication effort (Klein et al., 2022) overall found little evidence that the selected study replicated, however, a re-analysis (Chatard et al., 2020) that focused only on the studies that met preregistered criteria found significant evidence for replication when the studies followed the guidance of the original authors and one set of exclusion criteria were applied. In contrast, other sets of exclusion criteria in this re-analysis did not reveal evidence of replication. Overall, the multisite replication effort did not find encouraging results for TMT, which is perhaps similar to the discouraging results this method has identified for other social psychological studies (see Baumeister et al., in press).

**Multitool Assessment**

To summarize our findings, the respective analytic tools point to different conclusions that likely reflect the differences in the philosophies and methodologies of each analytic tool (see Table 1 for an overview). This is despite that many of these tools were directed at the same, extremely large, set of published studies. Initially, we had assumed that focusing our efforts on a very large literature would provide more convergence across methods than what one would expect for a small literature as it would seem the large literature would be less likely to be affected by sampling bias. Yet the findings point to a lack of convergence across analytic tools which demands careful and critical synthesis of the evidence. The multitool approach that we have employed here provides a useful means for interpreting the findings because we can compare across the different methods. Researchers who attempt to assess a literature with only one of these meta-analytic tools is only going to see one perspective, and one that is limited by the respective tradeoffs of the given tool. Given the conflicting findings that emerged across tools, we caution researchers against drawing firm conclusions about the evidential value of a literature through any individual analytic tool.
Meta-analyses, like the ones we conducted, typically strive to be comprehensive by including all possible tests of a hypothesis, even those that might have been poorly designed. However, in a literature that is as large and diverse as TMT, such an average may not be informative of the typical study. Indeed, an average across a sample that includes both well-designed and inadequately-designed TMT studies may be akin to calculating an average of both real effects and false ones, leading to a misrepresentation of the true average effect. One recommended solution is to critically evaluate each paper in terms of its quality, and draw a subjective conclusion about whether to include those results in a meta-analysis (Simonsohn et al., 2022). Given the enormous number of studies that we investigated here, this was not a feasible option for us.

We attempted to address the problems that arise from averaging across a diverse literature in two ways. First, we conducted sensitivity analyses for our \( p \)-curve analysis which found it to be fairly robust to the removal of a large number of the studies with the largest effect sizes. Second, we conducted all of our analyses on separate subgroups of studies that are reported in the SOM, although we note that the different analytic tools did not converge in their conclusions for these smaller subgroups of studies either.

We released a Shiny application to accompany this paper which enables users to define their own subsets of data based on any combination of moderators. Our intention is to provide an open, interactive implementation of the multiverse analyses (Steegen, Tuerlinckx, Gelman, & Vanpaemel, 2016), to allow readers to critique our conclusions as well as to draw their own. We acknowledge this approach may facilitate HARKing if the application is used in an exploratory way; therefore, we strongly encourage people to use the application with caution, and avoid framing findings from exploratory investigations as confirmation of their \textit{a priori} predictions. We hope that responsible use of this application will enable the research community to assess whether any particular conclusion is sensitive to arbitrary researcher decisions. We have made the open source code for this app available on the OSF, and we encourage researchers to adapt this code to allow them to analyze any possible subset of these studies to assess how robust TMT effects may be to various standards of research quality.

As we mentioned above, one reason why the tools we employed yielded different answers is that they are designed to answer slightly different questions. Below we consider two questions that the respective analytic tools speak to. For each we provide a liberal answer (in support of the MS hypothesis), a conservative answer (against the MS hypothesis), as well as an overall conclusion about the literature.

**Question 1: Is the Underlying Effect of the MS Hypothesis Real?**

From a theoretical standpoint, this is perhaps the most important question to consider. When people are reminded about their mortality, do they tend to engage in worldview defense? Evidence that provides an affirmative answer to this question comes from a variety of the measures reviewed above. The \( p \)-curve reveals that there is significant evidential value for this literature, and the selection model and WAAP-WLS also identified a significant overall effect. The \( z \)-curve's calculation of the conditional power for this literature also suggests overall evidential value. In addition, 50% of the preregistered highly-powered studies that tested individual studies of the MS hypothesis replicated this effect. Moreover, the question
of whether there is an underlying real effect should be best addressed by studies that are most appropriately designed to test the MS hypothesis. Studies that were published after many methodological reforms were beginning to be introduced in 2011 revealed significant and adequate evidential value by the \( p \)-curve, and nominally higher estimates of power by the \( z \)-curve. Likewise, subgroups of studies that adhered most closely to the underlying theoretical arguments of the effect (longer delays enhance the effect; Pyszczynski et al., 1999), revealed significant and adequate evidential value by the \( p \)-curve, nominally higher estimates of power by the \( z \)-curve, and nominally stronger effect size estimates by the selection model. Moreover, in the multisite replication effort (Klein et al., 2022), the effect size estimates were nominally higher in the author-advised locations than in those that followed an in-house protocol (and were significantly different from zero with one particular set of exclusion criteria for those studies that met all preregistered criteria; Chatard et al., 2020), which is also consistent with the trend of stronger effects for studies that were better designed.

On the other hand, evidence that provides a negative answer to this question includes the \( z \)-curve’s estimate of unconditional power, with a narrow confidence interval that overlaps with the unconditional power under a null effect. Likewise, the PET-PEESE analyses did not find significant evidence for the MS hypothesis in any of the subgroups that were analyzed. Moreover, the multisite replication effort did not successfully replicate the targeted study (Klein et al., 2022), and 50% of the preregistered high powered replication studies failed to replicate the original effect.

Though on their face the liberal and conservative interpretations feel contradictory, some observations are uncontroversial. The first observation is that the TMT literature consists of highly heterogeneous effects. For our entire universe of studies, heterogeneity is estimated at \( \tau = .53 \) under the selection model, which means that for the selection model estimate of \( g = .21 \) (\( r = .10 \)) for the entire literature, 95% of the effects underlying studies of MS hypothesis fall between \( g = -0.83 \) and 1.25 (or \( r = -.38 \) and .53); an extremely wide range of possible effect sizes arising from differences in study design. The second observation is that a large majority of studies in the TMT literature are underpowered. The \( p \)-curve estimates the average power of the significant findings to be 26%, roughly 1 out of 4. Even if we disregard the much lower power estimates from the \( z \)-curve (especially for unconditional power), this optimistic \( p \)-curve power estimate would suggest that for every published significant effect, on average, there are three studies with the same design that did not lead to a significant finding. Further supporting this observation is the WAAP-WLS estimate that only 13% of the simple main effects we included in our meta-analysis were associated with adequate power (80% or higher). Though the overall power of the TMT literature is low, at present we are unable to determine whether it is uniquely low in comparison to other social psychological phenomena. For example, a recent hand-coded \( z \)-curve analysis of the entire literature of system justification found an ERR of .18 (Sotola & Crede, 2022), which is similar to the ERR obtained for TMT (.19). Going forward, we hope that other literatures will also be hand-coded and assessed using this multitool approach such that we can compare the relative evidential value for different psychological phenomena.
With these observations in mind, we reconcile the liberal and conservative interpretations by concluding that there must be some true underlying effects in the studies we examined, yet overall, the heterogeneous set of significant findings in TMT likely include either inflated estimates or false positives, due to a combination of insufficient statistical power and publication bias. In direct replication attempts, we find that some MS effects appear to be highly contingent on strict adherence to the protocols and expert advice. The z-curve’s estimate of 8% unconditional power leads to the bleakest version of this interpretation of our observations, that the vast majority of studies on the MS hypothesis are either investigating null effects or are so underpowered that they are unable to meaningfully advance our knowledge in the hypothesis.

**Question 2: Are TMT Studies Likely to Replicate?**

This is an important question for researchers trying to reliably capture an effect of MS. If a study is conducted that is a direct replication of a published TMT study, how likely is it that it will replicate? On the liberal side, preregistered highly powered studies suggest effects will replicate roughly half the time, and this may even be an underestimate given that these studies assumed a larger effect than our meta-analyses suggest, rendering some of these replication efforts underpowered. Our p-curve and z-curve analyses of the subgroups of studies suggest that replications will be more successful if they include at least 2 distractors and employ the new methodological conventions that have emerged since 2011. The reanalysis of the multisite replication effort by Chatard et al. (2020) also suggests that a study will be more likely to replicate if the authors receive more detailed guidance from the original researchers.

On the conservative side, average conditional power of studies that lead to significant MS effects is very low at 19-26% (as estimated by the p-curve and z-curve). However, even our more conservative tests suggest replicability might increase if researchers plan for smaller effects. The average per cell sample size of past MS studies is around 28, but a much larger per cell sample size of \( n=200 \) should theoretically produce more successful replications. That said, we must keep in mind the heterogeneity of the effects and the low unconditional power. Not all MS effects are equal, and researchers should expect that many of the studies in the literature are unlikely to be successfully replicated. Further, conceptual replication of TMT effects, under the current state of the literature, are less likely to succeed without expert involvement.

Reconciling the more liberal and more conservative interpretations of the literature’s reproducibility hinges on planning the right replication study. We advise that researchers conduct direct replications of well-planned studies, seek expert advice where possible, and anticipate smaller effect sizes than what is documented in the literature (we suggest \( r = .10 \sim .15 \)). Though we suspect that the high heterogeneity obscures the magnitude of any true underlying effect in the universe of studies, we encourage researchers to conduct their own subgroup analyses of our data to benchmark how much power they can expect in their planned studies.

**Conclusion**
Our efforts to assess the evidential value of TMT point to two main conclusions. First, the literature investigating the MS hypothesis contain studies that appear to be testing real effects, although the literature is highly heterogeneous and underpowered, rendering many of the individual effects to be likely spurious. The size of the underlying effect is considerably smaller than the published literature suggests, and any efforts to replicate these effects would fare best by expecting a small effect size and following expert guidance. Second, our analysis suggests that any attempts to evaluate other psychological literatures would benefit by employing a multitool assessment approach, as a reliance on any single tool may lead to conclusions that are at odds with those that would have been drawn by using a different tool.

Methods

Our meta-analytic study consisted of the following steps. First, we searched for studies related to the MS hypothesis, and selected viable ones based on a set of pre-specified criteria adapted from Burke et al. (2010). Next, we chose effects in those studies that are appropriate for each analytic technique and hand coded them into disclosure tables containing relevant statistics (such as t-values, F-values, degrees of freedom, group means, group standard deviations), which are available on OSF. For the p-curve and z-curve analysis, we selected effects for focal hypotheses following the p-curve selection criteria given by Simonsohn et al. (2014a). For the other meta-analytic techniques, we selected simple main effects following selection criteria adapted from Burke et al. (2010). A team of research assistants were trained by the authors of this paper to independently perform the hand coding, while consulting the authors where the applications of the rules are unclear. To ensure that these rules were applied consistently, a research assistant and at least one author on this manuscript reviewed each hand-coded effect. Then, we conduct meta-analyses on the coded effects. P-curve and z-curve analyses were performed in the R statistical programming language (version 4.2.3; R Core Team, 2021), using code provided by their authors (Simonsohn et al., 2014a; Bartoš & Schimmack, 2020). For the traditional meta-analyses, including the selection model, PET-PEESE, and WAAP-WLS, we use R to convert statistics of the selected effects were converted into Hedge's g, which were then analyzed using JASP (JASP Team, 2022). The code for the p-curve and z-curve analyses, effect size calculations, as well as the JASP analysis files for the selection model, PET-PEESE, and WAAP-WLS, are all available on OSF. While we did not formally register our review protocol, we used our analysis plan on OSF as well as previous studies to guide our review process, and a PRISMA checklist (Page et al., 2021) is also available on OSF. Below we provide technical details to these study steps.

Study Search and Screening

We followed the study search and selection procedures used in Burke et al. (2010), adapting it for the p-curve analysis where necessary. We conducted a database search on PsycINFO using terror management theory and/or MS as key phrases. We also obtained papers from reference sections of prior reviews, and from a well-established TMT website (www.tmt.missouri.edu). This process yielded 1373 unique
manuscripts, many of which contained multiple studies. We then screened papers that were not (1) published in peer-reviewed journals, (2) available in English, or (3) reporting novel empirical tests of the MS hypothesis (i.e., no review papers or meta-analyses). We did not include studies published exclusively as part of a dissertation or thesis. We screened 527 manuscripts, then after closer review we excluded an additional 174 on the basis of these inclusion criteria, totaling 701 manuscripts (see Figure 4). The remaining manuscripts contained 1396 studies which were subjected to stricter inclusion criteria.

**P-curve and Z-curve Analyses**

We applied the same inclusion and selection rules for the p-curve and z-curve. Since p-curve has provided prescriptive rules on which studies can be included and which statistics should be selected, we followed the p-curve guidelines closely.

**Study Inclusion Criteria**

A full list of excluded and included studies is available on OSF. To be included in the analysis, studies needed to (1) test the MS hypothesis on distal defenses (this means that proximal defense and death-thought accessibility dependent measures do not qualify); (2) report a true experimental manipulation of MS with random assignment to condition; and (3) be compatible with the p-curve method such that at least one statistical test, using statistics accepted by the tool, that was unambiguously attached to the primary hypothesis was \( p < .05 \). Because the MS hypothesis follows specific rules, some studies that otherwise might have been included were not; for example, those whose primary predicted effect was a proximal effect and not a distal one were excluded (e.g., Bessarabova & Massey, 2019, study 2). Also, because of the criterion that the predicted effect is p-curvable, studies whose primary prediction was that the effect was mediated were also excluded (e.g., Agronskin & Jonas, 2013, study 2).

Altogether, we excluded a further 578 studies that did not meet these criteria, leaving 826 studies with complete statistical tests that were suitable for the p-curve.

**Effect Selection Rules**

For each study, we selected one or two effects associated with the focal hypothesis. The fundamental considerations of the p-curve are that p-values correspond to tests of the focal hypothesis, and that they have been published under the criterion of \( p < .05 \). If the focal hypothesis involves an interaction, consideration must be made of where this \( p < .05 \) criterion applies: When the interaction is attenuated – that is, the associated simple main effects are all in the same direction – the p-value associated with the interaction is selected (see Simonsohn et al., 2014b). This is because an attenuating interaction is always smaller than the largest simple main effect. Therefore, if a paper is deemed publishable because an attenuating interaction yields a \( p < .05 \), the accompanying simple main effect essentially contains a p-
value that was selected based on a stricter criterion. In contrast, when the interaction is reversing; that is, the associated simple main effects are in different directions; the \( p \)-values associated with both simple main effects are selected. This is because the reversing interaction is always greater than the simple main effects. The full description of the selection rules can be found in the \( p \)-curve user guide (see Simonsohn, Nelson, & Simmons, 2015).

We made small deviations from the \( p \)-curve authors’ guidelines, which are disclosed in our OSF document\(^1\). Most notably, when the most appropriate test was not reported (e.g., reversing trends should be reported in the case of \( 2 \times 3 \) interactions), we reported what the authors chose as the test of their key prediction; these were either interaction terms or main effects, depending on the nature of the hypothesis. Also, when reversing interactions were predicted, we reported two simple effects in accordance with recommendations from Simonsohn et al. (2015), unless the reversal was reported separately for the control and MS conditions (not separated by levels of the interaction variable) in which case the simple effect in the control condition did not test the MS hypothesis. Here, we reported the interaction term. We made these deviations unless the context of the reporting suggested the publication criterion was not \( p < .05 \). Our goal was to be more inclusive of the entire literature, and to select as many \( p \)-values as possible that are associated with the main hypotheses of published papers. Last, when the paper reported tests for two or more non-independent hypotheses that were (1) equally prominent in the paper, (2) equally likely to lead to publication, and (3) equally related to the MS hypothesis, we only included the first listed test in the results section.

Finally, in accordance with the \( p \)-curve procedure by Simonsohn et al. (2015), to account for any inaccurate and incorrect reporting, all \( p \)-values were recomputed using the test statistics (e.g., the \( t \)-statistics and the degree of freedom). Any study with recomputed \( p \)-values of .05 or above were excluded from the analyses. We provide an overview of the \( p \)-curve analyses in the SOM.

**Overview of \( p \)-curve Analyses**

**Inferential tests for evidential value.** The shape of the \( p \)-curve provides a visual diagnostic for the evidential value of the body of literature; a right skew suggests that at least some studies have underlying true effects. In order to test the shape of \( p \)-curve statistically to control for sampling error, Simonsohn et al. (2015) proposed two robust Stouffer’s Z-tests, including a test for right skew and a test for flatness.

**Test for right skew.** When the \( p \)-curve appears to have at least some right skew visually, the right skew test can be used to statistically examine this conclusion. The Stouffer’s method is used to compute a \( z \)-score on the deviation from the flat distribution. If the \( p \)-value associated with this \( z \)-score is significant, we conclude that some findings in our sample have underlying nonzero effects.

**Test for flatness.** In the absence of significant right skew, the flatness test is used to obtain significance for inadequate evidential value. Simonsohn et al. (2014a) define a sufficient right skew as one generated by studies with at least 33% power. Similar to the skewness test, the flatness test produces a \( z \)-score that
quantifies deviation from 33% power. An associated \( p \)-value less than .05 indicates that, even if an effect were to exist, the extant body of literature is too insufficiently powered overall (significantly below 33%) to provide satisfactory evidence for this effect, and direct replications are unlikely to succeed.

**Binomial tests for right skew and flatness.** In addition to these continuous tests, the \( p \)-curve also conducts binomial tests of right skew and flatness. Measuring right skew, the binomial test evaluates if the proportion of \( p \)-values above and below <.025 is significantly different from what would be expected if power was zero (because the distribution is flat, this value is 50% in both bins). Measuring flatness, the binomial test compares the distribution below .025 to one expected if power was 33% (approximately 71%). The binomial test is intuitive, but is also insensitive to the distribution of \( p \)-values within each bin, and therefore is less diagnostic than the continuous tests (see Simonsohn et al., 2014a).

**Robust \( p \)-curve analysis.** A critical assumption of the \( p \)-curve is that the publication criterion is based on \( p < .05 \). Ulrich and Miller (2015) posited that the \( p \)-curve inferential tests may be biased by ambitious \( p \)-hacking: the act of selectively reporting based on criteria that are more stringent than \( p < .05 \) (e.g., reporting all results where \( p < .04 \), or simply always reporting only the lowest \( p \)-value). As a sensitivity analysis to account for moderately ambitious \( p \)-hacking such as \( p < .03 \) and \( p < .04 \), Simonsohn et al. (2015) implemented additional inferential tests based on the criterion of \( p < .025 \), known as the half \( p \)-curve analysis. These additional inference tests are similar to the right skew and flatness tests in the main analysis, except they are performed only on studies with \( p \)-values below .025.

For a robust test for right skew, if the half \( p \)-curve test yields \( p < .05 \), or if both the half \( p \)-curve and the full \( p \)-curve yield \( p < .10 \), we consider there to be significant right skew, and take the result as evidence of the presence of evidential value. For a robust test for flatness, if the full \( p \)-curve yields \( p < .05 \) or the half \( p \)-curve and binomial tests both yield \( p < .10 \), we consider the curve to be significantly flatter than one that has 33% power, and take the result as evidence for inadequate evidential value.

**Power and effect size estimates.** A distinguishing feature of the \( p \)-curve is that it can be used to aggregate studies with different designs and analyses. A direct consequence, however, is that \( p \)-curve typically does not produce a straightforward average effect size estimate for the studies included. In place of an effect size estimate, the \( p \)-curve produces an estimate of the average power of all the studies in the literature by fitting the average power most likely to produce the observed \( p \)-curve (Simonsohn et al., 2014b).

**Between-study moderators.** We point to some between-study moderators in our analysis plans on OSF (note we did not make specific predictions) and included others that may impact estimated power, which we identified based on theory and past meta-analytic evidence (see Burke et al., 2010; Pyszczynski et al., 2015). As results from these additional analyses did not substantially alter our primary findings, they are not reported here, and further details are provided in the SOM.

**Accompanying application and disclosure table.** We developed a Shiny application via the shiny package (Chang et al., 2020) in the R statistical language (R Core Team, 2021). This application loads the
disclosure table that will enable readers to explore \( p \)-curves of subgroups which, for methodological and theory-driven reasons, might have more homogeneous effect size estimates. The code of the application, the requisite data, as well as instructions on how to run it offline, are available on OSF\(^8\). Once again, we caution against drawing conclusions after making multiple comparisons, which can inflate Type 1 error rates. We encourage users to pre-register their predictions and their theoretical or methodological reasoning (even privately) before using the accompanying app. We also included a \( p \)-curve disclosure table on OSF\(^8\) (this is a formatted table with all available information needed to derive the \( p \)-curve).

**Overview of Z-curve Analysis**

We conducted the \( z \)-curve analysis using the `zcurve` package (Bartoš & Schimmack, 2020) in the R statistical programming language. First, we converted the \( t \), \( F \), and \( \chi^2 \) statistics we collected for the \( p \)-curve analysis into \( z \)-scores using a function provided by the package. In order for the procedure to accept \( p \)-values as input, the package automatically converted a negative \( z \)-scores to positive. This is akin to treating all “surprising” findings, where the mean difference is in the opposite direction as the original prediction, as predicted findings; it does not mathematically affect the critical statistic used by the \( z \)-curve for power estimation (Bartoš & Schimmack, 2022). Next, the \( z \)-curve analysis modeled the population effects using a mixture of \( K \) normal distributions, where \( K \) is the number of studies included, and this mixture model was fitted to observed \( z \)-scores using either a density-based approach or an EM algorithm (for mathematical details, see Bartoš & Schimmack, 2022). From the best fit model, three estimates were produced: 1) the estimated unconditional mean power to obtain similar effects in the literature, called the *expected discovery rate* (EDR), 2) the proportion of published studies that are significant, the *observed discovery rate* (ODR), 3) the conditional mean power associated with published significant effects, called *expected replication rate* (EDR), which is the probability of the exact replication of observed significant effects, analogous to the \( p \)-curve power estimate.

Under no publication bias, EDR and ODR should be equal. The discrepancy of ODR and EDR would show the degree of selective publication – the larger the difference, the more severe the publication bias. However, since we applied \( z \)-curve to our \( p \)-curve data, all observed results were significant, and the ODR was not informative (the sample ODR is always be equal to 1 when only significant effects are included). We therefore focused our analysis on the EDR, the estimated unconditional power to discover MS effects in the literature overall. If the \( z \)-curve EDR estimate agreed with the \( p \)-curve average power estimate, it would lend support to generalizing the \( p \)-curve findings to variations of MS study designs not included in our analysis. Further, \( z \)-curve estimates the expected replication rate (ERR), which is the average power of all the studies that yield significant results. The ERR predicts the average probability of success should a researcher attempt a replication on one of the significant studies. The ERR estimate should be similar to the \( p \)-curve power estimate, as the two estimates target conceptually equivalent population quantities, albeit modeled differently and estimated via different algorithms.
Traditional Meta-Analysis

The remaining meta-analytic tools rely on the logic of traditional meta-analysis which led us to conduct separate independent coding following a procedure adapted from Burke et al. (2010). Below we review deviations from the $p$-curve coding procedure, and how they changed inclusion and selection criteria.

**Effects Selection Rules**

We were able to include in our classical meta-analysis 120 statistical tests that could not be in the $p$-curve and z-curve; of these, some had obtained nonsignificant primary predicted effects (e.g., Norenzayan et al., 2009, study 2) and others reported primary predicted effects that could not be subjected to the $p$-curve, though they included a straightforward experimental vs. control statistic (e.g., Lambert et al., 2014, study 2). However, 285 statistical tests included in the $p$-curve and z-curve could not be included in the sensitivity analyses for various reporting idiosyncrasies (e.g., when simple effects were reported within MS and control conditions, and not at the level of the moderator, as in Magee, Robert, & Woljdymski, 2012). The core difference that caused these inconsistencies with the $p$-curve is that classical meta-analytic methods require comparable effects to be reported in full. If authors did not report main or simple effects that included an MS experimental vs. control comparison, those studies were excluded. An effect was included in our meta-analysis if it fully reports any set of the following information: (1) The means, standard deviations, and sample sizes of the dependent variable in the experimental and control groups, (2) The t-test statistics and associated degree of freedom for the mean difference between the experimental and control groups, or (3) the F-test statistics and associated degrees of freedom for the simple main effect containing only the experimental and control groups. We then excluded four studies with effect sizes greater than $g=10$, for a total of 687 statistical tests (note that the traditional meta-analysis often included more simple effects per study than the $p$-curve and z-curve did, so these statistical tests correspond to 564 individual studies).

Our strategy for avoiding data dependencies as in the case of multiple dependent variables were consistent with the $p$-curve rules. Also consistent, we selected the most important test of the primary hypothesis when choosing between main effects and interactions. A key difference for interactions is that we reported simple effects, regardless of whether authors predicted attenuating or reversing interactions, but excluded the predicted null statistics in attenuating interactions. This decision enabled us to avoid describing predicted null and significant effects as comparable when authors’ rationale for making these distinctions was theory-consistent; therefore, the effects would not be comparable tests of the MS hypothesis. We extended this logic to more complicated models (e.g., three-way interactions) so we included independent simple effects that were not predicted to be null. Consistent with the $p$-curve, when authors predicted main effects, we reported statistics for comparisons between two groups only; this means that for more complex models we chose experimental and control conditions that were most key to the authors’ hypotheses. Or, when conditions were equally important, we chose the first reported comparison, or the most conservative control condition (e.g., if controls were dental pain and uncertainty
salience, we chose uncertainty salience because it was designed to be more similar to the MS prime). Consistent with the p-curve, at least one author on this manuscript reviewed each statistic we included.

**Overview of Bias Correction with Traditional Meta-Analysis**

**Effect size calculation.** For the traditional meta-analyses, we first converted the coded statistics into Hedge’s $g$ effect sizes and their associated standard errors (see Lipsey & Wilson, 2001), using the *metafor* package (Viechtbauer, 2010) and the *esc* package (Lüdecke, 2019). We re-coded negative effects the authors predicted to be in the positive direction, while re-coding positive effects that run in contrary to the author’s prediction as negative. Unlike in p-curve and z-curve, the direction of effects influences the aggregate effect size of traditional meta-analyses. Without this treatment, we may have obscured a true underlying effect; for example, by directly aggregating large positive and large negative effects all supporting the MS hypothesis. Conversely, we negatively coded positive effects that were in the opposite of the hypothesized direction. We are aware that this approach can be adversely affected by HARKing, as researchers may reformulate their hypothesis to the direction of the observed effect after it has been observed, introducing bias into the meta-analysis. However, without pre-registration of directional hypotheses, we are not aware of a better solution to this problem.

**Selection model.** We conducted selection model on the computed Hedge’s $g$s and standard errors using the JASP software (version 0.16.4; JASP Team, 2022; also see tutorial by Bartoš, Maier, Quintana, & Wagenmakers, 2022). By default, the software defaults to two-tail thresholds of .05 and .10, which can be described as: 1) one-tail $p < .025$, corresponding to two-tail $p < .05$, the significant findings, 2) one-tail $0.025 \leq p < .05$, corresponding to two-tail $0.05 \leq p < .10$, the “marginally” significant findings, 3) one-tail $0.05 \leq p < .05$, the nonsignificant findings, and 4) one tail $p \geq .95$, the “surprising” findings. We did not deviate from these defaults in our analyses, because we have no evidence or *a priori* reasoning that the publication standards of the TMT literature would be better described by an alternative specification.

While we expected heterogeneity effects due to the wide range of studies the MS literature contains, we followed the test of heterogeneity (Cochran’s Q) conducted by JASP. In the event of a nonsignificant test result, we would follow the fixed effects estimates; whereas, if the heterogeneity test was significant as expected, we would proceed to use the random effects estimates. JASP outputs unadjusted estimates (traditional random effects model without bias correction) and adjusted estimates (selection model), which we would both report in the appropriate sections.

**PET-PEESE.** We performed the PET-PEESE analysis on the same set of effect sizes and standard errors using JASP. Following the recommended procedure outlined by Stanley and Douliagos (2014), we would first examine the $t$-test for the weighted meta-regression intercept of the PET model predicting effect size with the standard error. If the $p$-value was below .10 and the effect was positive, we would interpret the effect size estimate for PEESE which swaps the predictor to the squared standard error. If instead we found no significant positive effect at the threshold $\alpha = .10$, we would ignore the PEESE output and conclude that there was no evidence in support of nonzero true effect of MS.
**WAAP-WLS.** The WAAP-WLS analysis was also conducted using JASP on the effect sizes and standard errors. In the unlikely event that we did not have more than 2 studies with power of 80%, we would report the WLS effect size estimate. Otherwise, we would proceed to interpret the WAAP effect size estimate and its associated statistical test.

**Declarations**

Competing interests: The authors declare no competing interests.

**Acknowledgments**

We are extremely grateful to the large team of research assistants who painstakingly hand-coded the immense number of papers that made up this review. Thanks to Sarah Ponton, Francis Yuen, Charul Maheshka, Rohin Gupta, Radha Bazaz, Emelia Catling, Edwin Chang, Jeneesha Dhaliwal, Yau Fan, Vanessa Kong, Raymond Li, Diana Lim, Diana Lee, Francesca Oldham, Paniz Radjaee, Sarah Rosenbloom, Julio Silva, Gabriel Suessenbacher-Fu, Fiza Syal, Stephanie Yu, and Fares Zubari. This project was funded by grants from the Social Sciences and Humanities Research Council 435-2019-0480 and 430-2020-00754 to Heine.

**References**


28. JASP Team (2022). JASP (Version 0.16.4) [Computer software].


**Footnotes**

1. We initially planned the study as a *p*-curve analysis, but later expanded it to include a diverse range of approaches that have shown good performance in past simulation studies, under the helpful suggestion of our editor and reviewers.

2. For convenience and consistency with the app, we refer to *pp*-values as *p*-values in our results.

3. “Statistical power” commonly refers to the probability of obtaining a significant result under a nonzero population true effect. However, since a literature contains studies investigating an unknown mix of nonzero and null effects, *unconditional power* in this context refers to the overall probability of obtaining a significant effect from this mix of effects. *P*-curve does not explicitly model a mix of effects, and its authors discuss power in terms of conditional power (ERR).

4. This analysis was first conducted and shared in a Twitter post by Dr. Uli Schimmack, a primary author of the *z*-curve, in response to a preprint of an earlier version of this article.

5. Likewise, one critic of PET-PEESE stated: “In social psych, I informally count 13 published uses of PET, a handful of submitted versions, and in every single instance I've found, PET has returned an effect size estimate that doesn't differ from zero” (Gervais, 2016).

6. As expected, Cochran's Q-test suggested the presence of heterogeneity, *Q*(687)=3976.68, *p* < .001, which means random effects model, rather than fixed effects model, is appropriate here.

7. Available at https://tmtpcurve.shinyapps.io/main/

8. Available at www.p-curve.com/

9. See https://osf.io/rdsm4/?view_only=9da1be8b55e7450e9c7ce82356d5b7b9

10. We also excluded studies whose primary hypothesis required a test of mediation, and no other result would have plausibly explained why the study was published, as there is no straightforward way to
include a mediation model in the p-curve (which requires a single test statistic, whereas indirect effects are typically evaluated only confidence intervals due to nonnormality).

11 We originally planned to analyze subgroups of sample sizes; instead, we present estimated effect sizes based on each subset’s average N and number of cells

12 Technically, “surprising” findings are also separated into two groups, \( .95 \leq p < .975 \) and \( p \geq .975 \), but due to the low number of studies in these groups, they have been combined.

**Table**

**Table 1** *Summary of conclusions across meta-analytic and bias correction tools.*
<table>
<thead>
<tr>
<th>Approach</th>
<th>Estimates</th>
<th>Notes</th>
<th>Primary Conclusions</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P</em>-Curve</td>
<td>Conditional power: 26%</td>
<td><em>P</em>-curve does not test for overall average effects, and it is not</td>
<td>At least <em>some</em> published MS effects are real; The average</td>
</tr>
<tr>
<td></td>
<td></td>
<td>meant to generalize beyond designs included in the analysis.</td>
<td>power of published MS effects is lacking.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z-Curve</td>
<td>Conditional power: 19%</td>
<td>This newer approach has not been investigated by many independent</td>
<td>The average power of published MS effects is lacking; The</td>
</tr>
<tr>
<td></td>
<td></td>
<td>researchers; Initial simulation results by original authors appear</td>
<td>typical MS study (accounting for those unpublished) is</td>
</tr>
<tr>
<td></td>
<td></td>
<td>promising.</td>
<td>expected to yield null or trivial effects.</td>
</tr>
<tr>
<td></td>
<td>Unconditional power: 8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traditional Meta-</td>
<td><em>g</em> = .66</td>
<td>This commonly used method is vulnerable to publication bias and</td>
<td>The average effect is likely an inflated estimate and we don't</td>
</tr>
<tr>
<td>Analysis</td>
<td></td>
<td>aggregation of <em>p</em>-hacking efforts.</td>
<td>recommend accepting it at face value.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Selection Model</td>
<td><em>g</em> = .21</td>
<td>Estimates can be sensitive to the selection model specification,</td>
<td>There is a small overall MS effect (roughly <em>r</em> = .10).</td>
</tr>
<tr>
<td></td>
<td></td>
<td>which may not be able to adequately account for <em>p</em>-hacking, leading</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>to bias.</td>
<td></td>
</tr>
<tr>
<td>PET-PEESE</td>
<td><em>g</em> = -.08</td>
<td>Estimate can be imprecise when study sample sizes are small and</td>
<td>There is no evidence for an overall MS effect.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>effects are heterogeneous; Estimate can be biased under publication</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>bias and <em>p</em>-hacking.</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WAAP-WLS</td>
<td><em>g</em> = .30</td>
<td>High false positive rates under publication bias and <em>p</em>-hacking;</td>
<td>There is a small overall MS effect (roughly <em>r</em> = .15).</td>
</tr>
<tr>
<td></td>
<td>(Estimated from the 13%</td>
<td>Yields more accurate estimates compared to PET-PEESE when there are</td>
<td></td>
</tr>
<tr>
<td></td>
<td>of effects with 80%+</td>
<td>real effects.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>power)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Replication</td>
<td>Overall 50% success rate</td>
<td>Successful replication dependent on strict adherence to original</td>
<td>Variations on the original design and conceptual replications</td>
</tr>
<tr>
<td>Effort</td>
<td></td>
<td>author's guidelines.</td>
<td>are less likely to succeed.</td>
</tr>
</tbody>
</table>

**Figures**
Figure 1

Potential p-curve distributions using simulated studies

Note. Illustrations of potential p-curves generated by 50,000 simulated experiments with $N = 87$. (a) 5% power (null distribution); (b) 33% power; and (c) 80% power.
Figure 2

P-curve of the entire universe of studies (K = 826)
Figure 3

*Potential z-curve distributions using simulated studies*

*Note.* Illustrations of potential z-curves generated by 50,000 simulated experiments with $N = 87$. (a) 5% power (null distribution); (b) 33% power; and (c) 80% power.
Figure 4

Flowchart of the application of exclusion and inclusion criteria
Figure 5

Z-curve of the entire universe of studies (K = 826)

Note. The error bar indicates the lower and upper bounds of the 90% confidence interval. K is the number of studies included in the estimate, and is the average sample size of those studies.
Figure 6

*Funnel plot of the traditional meta-analysis*

**Supplementary Files**

This is a list of supplementary files associated with this preprint. Click to download.