Assessing the Trustworthiness of our Cumulative Knowledge in Learning, Behavior, and Performance

Authors

Sheila List, Virginia Commonwealth U., lists@vcu.edu
Sven Kepes, Virginia Commonwealth U., skepes@vcu.edu
Michael A McDaniel, Virginia Commonwealth U., mamcdani@vcu.edu
Xavier MacDaniel, Independent Researcher, xaviermacdaniel@gmail.com

Sheila List, Virginia Commonwealth U., lists@vcu.edu
Sven Kepes, Michael A McDaniel, Xavier MacDaniel

Assessing the Trustworthiness of our Cumulative Knowledge in Learning, Behavior, and Performance

ABSTRACT

Meta-analytic studies are the primary way for systematically synthesizing quantitative research findings to cumulate knowledge. As such, they have substantial influence on research and practice. Recently, however, the robustness of results from some meta-analytic studies in management has been questioned. Despite this, very few studies assess the presence and impact of publication bias (PB) and outliers, two factors influencing the non-robustness of meta-analytic results. In this study, we use a comprehensive sensitivity analysis approach to reexamine datasets from nine meta-analyses of correlation coefficients published in *Psychological Bulletin* that were categorized in the area of learning, behavior, and performance. We reexamined 123 distributions from these nine meta-analytic studies. Our results indicate that 88% of the meta-analytic results reported in the nine meta-analytic studies are unlikely to be robust. The degree of the non-robustness was classified as being 'severe' (i.e., > 40%) in 78% of the meta-analytic distributions. These results suggest that most of the meta-analytic results and associated conclusions and recommendations in this area may not be trustworthy. This adds to a growing body of evidence suggesting that our research practices may need to be revised to improve the trustworthiness of our cumulative knowledge in management and the social sciences.

Keywords:

Meta-analysis, outliers, publication bias

Assessing the Trustworthiness of our Cumulative Knowledge in Learning, Behavior, and Performance

The trustworthiness of our cumulative scientific knowledge has come under scrutiny across many areas of science (Goldstein, 2010), ranging from physics, biology, and medicine to the social sciences (e.g., Fanelli, 2010; NSF SBE Advisory Committee, 2015). According to some, our scientific disciplines may be experiencing a crisis of confidence (Earp & Trafimow, 2015; Ioannidis, 2005; Pashler & Wagenmakers, 2012). Numerous factors related to our scientific process, including our publication norms and standards, seem to lie at the heart of the potential crisis (Kepes & McDaniel, 2013; O'Boyle, Banks & Gonzalez-Mulé, 2017; Pashler & Wagenmakers, 2012; Simmons, Nelson & Simonsohn, 2011). However, before speculating about the potential factors which may have led to the potential crisis, it is important to first determine if there really have been adverse effects on our cumulative knowledge in a particular research area, and, to what extent these adverse effects exist. We will explore this by examining the robustness of meta-analytic results from studies published in *Psychological Bulletin* in one management-relevant literature area – learning, behavior, and performance.

HOW TRUSTWORTHY IS OUR CUMULATIVE SCIENTIFIC KNOWLEDGE?

Meta-analytic studies are the primary way for systematically synthesizing quantitative research findings to generate cumulative scientific knowledge (Borenstein, Hedges, Higgins & Rothstein, 2009; Kepes, McDaniel, Brannick & Banks, 2013). It is therefore troublesome that the accuracy of select meta-analytic studies has been questioned (e.g., Banks, Kepes & McDaniel, 2012; Banks, Kepes & McDaniel, 2015; Kepes, Banks & Oh, 2014; Kepes & McDaniel, 2015). Potential reasons for inaccuracy in our literature include the presence of publication bias and outliers. *Publication bias* (PB) denotes a situation in which the publicly available literature on a

particular relation of interest is not representative of all studies on that relation (Banks et al., 2015; Kepes, Banks, McDaniel & Whetzel, 2012). Thus, PB refers to data suppression, typically with respect to effect size data from primary samples.

Such suppression is not necessarily due to purposeful intent; researchers, including authors, reviewers, and editors, make decisions during the scientific and publishing process that can inadvertently lead to the suppression of results, typically statistically non-significant ones (Banks et al., 2015; Chalmers & Dickersin, 2013; Kepes & McDaniel, 2013). Therefore, PB may denote 'errors of exclusion;' that is, some results are not reported but should be. Most suppressed research likely contains null results, or results that were not hypothesized (Fanelli, 2010; Sterling & Rosenbaum, 1995). Furthermore, some evidence indicates that this may be especially pronounced in our most prestigious journals (Eisend & Tarrahi, 2014; Murtaugh, 2002). Because statistically non-significant results (most likely small magnitude effect sizes) are often suppressed, meta-analytic mean estimates can be misestimated, typically overestimated (Kepes et al., 2012; Rothstein, Sutton & Borenstein, 2005b). This misestimation can lead to the use of ineffective interventions, practices, or policies as well as the misdirection of research agendas. Put differently, at best, this leads to ineffective allocation of limited resources for research and practice.

Outliers are another phenomenon that can distort empirical results. *Outliers* or extreme effect size values, refer to effect size data points that are inconsistent with the vast majority of the data (Orr, Sackett & Dubois, 1991; Viechtbauer & Cheung, 2010). Therefore, they can be considered 'errors of inclusion,' data that are in one's dataset although such data probably should not be included. Such data points can be due to random sampling error, transcription errors, or reflect some unique characteristic of a particular sample (e.g., an uncommon operationalization

of a construct). Just like PB, the presence of outliers can adversely affect the results and conclusions of meta-analytic studies (Ada, Sharman & Balkundi, 2012; Viechtbauer & Cheung, 2010). Unfortunately, only 4% of all meta-analyses conduct PB analyses with appropriate methods (Banks et al., 2012) and less than 3% examine the effects of potential outliers (Aguinis, Dalton, Bosco, Pierce & Dalton, 2011). Therefore, we do not have a good understanding of the extent to which these phenomena have affected meta-analytic results. One of our goals is therefore to use more appropriate methods to obtain a better estimate regarding the presence and severity of PB and outliers in the management-relevant literature area of learning, behavior, and performance.

Causes of a Systematically Unrepresentative Literature

Because the probability of publishing a paper is associated with the number of statistically significant results (Cucina & McDaniel, 2016; Fanelli, 2010; Kepes & McDaniel, 2013; Sterling & Rosenbaum, 1995), researchers may be motivated to marshal the available "methodological flexibility" to obtain the desired statistically significant results (Kepes & McDaniel, 2013, p. 255; see also, e.g., Ferguson & Heene, 2012; Simmons et al., 2011). Thus, factors such as our journals' obsession with statistically significant results (Fanelli, 2010) may motivate researchers to engage in questionable research practices (QRPs) to increase their publication count.

The combination of our field's theory-orientation (Campbell & Wilmot, 2018), the lack of research registries, reproducibility assessments, and replications, allows and even motivates researchers to engage in QRPs. For instance, researchers can manipulate their data and analyses until they find statistically significant results that are interesting, popular, or newsworthy, and thus deemed worthy of publication in our journals (Davis, 1971; Hartshorne & Schachner, 2012;

LeBel & Peters, 2011; Pfeffer, 2007). Further, without research registries, researchers can hide null or otherwise uninteresting results from the scientific community. In addition, the lack of reproducibility assessments and direct replications in our journals (Makel, Plucker & Hegarty, 2012; Wicherts, Borsboom, Kats & Molenaar, 2006) suggests that non-robust and potentially erroneous results are not identified. Hence, PB and outliers are likely to exist in our published literature.

The prevalence of underpowered samples is an additional factor that should be addressed. Sample sizes throughout the 1990's and 2000's have remained relatively stable (Shen, Kiger, Davies, Rasch R. L., Simon K. M. & Ones, 2011). Yet, the complexity of our theoretical models (e.g., moderated mediation) has continued to increase, likely raising the prevalence of underpowered samples (Thoemmes, MacKinnon & Reiser, 2010) and thus the likelihood of type II errors. Taken together, factors such as the ones described above yield meta-analytic datasets that can be systematically biased due to suppressed results that are missing from the publicly available literature (i.e., PB) as well as the publication of extreme data points (i.e., outliers). As Kepes et al. (2012) noted, these dynamics are not random; they are systematic. Therefore, methods specifically designed to take these phenomena, PB and outliers, into account are needed to assess their effects on meta-analytic results and, thus, our cumulative knowledge.

Current Study

In this paper, we assess the trustworthiness of cumulative knowledge in one management-related area (learning, behavior, and performance) as published in *Psychological Bulletin,* the premier journal for meta-analyses in the social sciences, including applied psychology and management. Specifically, we examine whether errors of exclusion (i.e., PB) or errors of inclusion (i.e., outliers) have adversely affected published meta-analytic results in this

journal. If both types of errors (i.e., exclusion and inclusion) have affected meta-analytic results, we will assess which type had more adverse effects.

METHOD

Publication bias and outlier analyses are best classified as sensitivity analyses. We performed a comprehensive battery of such analyses on meta-analytic datasets published in *Psychological Bulletin*. We searched APA's PsycNET databases for all meta-analytic reviews published in this journal between 2000 and 2015 using search terms such as *meta-analysis*, *meta-analytic*, *meta-analyses*, and *systematic review.* To be included in our study, a meta-analytic study had to fulfill four criteria. First, the study had to provide a clear description of the methods and meta-analytic distributions. Second, the study needed to include the raw data necessary for our re-analysis. Third, we limited our study to meta-analytic studies with correlation coefficients (e.g., Pearson's *r*). Fourth, the study had to be categorized in the area of learning, behavior, and performance by three individuals with advanced degrees in psychology. Applying these decisions rules resulted in nine meta-analyses for re-analysis. We analyze only those original distributions with 10 or more effect sizes (i.e., $k \geq 10$) because results from distributions with fewer effects may not be trustworthy due to inadequate statistical power and second-order sampling error (Kepes et al., 2012; Sterne et al., 2011). This left us with 123 meta-analytic distributions to be analyzed. Our search and winnowing processes are illustrated in Figure 1.

Insert Figure 1 about here

Meta-analytic and Sensitivity Approach

We used the Hedges and Olkin (1985; Hedges & Vevea, 1998) meta-analytic approach, which allows for the implementation of comprehensive sensitivity analyses (Kepes & McDaniel,

2015; Kepes et al., 2013). We note that PB analyses have not been developed for psychometric meta-analysis. We used *R* version 3.2.4 (R Studio, 2017) and *R Studio* version 1.0.143 (R Studio, 2017) with the *metafor* (Viechtbauer, 2015) and *meta* (Schwarzer, 2015) packages. We refer to the meta-analytic mean without any adjustment for potential biases as the 'naïve' meta-analytic mean (Copas & Shi, 2000) and to the mean estimates obtained from any sensitivity analysis as 'adjusted' ones (Kepes et al., 2012).

When implementing our meta-analytic and sensitivity analysis approach, we followed best practice recommendations (e.g., Kepes & McDaniel, 2015; Kepes et al., 2013; Rothstein et al., 2005b; Viechtbauer & Cheung, 2010) and used several methods to assess the potential presence of PB and outliers. We used the fixed-effects (FE) model and L_0 estimator for the implementation of trim and fill (Kepes et al., 2012; Sutton, 2005) and assess the robustness of these results by also using the random-effects (RE) model with the same estimator (Moreno et al., 2009). For the cumulative meta-analysis by precision, we examined the entire cumulative meta-analysis in a forest plot and, aligned with ideas from Stanley, Jarrell and Doucouliagos (2010), present the cumulative meta-analytic mean of the five most precise effect sizes (Kepes, Bushman & Anderson, 2017). For selection models, we used the a priori approach with *p*-value cut-points to model moderate and severe instances of PB recommended by Vevea and Woods (2005). PET-PEESE (precision-effect test, precision effect estimate with standard error; Stanley & Doucouliagos, 2014) was implemented with the conditional framework for selecting which model, PET or PEESE, to use to obtain the for bias 'adjusted' mean estimate (we used a one-tailed significance test in this framework). For the assessment of outliers, we used the one-sample removed technique (Borenstein et al., 2009) as well as Viechtbauer and Cheung's (2010; Viechtbauer, 2015) battery of multivariate influence diagnostics. Detailed descriptions of these

methods are provided elsewhere (e.g., Banks et al., 2015; Kepes et al., 2012; Rothstein et al., 2005b; Stanley & Doucouliagos, 2014; Vevea & Woods, 2005; Viechtbauer & Cheung, 2010).

When assessing the degree of any potential bias, we followed Kepes et al.'s (2012) decision rules (see also Kepes & McDaniel, 2015). When comparing the originally obtained naïve meta-analytic mean to an adjusted one, adjusted for outliers, PB, or both, a relative change in the naïve meta-analytic mean estimate of less than or equal to 20% indicated that bias is negligible, a change greater than 20% but less than or equal to 40% was interpreted as moderate, and a change of more than 40% was considered severe. Furthermore, we did not rely on any single point estimate but, instead, adopted the triangulation approach (Jick, 1979; Orlitzky, 2012) to locate the position of the 'true' mean effect and to assess the overall robustness of the originally obtained naïve meta-analytic mean estimate (Kepes et al., 2012; Kepes & McDaniel, 2015). Specifically, we perform all PB methods before and after the removal of potential outliers and, thus, use 16 mean estimates, eight before and eight after the removal of potential outliers (the meta-analytic mean, one estimate from the one-sample removed [osr] analysis, and six estimates from the PB analyses all before and after the removal of identified outliers), when estimating the likely location of the 'true' underlying mean effect. Therefore, we urge caution when interpreting any one result in isolation. Instead, one should look for convergence across the different methods. If sensitivity analyses results converge on a mean estimate that differs noticeably (i.e., $\geq$ 20%; Kepes et al., 2012) from the naïve meta-analytic mean effect size estimate, evidence for bias and non-robustness is provided.

Because it is important to account for outliers when assessing publication bias, and for publication bias when assessing outliers (Kepes & McDaniel, 2015), we conducted all PB analyses before and after the removal of the identified outliers. To summarize our results and

assess the potential severity of the bias and the robustness of the naïve meta-analytic mean ($\bar{r}_o$), we follow the procedures outlined by Kepes and McDaniel (2015) and calculate the average range estimate (ARE), baseline range estimate (BRE), and maximum range estimate (MRE). In brief, the ARE was calculated using the difference between the naïve meta-analytic mean estimate ($\bar{r}_o$) and the average of the other estimates (the meta-analytic mean after the removal of outliers as well as one estimate from the one study removed [osr] analysis and six estimates from the publication bias analyses, both before and after the removal of outliers). The BRE was defined as the absolute difference between the naïve mean ($\bar{r}_o$), the potentially best mean estimate of the original meta-analytic dataset, and the mean estimate across all sensitivity analyses that is farthest away from the naïve mean estimate. The MRE was operationalized as the absolute difference between the lowest and the highest value from any of the results of the naïve meta-analysis and the battery of sensitivity analyses.

We then calculated the relative differences for the three range estimates. To calculate the relative difference of the range estimates, we used the naïve mean ($\bar{r}_o$) as the base (i.e., as 100%). We used benchmarks recommended by Kepes et al. (2012; see also, Kepes & McDaniel, 2015) to infer the magnitude of bias. A negligible degree of bias was observed if the relative range (ARE, BRE, or MRE) was at least .02 and less than or equal to 20%, a moderate degree if it was at least .02 and greater than 20% and less than or equal to 40%, and a large degree if it was at least .02 and greater than 40%. Overall conclusions regarding the robustness of the obtained results were determined using the conclusions from the three range estimates, ARE, BRE, and MRE. If the conclusions of the three estimates were in agreement, the overall conclusion resulted in the practical difference noted by the three ranges (i.e., negligible, moderate, or large

difference). If the practical differences did not converge, we reported the range of the range estimates (e.g., moderate to large difference).

When determining whether any non-robustness, if present, was due to publication bias, outliers, or their interaction (i.e., a combination of both), we examined the differences between the naïve meta-analytic estimate ($\bar{r}_o$) from the original distribution and the meta-analytic mean estimate after the removal of identified outlier(s) as well as the 14 estimates from our battery of sensitivity analyses before and after the removal of outliers. Publication bias was considered to be a source of non-robustness if (a) the difference between the naïve meta-analytic mean and the publication bias analyses before or after the removal of outliers was at least .02 in its magnitude and greater than 20% or (b) the difference between the meta-analytic mean estimate after outlier removal and the estimates the publication bias analyses before or after the removal of outliers was at least .02 in its magnitude and greater than 20%. Outliers were considered to contribute to the non-robustness of a naïve meta-analytic mean estimate to a noticeable degree if (a) the difference between the naïve meta-analytic mean estimate, the meta-analytic mean before the removal of identified outlier(s), and any of the osr analyses prior to the removal of outliers was at least .02 in its magnitude and greater than 20%, or (b) the difference between the naïve meta-analytic mean estimate and the meta-analytic mean after the removal of identified outlier(s) was at least .02 in its magnitude and greater than 20%. A combined effect of outliers and publication bias was considered to contribute to the observed non-robustness if the difference between the naïve meta-analytic mean before outlier removal and any estimate of our sensitivity analysis (i.e., osr analysis or publication bias analyses) after the removal of outliers was at least .02 in its magnitude and greater than 20%.

RESULTS

Table 1 contains the results from our meta-analytic and sensitivity analyses for one meta-analytic study. The top part of the table displays the results for the original distributions; the bottom part contains the results with the identified outliers removed. The first three columns in Table 1 report what distribution was analyzed as well as its number of its effect sizes (*k*) and individual observations (*N*). Columns 4-10 show the naïve meta-analytic results, including the random-effects (RE) meta-analytic mean (the naïve mean; $\bar{r}_o$), the 95% confidence interval (95% CI), the 90% prediction interval (90% PI), Cochran's *Q*, I^2, tau (τ), and the osr analysis (osr $\bar{r}_o$; minimum, maximum, and median $\bar{r}_o$ estimates). Columns 11-18 display the results from the trim and fill analyses; for the recommended fixed-effects (FE) as well as the RE model, respectively. For each model, the table includes the side of the funnel plot on which the imputed samples are located (FPS), the number of imputed effect sizes (*ik*), the respective trim and fill adjusted mean effect size (t&f$_{FE}$ $\bar{r}_o$ or t&f$_{RE}$ $\bar{r}_o$), and the corresponding 95% CI. Column 19 contains the cumulative mean for the five most precise samples (pr_5 $\bar{r}_o$). Columns 20 and 21 contain the results from the moderate (sm$_m$ $\bar{r}_o$) and severe selection (sm$_s$ $\bar{r}_o$) models. Finally, column 22 contains the result of the PET-PEESE (pp $\bar{r}_o$) analysis. Due to space limitations, our discussion focuses on two distributions from one meta-analytic study, the study by Karlin, Zinger and Ford (2015). Although we did not include all funnel and forest plots due to space considerations, we have included the ones relevant to the discussed distributions. A summary of results for all nine meta-analyses is also provided.

Robustness of Karlin et al.'s (2015) Meta-analytic Results

Table 1 contains the results for the meta-analytic study by Karlin et al. (2015), which assessed the effects of feedback on performance, as measured by energy conservation. The

Overall effect was estimated to be .04 (k = 42), indicating that, in general, feedback has a small but positive effect on performance (i.e., energy consumption). However, this estimate is unlikely to be robust, as indicated by the results from our PB methods. The FE trim and fill model imputed 12 effect sizes on the left side of the distribution (see Figure 2a, left panel). Consequently, the mean estimate was adjusted downward to .02 (Δ = .02 or 50%). The estimate from the RE trim and fill model was even further away from the naïve meta-analytic mean (t&f$_{RE}$ $\bar{r}_o$ = -.01; Δ = .05 or 125%) and the estimates from the moderate selection model (sm$_m$ $\bar{r}_o$ = .01, Δ = .03 or 75%) as well as the PET-PEESE model (pp $\bar{r}_o$ = .03, Δ = .01 or 25%) were in close proximity to the FE trim and fill estimate. Only the cumulative mean estimate of the five most precise samples suggested that the naïve meta-analytic mean could be underestimated (pr_5 $\bar{r}_o$ = .05, Δ = .01 or 25%). The forest plot of the cumulative meta-analysis showed a sharp discontinuity after around half of the effect sizes were added (see Figure 2a, right panel), suggesting a non-robust naïve mean. The severe selection model did not provide a credible mean estimate. These findings could be partly due to the relatively large degrees of heterogeneity (90% PI = -.07, .16; Q = 814.81; I^2 = 94.97, τ = .07). The contour-enhanced funnel plot showed the effect size that is the likely culprit of the discontinuity and the large degree of heterogeneity (see Figure 3a). In combination with the varying osr estimates, this effect size as well as others could be outliers.

We identified five outliers in the distribution and removed them. Consequently, the amount of heterogeneity decreased noticeably (90% PI = .01, .16; Q = 50.54; I^2 = 28.78; τ = .05) and the adjusted mean was estimated to be .09, .05 (125%) larger in magnitude than the original naïve mean estimate. All but one PB method yielded relatively consistent adjusted mean estimates, either .07 (t&f$_{FE}$ $\bar{r}_o$, t&f$_{RE}$ $\bar{r}_o$, and sm$_m$ $\bar{r}_o$; Δ = .03 or 75%) or .05 (pr_5 $\bar{r}_o$ and sm$_s$ $\bar{r}_o$; Δ

= .01 or 25%). The contour-enhanced funnel plot as well as the cumulative meta-analysis by precision supported these findings (see Figure 2b, left panel). Only the PET-PEESE estimate did not converge with these results (pp $\bar{r}_o$ = .01, Δ = .03 or 75%). Excluding this latter estimate, our results suggested that the originally estimated naïve meta-analytic mean of .04 is somewhat underestimated due to PB and outliers. According to our results, the 'true' underlying observed mean for the effect of feedback on energy consumption is likely to be somewhat larger in magnitude than the originally estimated .04, probably around .06 (Δ = .02 or 55%).

As another example, the naïve meta-analytic mean for effect sizes from studies where the feedback frequency was continuous (*Treatment variables: Feedback frequency: Continuous*, *k* = 17) was estimated to be .05 and the data appeared to be very heterogeneous (90% PI = -.46, .54; Q = 368.57; I^2 = 95.66; τ = .33). Both trim and fill models (t&f$_{FE}$ $\bar{r}_o$ = -.05, Δ = .10 or 200%; t&f$_{RE}$ $\bar{r}_o$ = -.07, Δ = .12 or 240%) as well as the moderate selection model (sm$_m$ $\bar{r}_o$ = -.04, Δ = .09 or 180%) yielded results of similar magnitude but in the opposite direction to the naïve mean. This suggests the presence of rather severe bias. However, the cumulative mean estimate of the five most precise samples (pr_5 $\bar{r}_o$ = .07, Δ = .02 or 40%) suggested that the naïve mean was somewhat underestimated while, once again, the PET-PEESE estimate (pp $\bar{r}_o$ = -.37, Δ = .42 or 840%) did not converge well with any of the other results. Similar to the previously discussed distribution from Karlin et al. (2015), the forest plot depicting the cumulative meta-analysis showed a sharp discontinuity (see Figure 2c, right panel). The contour-enhanced funnel plot (see Figure 2c, left panel) showed the culprit of the discontinuity and the widely varying osr estimates suggested that outliers have a noticeable influence on the obtained results, which is also supported by the large degrees of heterogeneity.

After the removal of one identified outlier (k = 16), the distribution was substantially less heterogeneous (90% PI = -.01, .18; Q = 23.15; I^2 = 36.15; τ = .05) and the for outlier adjusted mean was estimated to be .09 (Δ = .04 or 80%). All but one PB method yielded very similar but somewhat smaller magnitude results (t&f$_{FE}$, t&f$_{RE}$ $\bar{r}_o$, pr_5 $\bar{r}_o$, and sm$_m$ $\bar{r}_o$ = .07, Δ = .02 or 40%; sm$_s$ $\bar{r}_o$ = .06 Δ = .01 or 20%). The funnel plot distribution was more symmetrical and the forest plot depicting the cumulative meta-analysis by precision did not contain the discontinuity (see Figure 2d, left and right panel respectively). Still, the positive drift of the cumulative mean in the forest plot suggested that PB is present in this distribution and, therefore, supported the aforementioned findings. But once again, the PET-PEESE estimate (pp $\bar{r}_o$ = -.02, Δ = .07 or 140%) did not converge with the results of the other PB methods. Taken together and excluding the PET-PEESE estimate, we concluded that the original naïve meta-analytic mean of .05 was underestimated; that the 'true' underlying observed mean is likely to be slightly larger in magnitude, potentially around .07 (Δ = .02 or 36% when compared to the originally estimated naïve meta-analytic mean).

Insert Tables 1 and 2 and Figures 2 and 3 about here

Table 2 summarizes our results regarding the robustness of Karlin et al.'s (2015) naïve meta-analytic mean estimates. Not surprisingly, given the results displayed in Table 1, all of these estimates seemed to be non-robust. According to our guidelines, at least one of the three range estimates indicated a 'large' (i.e., > 40%) difference for all 14 distributions. Furthermore, for over a third of the distributions (36%, 5/14), all three range estimates indicated that there was a 'large' difference. We also determined that outliers contributed to the non-robustness in 71% of the distributions (10/14) and PB was a cause in the observed non-robustness in all instances

(100%, 14/14). A combined effect was observed in 10 distributions (71%). Therefore, outliers, PB, and their interaction had relatively similar distorting effects on the naïve mean estimate. Put differently, errors of exclusion (i.e., PB), errors of inclusion (i.e., outliers), and a combination of both seem to have affected the naïve mean estimates.

Figure 3 shows the dispersion of the naïve meta-analytic mean effect size and the estimates from the battery of sensitivity analyses, before and after outlier removal (when applicable). It can be seen that the mean estimates for several meta-analytic distributions (e.g., *Treatment variables: Feedback frequency: Continuous*, *Treatment variables: Feedback medium: Monitor*, and T*reatment variables: Comparison message: No comparison message*) are widely dispersed. By contrast, the estimates from other distributions (e.g., *Treatment variables: Feedback medium: Card*, *Treatment variables: Comparison message: No comparison message*, and *Treatment variables: Feedback duration – 3-6 month*) converged well and are thus are clustered together. Therefore, the naïve meta-analytic mean estimates for the former distributions are likely to be non-robust and the naïve means for the latter distributions are likely to be robust. Consequently, our cumulative knowledge pertaining to the former distributions is unlikely to be trustworthy but our cumulative knowledge pertaining to the latter distributions is likely to be trustworthy.

Summary of All 123 Distributions

Table 3 summarizes the results of all nine meta-analytic studies and the 123 distributions we re-analyzed. The first two columns contain some general information about each meta-analytic study; the citation of the study and the number of the distributions we re-analyzed. Columns three and four display the number of the of the robust and non-robust estimates we obtained (differences greater than or equal to 20% between the naïve meta-analytic mean and

any one estimate from our sensitivity analyses were classified as non-robust; see the robustness tables). Columns five through seven show the cause of the observed non-robustness. Specifically, they illustrate for how many distributions PB, outliers, or their interaction contributed to the observed non-robustness. Finally, the last two columns display the number of distributions that exhibited 'moderate' (i.e., > 20%) and 'severe' (i.e., > 40%) non-robustness. We derived these numbers by adding the instances in which the overall conclusion from the robustness table associated with a meta-analytic study included the terms 'moderate' and 'large.' Hence, an overall conclusion of 'negligible to large differences' (or 'moderate to large differences') added an instance of one to the 'moderate' and 'severe' degrees of non-robustness. Conversely, an overall conclusion of a 'large difference' added an instance of one to the 'severe' degree of non-robustness column in Table 3. This explains why the sum of these numbers is unequal to the number shown in column four, our estimate for the number of non-robust naïve meta-analytic mean estimates. We did this because the three range estimates (ARE, BRE, and MRE) frequently reached different conclusions. Specifically, given how the ARE was calculated, it tended to provide a conclusion of a 'negligible' bias much more frequently than the other two range estimates. In the spirit of triangulation and to provide the most transparent and accurate overall conclusion, we considered all three range estimates rather than just one upon which to base our conclusion.

As can be seen from the table, only a few of the obtained naïve meta-analytic mean estimates were robust. For example, the naïve mean estimates for all 23 distributions of Judge et al.'s (2001) meta-analytic study were classified as being non-robust. Outliers contributed to this observed non-robustness in two distributions (2/23, 9%); PB was a factor in almost all of the distributions (22/23, 96%; their combined effect was noted in about three-quarters of the

distributions (17/23, 74%). Maybe most importantly, the observed non-robustness was classified as being 'severe' in most of the instances (21/23, 91%). Therefore, the cumulative knowledge related to this meta-analytic study is unlikely to be trustworthy. With regard to the previously discussed meta-analytic study, we can see that Karlin et al.'s (2015) study did not contain any mean estimates that can be categorized as robust. Specifically, 71% (10/14) of the reanalyzed distributions were noticeably affected by the presence of outliers, almost all of the distributions were affected by PB (13/14, 93%), and a combined effect of outliers and PB was noted in 10 of the distributions (10/14, 71%). Consequently, all (14/14, 100%) of the obtained naïve mean estimates were non-robust to a potentially 'severe' degree.

Overall, the naïve meta-analytic mean estimates for the vast majority of distributions in the meta-analytic studies we re-analyzed were classified as being non-robust (i.e., 108/123, 88%). Conversely, the naïve mean estimates of only 12% of the distributions (15/123) were deemed robust. Outliers contributed to the observed non-robustness in 24% (30/123) of the naïve means, PB in 88% (108/123), and their combined effect in 46% (56/123). Maybe most importantly, the degree of non-robustness was classified as being severe (i.e., > 40%) in over three-fourths (78%, 96/123) of the naïve mean estimates. Therefore, one may be able to conclude that our cumulative knowledge derived from the analyzed meta-analytic distributions is unlikely to be trustworthy in the vast majority of the distributions, which should be of major concern. Detailed results tables, robustness tables, and relevant figures for all nine meta-analyses are available upon request.

Insert Table 3 about here

DISCUSSION

The cumulation of research findings in recent years suggests that factors related to our scientific process may have had adverse effects on the trustworthiness of our cumulative knowledge. Of particular interest have been PB and outliers because both of these phenomena have been found to distort meta-analytic mean estimates and related statistics (Banks et al., 2015; Viechtbauer & Cheung, 2010). Unfortunately, most published meta-analyses in our sciences tend to ignore the potential threat stemming from PB and outliers. Therefore, one of our goals was to attain a sound estimate regarding the presence and severity of PB and outliers in meta-analyses in the area of learning, behavior, and performance. In addition, we sought to know which of the two phenomena has the greater effect on any observed non-robustness, and, therefore, is contributing more to the perceived crisis of confidence in our sciences (e.g., Earp & Trafimow, 2015). In our attempt to answer these and related questions, we also verified empirically whether one of these phenomena (e.g., outliers or errors of inclusion) can distort the results of methods to assess the presence of the other (e.g., PB or errors of exclusion). We address these and related issues next.

How Robust are Our Naïve Meta-analytic Results in the Literature Area of Learning, Behavior, and Performance and What Causes Any Observed Non-robustness?

To provide an accurate estimate of the presence and amount of PB and outliers in the literature area of learning, behavior, and performance we re-analyzed relevant publicly available datasets of nine meta-analytic studies with 123 meta-analytic distributions containing at least 10 effect sizes published in *Psychological Bulletin* between 2000 and 2015. Overall, we found that the vast majority (88%, 108/123) of the originally reported naïve meta-analytic means were non-robust.

With regard to our next question, which of the two examined phenomena, PB or outliers, had the greater distorting effect, we found that PB tended to have vastly greater adverse effects. Although we observed outliers to have affected 24% (30/123) of the naïve meta-analytic mean estimates, PB affected almost four times as many naïve mean estimates (88%, 108/123). In other words, according to our results, errors of exclusion (i.e., PB) tended to have a substantially larger distorting effect than errors of inclusion (i.e., outliers), providing credence to the notion that PB may present one of the greatest threats to the validity of meta-analytic results (Rothstein, Sutton & Borenstein, 2005a).

We also observed a combined effect of PB and outliers in roughly a third of the distributions (46%, 56/123). This finding is interesting as this rate of occurrence is larger in magnitude than the one for outliers alone. It seems as if both of these phenomena can have an interactive effect that may go beyond the main effect of outliers. Thus, we can conclude that both PB and outliers as well as their combined effect contributed to the high frequency and degrees of misestimation and non-robustness of the naïve meta-analytic mean estimates.

Do Outliers Distort Results of Publication Bias Analyses?

Just as with meta-analytic methods, many PB methods are sensitive to heterogeneity. Therefore, we not only wanted to know whether outliers can distort naïve meta-analytic mean estimates, but also whether outliers affect results from PB methods. Here, our results are based on the 72 distributions for which we obtained results before *and* after the removal of outliers. First, we observed that, as expected, the removal of outliers tended to reduce the observed heterogeneity, as indicated by the typically smaller values Q, I^2, and τ as well as the narrowing of the 90% PI, when compared to the distributions before outlier removal. Next, we assessed the frequency to which each of our PB methods yielded the same results before and after removal of

outliers (see Table 4). From the table, it can be seen that the magnitude of the naïve meta-analytic mean remained the same after outlier removal in only 6% (4/72) of the distributions. The estimates from the RE trim and fill model (14%, 10/72), PET-PEESE (7%, 5/72), and both selection models (14%, 10/72; 3%, 2/72) are similar in magnitude.

Insert Table 4 about here

The estimates from the FE trim and fill model seem to be substantially less affected by outliers as the rate of agreement before and after outlier removal is noticeably larger in magnitude (32%, 23/72). The estimate from the five most precise samples ($pr_5\ \bar{r}_o$) is more than twice as large (65%, 47/72). Therefore, it seems to be the method most robust to the influence of outliers. By contrast, the RE trim and fill model, both selection models, and PET-PEESE do not seem to be robust to the influence of outliers. Regardless, even agreements of 32% or 65% do not seem large enough to have confidence in the accuracy of PB results before the removal of outliers. We thus recommend that outliers are taken into account when estimating a meta-analytic mean and as well as the effect of PB on it. In brief, we clearly demonstrate that outliers can distort meta-analytic and PB analyses results.

Implications for Research and Practice

Based on our study, several implications for research and practice can be discerned. Maybe most importantly, our results unambiguously indicated that both PB and outliers can affect naïve mean estimates and related statistics and, therefore, the robustness of meta-analytic results and the trustworthiness of the associated conclusions and recommendations (see Table 3). Furthermore, both phenomena, PB and outliers, seemed to have an interactive effect on the robustness of obtained naïve results. Our findings even indicated that this combined effect is

likely to be larger in magnitude than the effect of outliers alone (see Table 3). Consequently, one should account for both phenomena when conducting a meta-analytic study. Without such a comprehensive assessment of obtained results, their robustness is unknown and, therefore, the delineated recommendations for future research and practice may be misleading if not erroneous. Consequently, a literature area may not be trustworthy.

With regard to the use of particular methods, the Meta-analysis Reporting Standards do not recommend any particular one (APA, 2008). Based on our study, we can provide more specific recommendations. We calculated the convergence rates of the results from the different methods to assess PB (see Table 4). The table illustrates convergence rates before and after outlier removal. It shows the frequency in which each method yielded a 'negligible,' 'moderate,' or 'severe' difference before and after outlier removal. In addition, the middle part of the table displays the inter-PB detection method convergence rates across the three levels of practical difference before and after outlier removal. As can be seen from the middle part of the table, convergence rates between the results tended to increase following outlier removal.

Overall, convergence (after outlier removal) was highest between the results from the FE and RE trim and fill model (86%), the FE trim and fill model and the estimate from the moderate selection model (64%), the RE trim and fill model and the moderate selection model (66%), the RE trim and fill model and the moderate selection model (63%), and the RE trim and fill model and the PET-PEESE estimate (63%). Therefore, it is a combination of these empirical methods that we recommend for future research. Specifically, in addition to the naïve meta-analytic mean (i.e., $\bar{r}_o$), we recommend the use of, at a minimum, the recommended FE trim and fill as well as the RE trim and fill model, the moderate selection model, and the PET-PEESE estimate to triangulate the likely location of the most robust estimate of the 'true' meta-analytic effect size.

However, it is important to note that the performance of PET-PEESE varied widely between meta-analyses and distributions. For instance, although we included the PET-PEESE estimate in our calculations of the ARE, BRE, and MRE, we excluded it from our overall conclusions (i.e., our estimate of the "true" effect size) in both distributions we discussed from Karlin et al.'s (2015) meta-analytic study because of its lack of convergence with the other estimates. Thus, our general recommendations may need to be adapted/adjusted to the particular distributions analyzed. In addition to these methods, we recommend the use of two graphical methods, the funnel plot to examine the distribution and the forest plot to display the cumulative meta-analysis by precision. To provide more generalizable recommendations, we echo prior research and recommend that future simulation studies assess the performance of PB methods under various conditions (e.g., number of effect sizes and degree of heterogeneity; Kepes et al., 2012; Kepes & McDaniel, 2015).

We also have recommendations related to practice. Because meta-analytic reviews are potentially the most important tool for advancing evidence-based practice, practitioners should be aware that phenomena such as PB and outliers can adversely affect meta-analytic results. Therefore, they may want to influence researchers, especially journal editors, to ensure that future meta-analytic studies include the results of comprehensive sensitivity analyses to assess the robustness of the reported naïve meta-analytic results as well as the trustworthiness of the associated conclusions and recommendations. Relatedly, we suggest that practitioners be skeptical about implementing interventions recommended by studies that did not include sensitivity analyses and always collect their own data post-intervention to ensure that the intervention is functioning as proposed.

Limitations and Future Research

As is the case in every study, our study has limitations. First, the trustworthiness of meta-analytic results in the area of learning, behavior, and performance is not solely predicated on the meta-analytic studies published in *Psychological Bulletin* (or any other particular journal). We selected *Psychological Bulletin* because of its visibility and prominence. Because of this, we thought that the review process at this journal is likely more thorough when compared to other journals. Hence, we reasoned that meta-analytic studies in this journal may be considered to be among the best in the area of learning, behavior, and performance.

Second, one may question the adequacy of one or more of the methods we used, especially because the methods did not yield identical results. Contrary to this view, we see this as a strength of our study. We argue that is bad practice to rely on just one PB detection method to come to conclusions regarding the robustness of the naïve results. Instead, aligned with the concept of triangulation (Jick, 1979), when trying to estimate the location of the 'true' underlying effect size, it has been recommended to use several methods, especially ones that rely on different statistical assumptions, to increase one's confidence that the obtained results are not due to one particular methodological approach (Kepes & McDaniel, 2015). Inevitably, such an approach with multiple methods will yield differing results.

Relatedly, the fact that some methods to assess PB did not yield credible results may be viewed as a limitation. For instance, the results of the severe selection model estimate were regularly not credible and thus omitted from the results table. This is likely due to inflated variance estimates associated with their respective mean estimate (Kepes et al., 2012; Vevea & Woods, 2005). Clearly, more research is needed to understand the situations under which this selection model does not perform well, especially because Vevea and Woods (2005) recommended the severe selection model in particular.

Conclusion

Literature in the area of learning, behavior, and performance, as summarized in nine meta-analyses with 123 distributions containing at least 10 effect sizes published in *Psychological Bulletin*, is largely not trustworthy. This literature suffers from errors of exclusion (distorted results due to publication bias) and errors of inclusion (distorted results due to inclusion of outliers). Therefore, some of the concerns regarding the credibility crisis in psychology seem to be justified. To regain its status as a trustworthy scientific discipline, it seems as if both primary and meta-analytic research in management and applied psychology needs to revise its practices.

References

Ada, S., Sharman, R., & Balkundi, P. 2012. Impact of meta-analytic decisions on the conclusions drawn on the business value of information technology. *Decision Support Systems, 54*: 521-533.

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. 2011. Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management, 37*: 5-38.

American Psychological Association. 2008. Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63*: 839-851. doi: 810.1037/0003-1066X.1063.1039.1839.

Banks, G. C., Kepes, S., & McDaniel, M. A. 2012. Publication bias: A call for improved meta-analytic practice in the organizational sciences. *International Journal of Selection and Assessment, 20*: 182-196.

Banks, G. C., Kepes, S., & McDaniel, M. A. 2015. Publication bias: Understanding the myths concerning threats to the advancement of science. In C. E. Lance and R. J. Vandenberg (Eds.), *More statistical and methodological myths and urban legends* 36-64. New York, NY: Routledge.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. 2009. *Introduction to meta-analysis*. West Sussex, UK: Wiley.

Campbell, J. P., & Wilmot, M. P. 2018. The functioning of theory in IWOP. In D. S. Ones, N. Anderson, C. Viswesvaran and H. Kepir (Eds.), *The sage handbook of industrial, work & organizational psychology*: 3-38. Thousand Oaks, CA: Sage.

Cerasoli, C. P., Nicklin, J. M., & Ford, M. T. 2014. Intrinsic motivation and extrinsic incentives jointly predict performance: A 40-year meta-analysis. *Psychological Bulletin, 140*: 980.

Chalmers, I., & Dickersin, K. 2013. Biased under-reporting of research reflects biased under-submission more than biased editorial rejection. *F1000Res, 2*.

Copas, J., & Shi, J. Q. 2000. Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics, 1*: 247-262.

Cucina, J. M., & McDaniel, M. A. 2016. Pseudotheory proliferation is damaging the organizational sciences. *Journal of Organizational Behavior, 37*: 1116-1125.

Davis, M. S. 1971. That's interesting! Towards a phenomenology of sociology and a sociology of phenomenology. *Philosophy of the Social Sciences, 1*: 309-344.

Earp, B. D., & Trafimow, D. 2015. Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology, 6*.

Eisend, M., & Tarrahi, F. 2014. Meta-analysis selection bias in marketing research. *International Journal of Research in Marketing, 31*: 317-326.

Fanelli, D. 2010. "Positive" results increase down the hierarchy of the sciences. *PLoS One, 5*: e10068.

Ferguson, C. J., & Heene, M. 2012. A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*: 555-561.

Glasman, L. R., & Albarracín, D. 2006. Forming attitudes that predict future behavior: A meta-analysis of the attitude-behavior relation. *Psychological Bulletin, 132*: 778-822.

Goldstein, D. 2010. *On fact and fraud: Cautionary tales from the front lines of science*. Princeton, NJ: Princeton University Press

Hartshorne, J., & Schachner, A. 2012. Tracking replicability as a method of post-publication open evaluation. *Frontiers in Computational Neuroscience, 6*: 1-14.

Hedges, L. V., & Olkin, I. 1985. *Statistical methods for meta-analysis*. New York, NY: Academic Press.

Hedges, L. V., & Vevea, J. L. 1998. Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*: 486-504.

Ioannidis, J. P. A. 2005. Why most published research findings are false. *PLoS Medicine, 2*: e124.

Jick, T. D. 1979. Mixing qualitative and quantitative methods: Triangulation in action. *Administrative Science Quarterly, 24*: 602-611.

Judge, T. A., Thoresen, C. J., Bono, J. E., & Patton, G. K. 2001. The job satisfaction–job performance relationship: A qualitative and quantitative review. *Psychological Bulletin, 127*: 376-407.

Karlin, B., Zinger, J. F., & Ford, R. 2015. The effects of feedback on energy conservation: A meta-analysis. *Psychological Bulletin, 141*: 1205-1227.

Kepes, S., Banks, G., C., McDaniel, M. A., & Whetzel, D. L. 2012. Publication bias in the organizational sciences. *Organizational Research Methods, 15*: 624-662.

Kepes, S., Banks, G. C., & Oh, I.-S. 2014. Avoiding bias in publication bias research: The value of "null" findings. *Journal of Business and Psychology, 29*: 183-203.

Kepes, S., Bushman, B. J., & Anderson, C. A. 2017. Violent video game effects remain a societal concern: Comment on hilgard, engelhardt, and rouder (2017). *Psychological Bulletin, 143*: 775-782.

Kepes, S., & McDaniel, M. A. 2013. How trustworthy is the scientific literature in industrial and organizational psychology. *Industrial and Organizational Psychology: Perspectives on Science and Practice, 6*: 252-268.

Kepes, S., & McDaniel, M. A. 2015. The validity of conscientiousness is overestimated in the prediction of job performance. *PLoS One, 10*: e0141468.

Kepes, S., McDaniel, M. A., Brannick, M. T., & Banks, G. C. 2013. Meta-analytic reviews in the organizational sciences: Two meta-analytic schools on the way to mars (the meta-analytic reporting standards). *Journal of Business and Psychology, 28*: 123-143.

LeBel, E. P., & Peters, K. R. 2011. Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology, 15*: 371-379.

Lee, E.-S., Park, T.-Y., & Koo, B. 2015. Identifying organizational identification as a basis for attitudes and behaviors: A meta-analytic review. *Psychological Bulletin, 141*: 1049-1080.

Makel, M. C., Plucker, J. A., & Hegarty, B. 2012. Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*: 537-542.

Moreno, S. G., Sutton, A., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. 2009. Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology, 9*.

Murtaugh, P. A. 2002. Journal quality, effect size, and publication bias in meta-analyses. *Ecology, 83*: 1162-1166.

Noar, S. M., Benac, C. N., & Harris, M. S. 2007. Does tailoring matter? Meta-analytic review of tailored print health behavior change interventions. *Psychological Bulletin, 133*: 673-693.

NSF SBE Advisory Committee. 2015. Social, behavioral, and economic sciences perspectives on robust and reliable science: Report of the subcommittee on replicability in science advisory committee to the national science foundation directorate for social, behavioral, and economic sciences. In N. S. Foundation (Ed.). Washington, D.C.: National Science Foundation.

O'Boyle, E. H., Banks, G. C., & Gonzalez-Mulé, E. 2017. The chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management, 43*: 376-399.

Orlitzky, M. 2012. How can significance tests be deinstitutionalized? *Organizational Research Methods, 15*: 199-228.

Orr, J. M., Sackett, P. R., & Dubois, C. L. 1991. Outlier detection and treatment in I/O psychology: A survey of researcher beliefs and an empirical illustration. *Personnel Psychology, 44*: 473-486.

Pashler, H., & Wagenmakers, E. J. 2012. Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*: 528-530.

Pfeffer, J. 2007. A modest proposal: How we might change the process and product of managerial research. *Academy of Management Journal, 50*: 1334-1345.

R Studio. 2017. *Release notes: R studio v1.0.143 - april 19th, 2017*, https://www.rstudio.com/products/rstudio/release-notes/.

Randall, J. G., Oswald, F. L., & Beier, M. E. 2014. Mind-wandering, cognition, and performance: A theory-driven meta-analysis of attention regulation. *Psychological Bulletin, 140*: 1411-1431.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R., & Carlstrom, A. 2004. Do psychosocial and study skill factors predict college outcomes? A meta-analysis. *Psychological Bulletin, 130*: 261-288.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. 2005a. Publication bias in meta-analyses. In H. R. Rothstein, A. J. Sutton and M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment, and adjustments*: 1-7. West Sussex, UK: Wiley.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. 2005b. ***Publication bias in meta-analysis: Prevention, assessment, and adjustments.*** West Sussex, UK: Wiley.

Schwarzer, G. 2015. Meta-analysis package for r: Package 'meta.' r package version 4.3-2.

Shen, W., Kiger, T. B., Davies, S. E., Rasch R. L., Simon K. M., & Ones, D. S. 2011. Samples in applied psychology: Over a decade of research in review. *Journal of Applied Psychology, 96*: 1055-1064.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. 2011. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*: 1359-1366.

Stanley, T. D., & Doucouliagos, H. 2014. Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*: 60-78.

Stanley, T. D., Jarrell, S. B., & Doucouliagos, H. 2010. Could it be better to discard 90% of the data? A statistical paradox. *The American Statistician, 64*: 70-77.

Sterling, T. D., & Rosenbaum, W. L. 1995. Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician, 49*: 108-112.

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., Carpenter, J., Rücker, G., Harbord, R. M., Schmid, C. H., Tetzlaff, J., Deeks, J. J., Peters, J., Macaskill, P., Schwarzer, G., Duval, S., Altman, D. G., Moher, D., & Higgins, J. P. T. 2011.

Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal, 343*: 302-307.

Sutton, A. J. 2005. Evidence concerning the consequences of publication and related biases. In H. R. Rothstein, A. J. Sutton and M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments*: 175-192. West Sussex, UK: Wiley.

Thoemmes, F., MacKinnon, D. P., & Reiser, M. R. 2010. Power analysis for complex mediational designs using monte carlo methods. *Structural Equation Modeling: A Multidisciplinary Journal, 17*: 510-534.

Vachon, D. D., Lynam, D. R., & Johnson, J. A. 2014. The (non)relation between empathy and aggression: Surprising results from a meta-analysis. *Psychological Bulletin, 140*: 751-773.

Vevea, J. L., & Woods, C. M. 2005. Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods, 10*: 428-443.

Viechtbauer, W. 2015. Meta-analysis package for r: Package 'metafor.' r package version 1.9-5.

Viechtbauer, W., & Cheung, M. W. L. 2010. Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*: 112-125.

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. 2006. The poor availability of psychological research data for reanalysis. *American Psychologist, 61*: 726-728.

TABLE 1
Meta-analytic and Sensitivity Analysis Results for Karlin et al.'s (2015) Meta-analytic Dataset

	Meta-analysis									Publication bias analyses											
										Trim and fill								CMA	Selection models		PET-PEESE
										FE trim and fill				RE trim and fill							
Distribution	*k*	*N*	$\bar{r}_o$	95% CI	90% PI	*Q*	I^2	τ	osr $\bar{r}_o$	FPS	*ik*	t&f$_{FE}$ $\bar{r}_o$	t&f$_{FE}$ 95% CI	FPS	*ik*	t&f$_{RE}$ $\bar{r}_o$	t&f$_{RE}$ 95% CI	pr_5 $\bar{r}_o$	sm_m $\bar{r}_o$	sm_s $\bar{r}_o$	pp $\bar{r}_o$
Original distributions																					
Overall effect	42	256536	.04	.01, .08	-.07, .16	814.81	94.97	.07	.04, .07, .04	L	12	.02	-.01, .05	L	18	-.01	-.04, .02	.05	.01	n/a	.03
Treatment variables																					
- Feedback frequency																					
- Continuous	17	3744	.05	-.11, .21	-.46, .54	368.57	95.66	.33	.03, .09, .05	L	7	-.05	-.19, .08	L	7	-.07	-.21, .07	.07	-.04	n/a	-.37
- Feedback medium																					
- Card	15	2772	.08	.03, .13	.00, .15	17.20	18.61	.04	.07, .09, .08	L	4	.05	-.01, .11	L	4	.05	-.01, .11	.06	.06	.04	.03
- Monitor	16	3734	.04	-.13, .20	-.47, .53	368.83	95.93	.33	.02, .09, .04	L	7	-.08	-.21, .06	L	7	-.09	-.23, .05	.07	n/a	n/a	-.38
- Energy measurement																					
- kWh and cost	23	251454	.07	.03, .10	-.02, .15	428.49	94.87	.05	.02, .08, .06	L	8	.04	.01, .08	L	6	.05	.01, .08	.04	.05	.04	.04
- Comparison message																					
- No comparison	17	4240	.02	-.13, .17	-.46, .49	367.97	95.65	.30	.01, .08, .02	L	7	-.08	-.20, .04	L	7	-.09	-.21, .03	.07	-.07	n/a	-.37
- Comparison message	13	79755	.08	.01, .15	-.05, .20	23.95	49.89	.07	.05, .12, .08	L	6	.01	-.06, .08	L	6	.01	-.06, .08	.05	.06	.01	.01
- Comparison																					
- Comparison goal																					
- No goal comparison	38	256343	.04	.01, .07	-.07, .15	803.35	95.39	.07	.03, .07, .04	L	10	.02	-.01, .05	L	17	-.02	-.05, .00	.05	-.01	n/a	.03
- Combined intervention																					
- Feedback only	34	255087	.06	.03, .09	-.02, .15	446.82	92.61	.05	.04, .07, .06	L	9	.05	.02, .07	L	6	.05	.03, .08	.05	.05	.04	.04
- Energy granularity																					
- Whole home	38	255631	.04	.01, .07	-.07, .15	807.54	95.42	.07	.03, .07, .04	L	10	.01	-.02, .04	L	17	-.03	-.06, .00	.05	.00	n/a	.03
- Feedback duration																					
- 3-6 months	10	2721	.05	.01, .10	-.03, .14	13.48	33.23	.04	.04, .07, .05	R	1	.06	.01, .11	R	1	.06	.01, .11	.07	.04	n/a	.09
- 6-12 months	11	249200	-.04	-.09, .01	-.16, .08	761.64	98.69	.07	-.07, .05, -.05			-.04	-.09, .01			-.04	-.09, .01	.03	-.12	n/a	.05
Publication bias																					
- Sample size																					
- <300	26	2295	.11	-.06, .27	-.53, .67	388.68	93.57	.42	.10, .14, .11	L	12	-.10	-.24, .04	L	11	-.01	-.14, .12	-.15	.02	n/a	-.54
- >300	16	254241	.05	.02, .08	-.04, .13	421.20	96.44	.05	.03, .05, .05	L	2	.04	.01, .07			.05	.02, .08	.05	.04	n/a	.04
Distributions without outliers																					
Overall effect	37	7491	.09	.06, .12	.01, .16	50.54	28.78	.05	.08, .09, .09	L	9	.07	.03, .10	L	9	.07	.03, .10	.05	.07	.05	.01
Treatment variables																					

	Meta-analysis									Publication bias analyses											
										Trim and fill								CMA	Selection models		PET-PEESE
										FE trim and fill				RE trim and fill							
Distribution	*k*	*N*	$\bar{r}_o$	95% CI	90% PI	*Q*	I^2	τ	osr $\bar{r}_o$	FPS	*ik*	t&f$_{FE}$ $\bar{r}_o$	t&f$_{FE}$ 95% CI	FPS	*ik*	t&f$_{RE}$ $\bar{r}_o$	t&f$_{RE}$ 95% CI	pr_5 $\bar{r}_o$	sm_m $\bar{r}_o$	sm_s $\bar{r}_o$	pp $\bar{r}_o$
- Feedback frequency																					
- Continuous	16	3475	.09	.04, .13	-.01, .18	23.49	36.15	.05	.08, .10, .08	L	5	.07	.02, .11	L	4	.07	.03, .12	.07	.07	.06	-.02
- Feedback medium																					
- Card	*No outliers detected*																				
- Monitor	15	3465	.09	.04, .13	-.01, .19	23.52	40.47	.06	.08, .10, .09	L	4	.07	.02, .12	L	3	.07	.02, .12	.07	.07	.06	-.02
- Energy measurement																					
- kWh and cost	13	770	.15	.08, .22	.09, .21	7.36	.00	.00	.14, .17, .15			.15	.08, .22			.15	.08, .22	.17	.13	.11	.16
- Comparison message																					
- No comparison	16	3971	.08	.04, .12	.00, .15	21.23	29.33	.04	.07, .09, .08	L	3	.06	.02, .10	L	3	.06	.02, .10	.07	.06	.04	.01
- Comparison message	*No outliers detected*																				
- Comparison																					
- Comparison goal																					
- No goal comparison	33	7298	.08	.05, .11	.01, .14	41.63	23.13	.04	.07, .08, .08	L	7	.07	.04, .10	L	7	.07	.04, .10	.05	.06	.05	.02
- Combined intervention																					
- Feedback only	26	4112	.06	.02, .09	-.02, .13	31.72	21.18	.04	.05, .06, .06	L	7	.04	-.01, .08	L	7	.04	-.01, .08	.02	.04	.02	-.05
- Energy granularity																					
- Whole home	34	7870	.08	.05, .11	.00, .15	46.45	28.95	.04	.07, .08, .08	L	8	.06	.03, .09	L	8	.06	.03, .09	.05	.06	.05	.02
- Feedback duration																					
- 3-6 months	*No outliers detected*																				
- 6-12 months	*No outliers detected*																				
Publication bias																					
- Sample size																					
- <300	22	1401	.17	.11, .22	.12, .21	15.87	.00	.00	.16, .18, .17	L	1	.16	.11, .21	L	1	.16	.11, .21	.19	.15	.12	.16
- >300	11	5676	.04	.02, .07	.01, .08	10.93	8.54	.01	.04, .05, .04	R	1	.05	.02, .07	R	1	.05	.02, .07	.04	.04	.01	.09

Note. *k* = number of correlation coefficients in the analyzed distribution; *N* = meta-analytic sample size; $\bar{r}_o$ = random-effects weighted mean observed correlation; 95% CI = 95% confidence interval; 90% PI = 90% prediction interval; *Q* = weighted sum of squared deviations from the mean; I^2 = ratio of true heterogeneity to total variation; τ = between-sample standard deviation; osr $\bar{r}_o$ = one-sample removed observed means, including the minimum and maximum observed mean as well as the median observed mean; Trim-and-fill = trim-and-fill analysis; FPS = funnel plot side (i.e., side of the funnel plot where samples were imputed; L = left, R = right); *ik* = number of trim-and-fill samples imputed; t&f$_{FE}$ $\bar{r}_o$ = fixed-effects trim and fill adjusted observed mean; t&f$_{FE}$ 95% CI = fixed-effects trim and fill adjusted 95% confidence interval; t&f$_{RE}$ $\bar{r}_o$ = random-effects trim-and-fill adjusted observed mean; t&f$_{RE}$ 95% CI = random-effects trim-and-fill adjusted 95% confidence interval; CMA = cumulative meta-analysis; pr_5 $\bar{r}_o$ = cumulative meta-analytic mean estimate of the five most precise effects; sm_m $\bar{r}_o$ = one-tailed moderate selection model's adjusted observed mean; sm_s $\bar{r}_o$ = one-tailed severe selection model's adjusted observed mean; PET-PEESE = precision-effect test-precision effect estimate with standard error; PET-PEESE $\bar{r}_o$ = PET-PEESE adjusted observed mean; n/a = not applicable (sm_s $\bar{r}_o$ was non-credible due to an inflated variance estimate; Vevea & Woods, 2005).

TABLE 2
Robustness of Karlin et al.'s (2015) Naïve Meta-analytic Mean Estimates and Conclusions of the Sensitivity Analyses

Distribution	Average value	Lowest value	Highest value	$\bar{r}_o$	ARE			BRE			MRE			O	PB	O+PB	Overall conclusion
					Value	%	Conclusion	Value	%	Conclusion	Value	%	Conclusion				
Overall effect	.05	-.01	.09	.04	.01		Negligible	.05	125%	Large	.10	250%	Large	Yes	Yes	Yes	Negligible to large differences
Treatment variables: Feedback frequency: Continuous	.01	-.37	.09	.05	.04	80%	Large	.42	840%	Large	.46	920%	Large	Yes	Yes	Yes	Large differences
Treatment variables: Feedback medium: Card	.06	.03	.08	.08	.02		Negligible	.05	63%	Large	.05	63%	Large	No	Yes	No	Negligible to large differences
Treatment variables: Feedback medium: Monitor	.00	-.38	.09	.04	.04	100%	Large	.42	1050%	Large	.47	1175%	Large	Yes	Yes	Yes	Large differences
Treatment variables: Energy measurement: kWh and cost	.10	.04	.17	.07	.03	43%	Large	.10	143%	Large	.13	186%	Large	Yes	Yes	Yes	Large differences
Treatment variables: Comparison message: No comparison message	.00	-.37	.08	.02	.02		Negligible	.39	1950%	Large	.45	2250%	Large	Yes	Yes	Yes	Negligible to large differences
Treatment variables: Comparison message: Comparison message	.04	.01	.08	.08	.04	50%	Large	.07	88%	Large	.07	88%	Large	Yes	Yes	No	Large differences
Treatment variables: Goal comparison: No goal comparison	.04	-.02	.08	.04	.00		Negligible	.06	150%	Large	.10	250%	Large	Yes	Yes	Yes	Negligible to large differences
Treatment variables: Combination intervention: Feedback only	.04	-.05	.06	.06	.02		Negligible	.11	183%	Large	.11	183%	Large	No	Yes	Yes	Negligible to large differences
Treatment variables: Energy granualarity: Whole home	.04	-.03	.08	.04	.00		Negligible	.07	175%	Large	.11	275%	Large	Yes	Yes	Yes	Negligible to large differences
Treatment variables: Feedback duration - 3-6 months	.06	.04	.09	.05	.01		Negligible	.04	80%	Large	.05	100%	Large	No	Yes	No	Negligible to large differences
Treatment variables: Feedback duration - 6-12 months	-.03	-.12	.05	-.04	.01		Negligible	.09	225%	Large	.17	425%	Large	Yes	Yes	No	Negligible to large differences
Publication bias: Sample size: < 300	.05	-.54	.19	.11	.06	55%	Large	.65	591%	Large	.73	664%	Large	Yes	Yes	Yes	Large differences
Publication bias: Sample size: > 300	.05	.01	.09	.05	.00		Negligible	.04	80%	Large	.08	160%	Large	No	Yes	Yes	Negligible to large differences

Note: Average value = average mean estimate from all analyses, with the exception of the $osr_{min}\,\bar{r}_o$ and $osr_{max}\,\bar{r}_o$ estimates; Lowest value = lowest mean estimate from all analyses, with the exception of the $osr_{min}\,\bar{r}_o$ and $osr_{max}\,\bar{r}_o$ estimates; Highest value = highest mean estimate from all analyses, with the exception of the $osr_{min}\,\bar{r}_o$ and $osr_{max}\,\bar{r}_o$ estimates; $\bar{r}_o$ = random-effects weighted mean observed correlation from the distribution with outliers included; ARE = average range estimate: the absolute range between $\bar{r}_o$ and the average value; BRE = Baseline range estimate: the absolute range between $\bar{r}_o$ and the estimate farthest away (either the lowest or highest value); MRE = Maximum range estimate: the absolute range between the lowest or highest value; When calculating the relative difference of the range estimates, we used the $\bar{r}_o$ from the original distribution, the potentially best mean estimate, as the base. We note that, in many instances, the moderate and severe selection models ($sm_m\bar{r}_o$ and $sm_s\bar{r}_o$) provided nonsensical estimates. We chose to omit these estimate in such instances in order to provide the most conservative estimates of ARE, BRE and MRE.

TABLE 3
Summary of the Results

Author(s)	Distribu-tions analyzed	Robust naïve mean estimates	Non-robust naïve mean estimates	Naïve mean estimates affected by			Degree of non-robustness	
				outliers	publication bias	outliers and pub. bias	moderate	severe
All distributions	123	15 (12%)	108 (98%)	21 (17%)	108 (98%)	56 (46%)	105 (85%)	96 (78%)
1. Cerasoli, Nicklin and Ford (2014)	4	0 (0%)	4 (100%)	0 (0%)	4 (100%)	4 (100%)	4 (100%)	2 (50%)
2. Glasman and Albarracín (2006)	25	7 (28%)	18 (72%)	0 (0%)	18 (72%)	3 (12%)	22 (88%)	9 (36%)
3. Judge et al. (2001)	23	0 (0%)	23 (100%)	2 (9%)	22 (96%)	17 (74%)	23 (100%)	21 (91%)
4. Karlin et al. (2015)	14	0 (0%)	14 (100%)	10 (71%)	13 (93%)	10 (71%)	9 (64%)	14 (100%)
5. Lee, Park and Koo (2015)	6	0 (0%)	6 (100%)	0 (0%)	6 (100%)	0 (0%)	6 (100%)	6 (100%)
6. Noar, Benac and Harris (2007)	13	5 (38%)	8 (62%)	1 (8%)	8 (62%)	6 (46%)	9 (69%)	10 (77%)
7. Randall, Oswald and Beier (2014)	17	0 (0%)	17 (100%)	3 (18%)	17 (100%)	10 (59%)	15 (88%)	15 (88%)
8. Robbins, Lauver, Le, Davis, Langley and Carlstrom (2004)	12	3 (25%)	9 (75%)	1 (8%)	11 (92%)	3 (25%)	12 (100%)	11 (92%)
9. Vachon, Lynam and Johnson (2014)	9	0 (0%)	9 (100%)	4 (44%)	9 (100%)	3 (33%)	5 (56%)	8 (89%)

Note: Mean estimates were considered non-robust if the BRE and MRE indicated 'moderate' or 'severe' bias. The degree of non-robustness was based on the range of the ARE, BRE, and MRE. Hence, an "overall conclusion" of 'negligible to large differences' (or 'moderate to large differences') added an instance of one to the 'moderate' and 'severe' degrees of non-robustness columns. Conversely, an "overall conclusion" of a 'large difference' added an instance of one to the 'severe' degree of non-robustness column.

TABLE 4
Agreement Between Mean Estimates Before and After Outlier Removal

Author(s)	$\bar{r}_o$	t&f$_{FE}$ $\bar{r}_o$	t&f$_{RE}$ $\bar{r}_o$	pr_5 $\bar{r}_o$	sm$_m$ $\bar{r}_o$	sm$_s$ $\bar{r}_o$	pp $\bar{r}_o$
All distributions	5 (7%)	23 (32%)	10 (14%)	47 (65%)	10 (14%)	2 (3%)	5 (7%)
1. Cerasoli et al. (2014)	0 (%)	3 (75%)	0 (%)	3 (75%)	1 (25%)	1 (33%)	1 (25%)
2. Glasman and Albarracín (2006)	0 (%)	3 (60%)	2 (40%)	3 (60%)	0 (%)	0 (%)	1 (20%)
3. Judge et al. (2001)	1 (5%)	7 (33%)	1 (5%)	18 (86%)	3 (14%)	0 (%)	0 (%)
4. Karlin et al. (2015)	1 (10%)	0 (%)	1 (10%)	6 (60%)	1 (10%)	0 (%)	0 (%)
5. Lee et al. (2015)	n/a	n/a	n/a	n/a	n/a	n/a	n/a
6. Noar et al. (2007)	3 (30%)	2 (20%)	0 (%)	4 (40%)	1 (10%)	1 (33%)	2 (20%)
7. Randall et al. (2014)	0 (%)	4 (36%)	5 (45%)	6 (55%)	1 (9%)	0 (%)	0 (%)
8. Robbins et al. (2004)	0 (%)	1 (20%)	0 (%)	3 (60%)	2 (40%)	0 (%)	1 (20%)
9. Vachon et al. (2014)	0 (%)	3 (50%)	1 (17%)	4 (67%)	1 (17%)	n/a	0 (%)

Note: Although 123 meta-analytic distributions were included in our study, after outlier removal, we were left with 72 (because outliers were not identified in some distributions and others had fewer than 10 effect sizes after outlier removal). n/a = no outliers were identified (Lee et al., 2015) or all severe selection model mean estimates were discarded (Vachon et al., 2014).

TABLE 5
Convergence Rates of the Practical Differences Before and After Outlier Removal

Publication bias method	Before outlier removal			1.	2.	3.	4.	5.	After outlier removal		
	negligible	moderate	severe						negligible	moderate	severe
1. FE trim and fill (t&f$_{FE}$ $\bar{r}_o$)	67 (54%)	28 (23%)	28 (23%)	- -	62 (86%)	46 (64%)	43 (60%)	45 (63%)	33 (46%)	22 (31%)	17 (24%)
2. RE trim and fill (t&f$_{RE}$ $\bar{r}_o$)	86 (70%)	20 (16%)	17 (14%)	93 (76%)	- -	45 (63%)	39 (54%)	38 (53%)	34 (47%)	25 (35%)	13 (18%)
3. Moderate selection model (sm$_m$ $\bar{r}_o$)	89 (72%)	16 (13%)	18 (15%)	74 (60%)	81 (66%)	- -	37 (51%)	32 (44%)	32 (44%)	30 (42%)	10 (14%)
4. Five most precise samples (pr $\bar{r}_o$)	74 (60%)	25 (20%)	24 (20%)	85 (69%)	75 (61%)	80 (65%)	- -	35 (49%)	36 (50%)	21 (29%)	15 (21%)
5. PET-PEESE (pp $\bar{r}_o$)	36 (29%)	45 (37%)	42 (34%)	59 (48%)	47 (38%)	50 (41%)	54 (44%)	- -	27 (38%)	12 (17%)	33 (46%)

Note. There were 123 meta-analytic distributions included in our study. After outlier removal, 72 distributions were analyzed because outliers were not identified in some distributions and others had fewer than 10 effect sizes after outlier removal. Columns 2 and 10 show the number of times each publication bias detection method observed a 'negligible' practical difference in the corresponding naïve meta-analytic mean effect size estimate before and after outlier removal, respectively. Columns 3 and 11 show the number of times each publication bias detection method observed a 'moderate' practical difference in the corresponding naïve meta-analytic mean effect size estimate before and after outlier removal, respectively. Columns 4 and 12 show the number of times each publication bias detection method observed a 'severe' practical difference in the corresponding naïve meta-analytic mean effect size estimate before and after outlier removal, respectively. Values in parentheses in Columns 2, 3, 4, 10, 11, and 12 show the number of times in percentages each practical difference was detected. Columns 5-9 report the inter-publication bias detection method convergence rates across the three levels of practical difference before (below diagonal, $k = 123$) and after (above diagonal, $k = 72$) outlier removal. Higher values below and above the diagonal represent higher rates of convergence between publication bias detection methods. A comparison of below and above the diagonal values indicates whether or not convergence improved following outlier removal.

FIGURE 1
Search and Winnowing Process

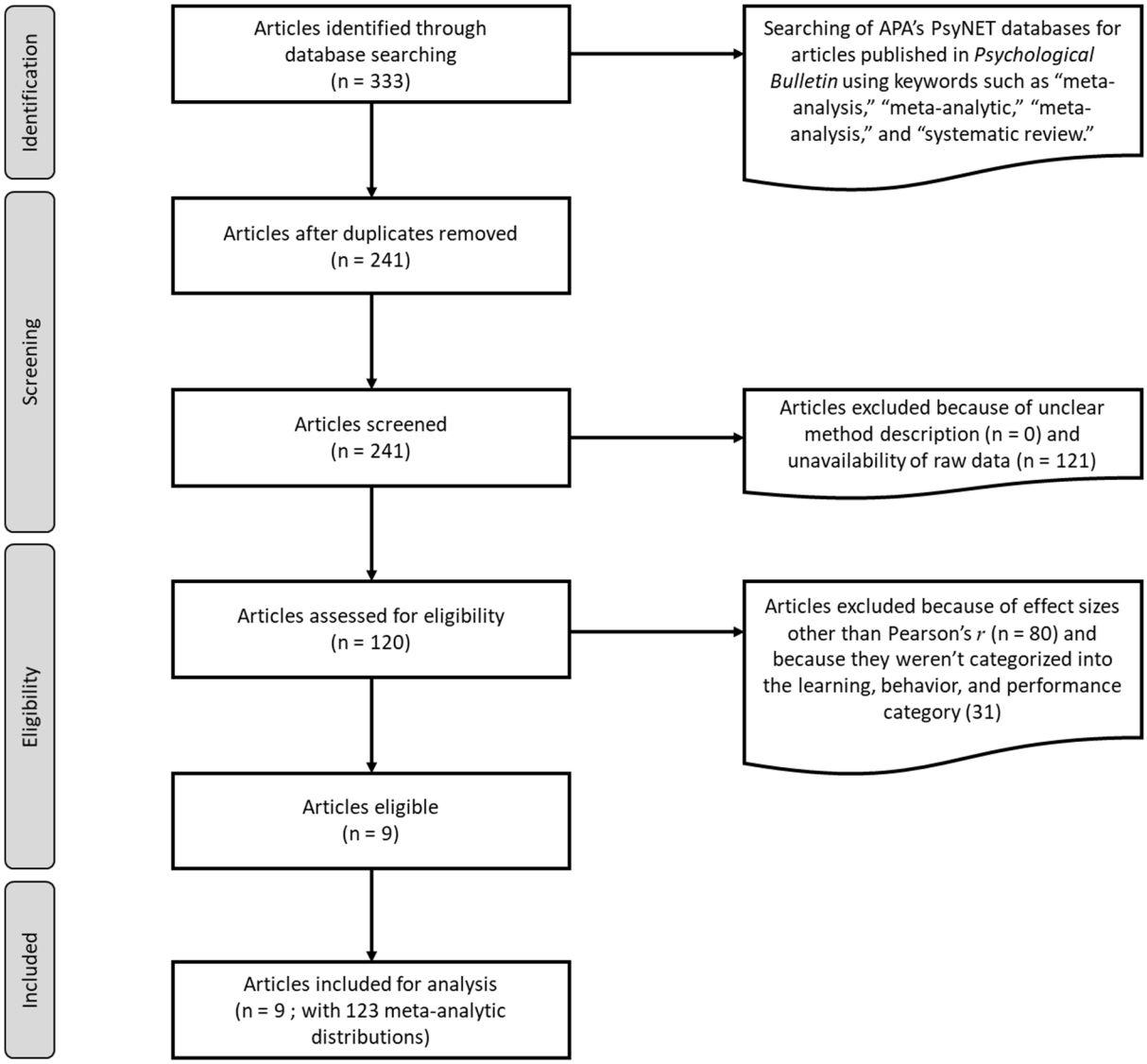
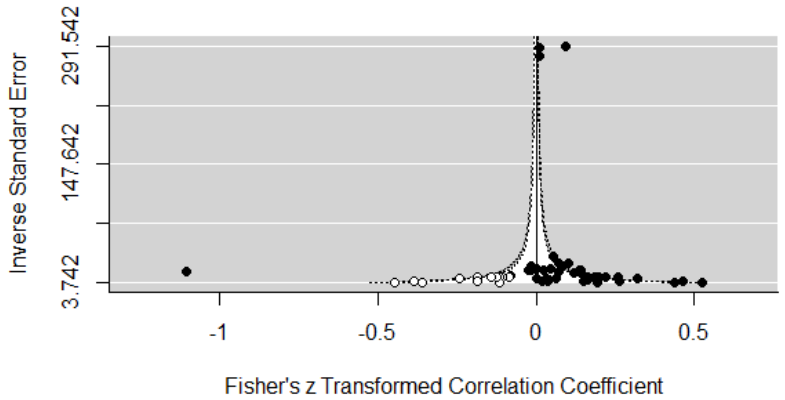


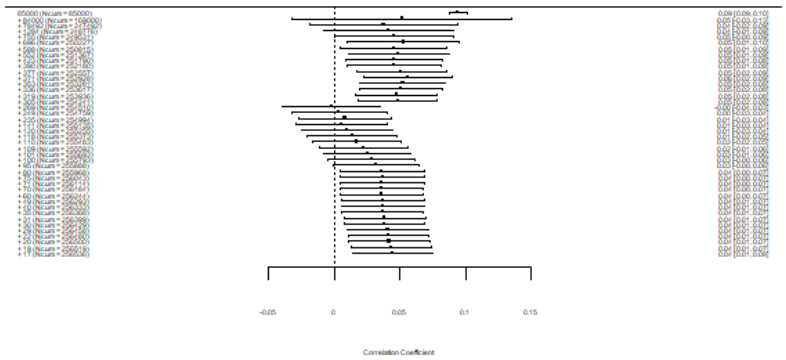
FIGURE 2
Contour-enhanced Funnel Plots (with FE Trim and Fill Imputations) and Forest Plots (Displaying the Cumulative Meta-analysis by Precision) for Select Distributions from Karlin et al. (2015)

(a) Overall effect (*k*=42)

Funnel plot (with trim and fill imputations)

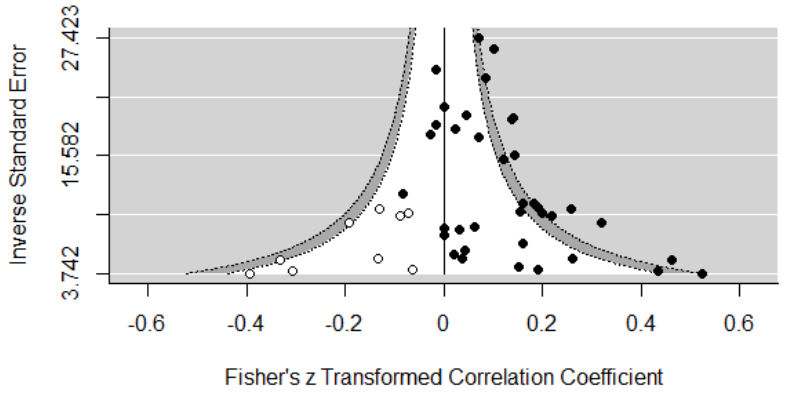


Cumulative meta-analysis by precision

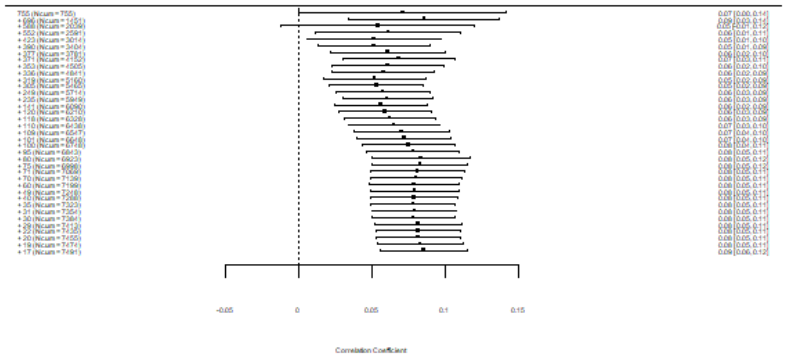


(b) Overall effect - without identified outliers (*k*=37)

Funnel plot (with trim and fill imputations)

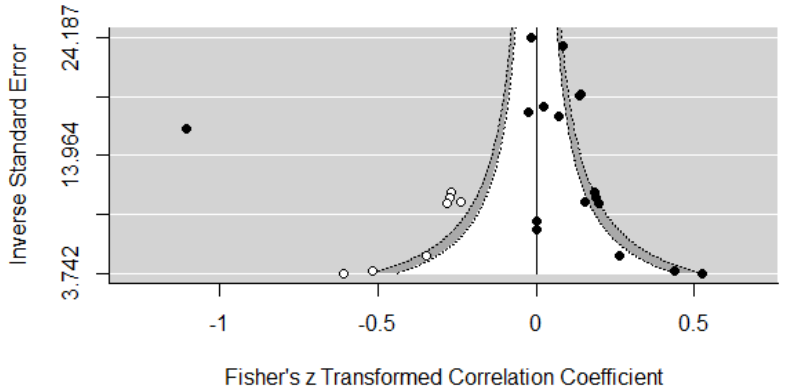


Cumulative meta-analysis by precision

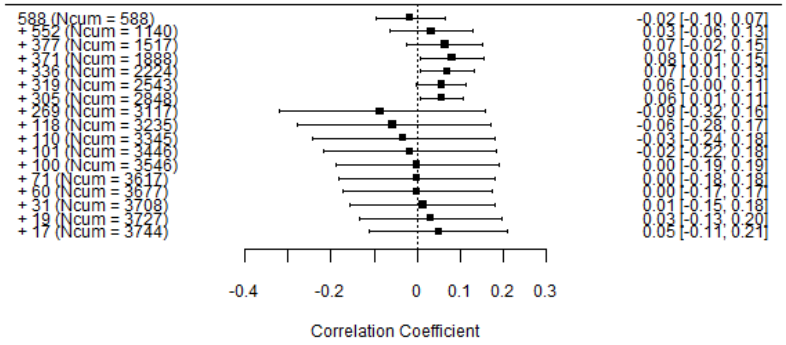


(c) Treatment variables: Feedback frequency: Continuous (*k*=17)

Funnel plot (with trim and fill imputations)

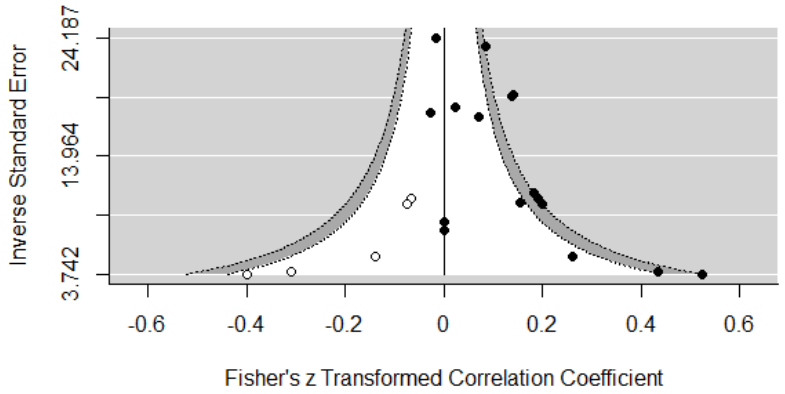


Cumulative meta-analysis by precision



(d) Treatment variables: Feedback frequency: Continuous - without identified outliers (*k*=16)

Funnel plot (with trim and fill imputations)



Cumulative meta-analysis by precision

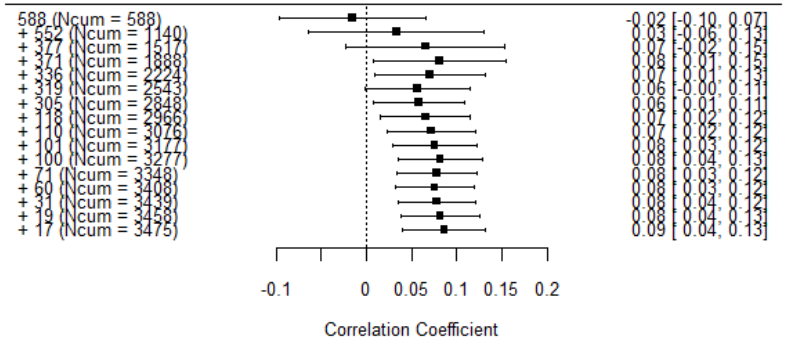


FIGURE 3
Dispersion of the Naïve Meta-analytic Mean Estimate and the Estimates from the Sensitivity Analyses Results Before and After Outlier Removal for Karlin et al.'s (2015) Distributions