

Is there evidence of publication biases in JDM research?

Frank Renkewitz*

Heather M. Fuchs*

Susann Fiedler†

Abstract

It is a long known problem that the preferential publication of statistically significant results (publication bias) may lead to incorrect estimates of the true effects being investigated. Even though other research areas (e.g., medicine, biology) are aware of the problem, and have identified strong publication biases, researchers in judgment and decision making (JDM) largely ignore it. We reanalyzed two current meta-analyses in this area. Both showed evidence of publication biases that may have led to a substantial overestimation of the true effects they investigated. A review of additional JDM meta-analyses shows that most meta-analyses conducted no or insufficient analyses of publication bias. However, given our results and the rareness of non-significant effects in the literature, we suspect that biases occur quite often. These findings suggest that (a) conclusions based on meta-analyses without reported tests of publication bias should be interpreted with caution and (b) publication policies and standard research practices should be revised to overcome the problem.

Keywords: meta-analysis, publication bias, funnel plot, SVO, cooperation.

1 Introduction

In comparison to statistically non-significant results, a larger proportion of significant results overestimate the underlying population effect. It is a long known and repeatedly discussed problem (Greenwald, 1975; Hedges & Olkin, 1985; Iyengar & Greenhouse, 1988; Light & Pillemer, 1984) that a preferential publication of significant studies will therefore lead to a literature that provides a false impression regarding the robustness and size of the effect in question. When non-significant results are largely or completely excluded from publication, even non-existing effects may appear substantial (Rosenthal, 1979; for a recent survey on other selection problems that may affect paradigmatic research and their possible consequences, see Fiedler, 2011).

Strong, direct evidence of an overrepresentation of significant results in scientific literature originates primarily from the area of medicine, where several representative samples of all studies investigating a specific research question have become available. These samples consist of studies that are registered with drug licensing agencies, funding agencies and institutional review boards. Several surveys compared the results of these registered studies with the data that were eventually published. A recent and particularly impressive example is the survey

by Turner et al. (2008). The data base of this survey consists of 74 clinical trials on the effect of antidepressant agents that were registered with the Food and Drug Administration (FDA) in the United States. According to the FDA, 38 studies reported statistically significant primary results. Thirty-seven of these studies were eventually published. In contrast, of the 36 studies reporting non-significant main results, 22 remained unpublished. An additional 11 of these studies appeared in scientific journals but reported—in contradiction to the FDA records—significant main outcomes (in these studies, the dependent variables considered to be the most relevant were exchanged). The combined effect size of the registered studies was $g=.31$. In the published literature, however, this combined effect size was inflated to $g=.41$.

In the field of psychology, surveys of this nature are rare, as individual studies and their results, in particular, are seldom documented in a systematic fashion. However, one survey did employ a similar procedure to assess publication biases (Cooper, de Neve, & Charlton, 1997). The data base consisted of all studies that were approved by the Department of Psychology Human Subjects Committee at a U.S. university between 1986 and 1988. Approximately 50% of the studies reporting significant results were published. Of the studies with non-significant outcomes, however, only 4% were submitted for publication.¹

Thus, there is strong evidence supporting the conclusion that publication biases affect scientific literature in several disciplines (see Palmer, 2000, for examples from

The authors would like to thank Berenike Waubert de Puiseau, Daniel Balliet and an anonymous reviewer for their helpful comments on a previous draft of this paper.

*Department of Psychology, University of Erfurt, Nordhäuser Strasse 63, D-99089 Erfurt, Phone: +49-(0)361/ 737 2223, E-mail: frank.renkewitz@uni-erfurt.de

†Max Planck Institute for Research on Collective Goods.

¹The final publication status of these studies remained unclear. Effect sizes of published or unpublished studies are not reported.

biology) including psychology. Even though the omnipresence of null hypothesis significance tests may be less pronounced in the area of JDM than elsewhere in psychology, it is still a widely used procedure. This and the rarity of non-significant results give good reason to assume that publication biases do occur. Thus, the main question we pursue in this article is whether there is evidence for inflated estimates of effect sizes in the JDM literature that are caused by such biases.

Generally, publication biases pose a threat to the validity of the body of scientific knowledge represented in the literature. However, in the absence of registered studies that may serve as a standard of comparison, the problem may only become apparent when study results are collected for a systematic, quantitative review. Any summary or review of extant literature, including meta-analyses, will inevitably produce an incorrect estimate of the true effect, if the available information represents a selective sample of the relevant research area.² At the same time, meta-analyses provide the opportunity to gauge the extent of the problem. Several methods have been developed that aim to assess whether a collection of effect sizes is affected by publication bias. In the areas of human medicine and biology, several examples of serious publication biases have been identified by using these methods to reanalyze published data (Palmer, 2000; Sutton, Duval, Tweedie, Abrams, & Jones, 2000b). However, in JDM (and psychology as a whole) this problem has been largely ignored.

In the following, we will first provide a brief overview of methods for the detection of publication biases. Then, we will use these methods to reanalyze in some detail one meta-analysis from the area of JDM. Finally, we will explore whether there is evidence of publication bias in other JDM meta-analyses.

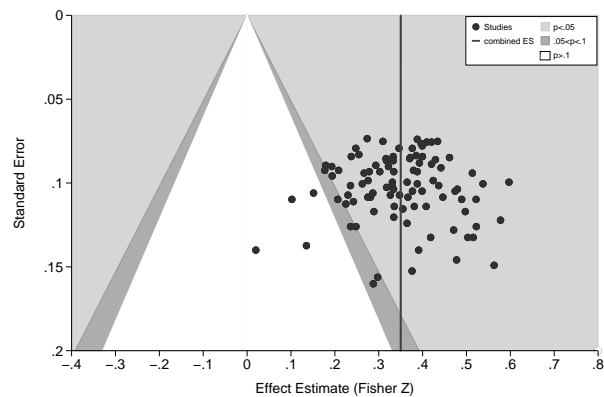
2 Method

Many common methods for the detection of publication bias are based on the funnel plot (Light & Pillemer, 1984)—a simple scatter plot of study effect sizes against a measure of study sample size. In the absence of bias, the data points are symmetrically distributed around the true population effect. The greater variability in effect sizes found in smaller and therefore less precise studies results in the typical inverted funnel shape illustrated in Figure 1.³ However, when significance testing induces a bias,

²Of course, this is also the case when the extant literature provides a representative sample of the relevant research area but the review in question includes only a selective portion of the available literature (cf. Hilbig, 2010).

³Common measures of study sample size used in funnel plots are standard error and precision. For Fisher Z values, which are used as an effect measure throughout this article, the relationship between sample

Figure 1: Funnel plot with a mean effect size of $r = .34$ and no bias. Contours mark the conventional 5% and 10% levels of significance.



some or all of the studies that either reported effect sizes near zero or found small to moderate effect sizes in conjunction with small sample sizes will be missing. Thus, in this case there will be a lack of studies in the lower left-hand side of the distribution of data points; and the plot will appear asymmetrical. Additionally, there will be an association between effect size and study precision, with less precise studies yielding larger effect sizes. Several statistical methods (e.g., Begg's rank correlation) aim to uncover publication bias by assessing this association; two of these (Egger's regression and trim-and-fill analysis) also provide estimates of the true effect adjusted for bias. A description of these methods can be found in the appendix.⁴

It is important to realize, however, that publication bias is not the only possible reason for funnel plot asymmetry. Statistical methods for the assessment of asymmetry are of a correlative nature and thus do not indicate causality. Therefore, a post hoc analysis of the presence of publication bias implies consideration of alternative explanations for such asymmetry and cannot yield definite "proof" of bias. In psychological data sets, the most plausible alternative explanation is typically heterogeneity. The studies comprised in a meta-analysis and in the respective funnel plot may (and often do) estimate different underlying population effects. If, additionally, the true effect is larger

size and standard error is given by $se_z = 1/\sqrt{(N-3)}$. The precision of a study is typically defined as $1/se$. In Figure 1 and all further funnel plots, we use se , following a recommendation by Sterne and Egger (2001). Therefore, the y-axis is inverted to ensure that the dispersion of effect sizes is larger in the bottom part of the funnel plot (as would be the case with sample size or precision).

⁴Funnel plots can be created and statistical tests calculated using both general purpose statistical packages such as Stata and R as well as Comprehensive Meta-Analysis—a stand-alone program for meta-analysis that is somewhat less flexible but has a user-friendly, menu-driven interface.

in smaller studies—for example due to appropriate use of a priori power analysis (e.g., Faul, Erdfelder, Buchner, & Lang, 2009)—this will lead to asymmetry. However, heterogeneity cannot account for a lack of non-significant results in particular from the published literature.⁵ Thus, if non-significant studies appear to be missing in a funnel plot, this lends further credence to the assumption that the asymmetry was indeed caused by publication bias. For this reason, we use contour-enhanced funnel plots (Peters et al., 2006) throughout this article. The contours in these plots denote the conventional 5% and 10% levels of significance based on a two-tailed Z-test (see Figure 1). The p -values resulting from this Z-test may not correspond exactly to the p -values reported in the original studies because these may have used different statistical procedures or have tested the effect of interest only in conjunction with the effect of other factors. Still, the contours will provide a fairly reliable impression of the level of significance of each effect estimate. In addition, we will apply an exploratory procedure recently proposed by Ioannidis and Trikalinos (2007) that tests for a lack of non-significant (or an excess of significant) studies in a body of research. This test is also briefly described in the appendix.

3 A reanalysis of a meta-analysis on the relationship between social value orientation and cooperation

As an example, we present in detail a reanalysis of one recent meta-analysis from the field of JDM. Balliet, Parks, and Joireman (2009) assessed the relationship between social value orientation (SVO, Messick & McClintock, 1968) and cooperation in social dilemmas. The SVO measure describes preferences for different distributions of payoffs to oneself and other persons. Based on these measures (Kuhlman & Marshello, 1975; Liebrand & McClintock, 1988; Van Lange, Otten, De Bruin, & Joireman, 1997), participants are classified either as proselves or prosocials. Whereas proselves attempt to maximize their own (absolute or relative) payoffs, prosocials are interested in maximizing common payoffs.

The meta-analysis comprises 48 reports including 82 lab studies that used experimental games such as the prisoner's, public goods or commons dilemmas to assess the correlation between SVO and cooperation. Twenty-one

⁵Assume that all researchers in a given area have valid intuitions about moderators and are thus able to make correct assumptions regarding the relevant true effect sizes in advance of their studies. Furthermore, all researchers perform a priori power analyses and aim at the conventional power level of 80%. Even under these circumstances, one would still expect 20% of all studies to be non-significant for any given sample size.

of these studies were unpublished. As the main result of a mixed-effects analysis, Balliet and colleagues (2009) report a combined effect size of $r=.30$, indicating that prosocials cooperate more than proselves. They also address the issue of publication bias by computing Orwin's fail-safe N (Orwin, 1983). In general, a fail-safe N represents the number of additional studies with a mean effect of zero⁶ that would be necessary to reduce the combined effect to statistical non-significance (Rosenthal, 1979) or to a size considered trivial (in this case, $r=.04$). Balliet and colleagues report a fail-safe N of 510, on the basis of which they conclude that "the effect size ... appears to be robust against the presence of a large number of unpublished studies finding a null result" (p. 538). However, the fail-safe statistic is deficient, as it does not assess whether the data set actually shows any evidence of publication bias and consequently does not indicate the extent to which the combined effect may have been affected by such bias. Recent reviews generally speak against the use of fail-safe N (Becker, 2005; Higgins & Green, 2009).

To begin a more appropriate analysis of publication bias, we created a funnel plot of the effect sizes reported in the original studies and their standard errors (Figure 2). Focusing only on the published studies, it is apparent from visual inspection that the distribution of the corresponding effect sizes is asymmetrical. The statistical methods confirm this assessment: Begg's rank correlation and Egger's regression find a significant association (α -level of 10%) between effect sizes and their standard errors (see Table 1). Trim-and-fill detects asymmetry with the estimator R_0 and indicates that 23 studies are missing (Table 1). Figure 3 includes the studies that are imputed by the trim-and-fill procedure to obtain a more symmetrical funnel plot. Twenty of the 23 imputed studies are located in the area of non-significance.

Balliet and colleagues (2009) identify two moderators of the combined effect size (payment of participants according to performance and type of game with the levels "give-some" and "take-some" games). However, these identified sources of heterogeneity, at least, cannot account for the observed asymmetry in the set of published studies. In the subsets of studies defined by the moderators, which are all more homogenous than the total set, we find descriptively similar levels of asymmetry. For instance, for the 30 published studies using outcome-dependent payment, Begg's rank correlation is $\tau=.27$ ($p=.02$); while Egger's regression yields $b_1=1.72$ ($p=.06$). For the 22 studies without outcome-dependent payment, the corresponding figures are $\tau=.19$ ($p=.10$) and $b_1=1.95$ ($p=.07$).⁷

⁶With Orwin's fail-safe N , one can also assume that the average magnitude of the additional effect sizes is some value different from zero.

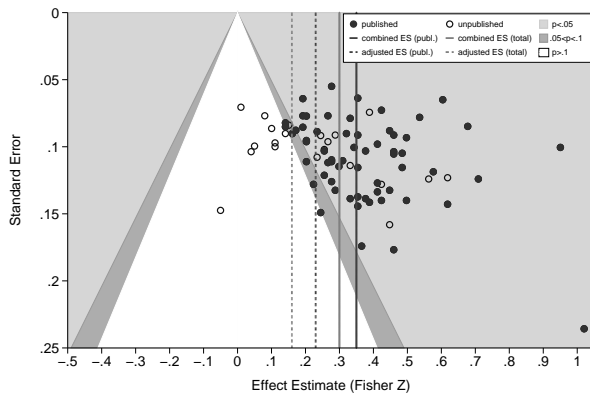
⁷In the two subsets defined by the moderator game type, the results

Table 1: Three indicators of publication bias in the meta-analyses of Balliet et al. (2009) and Dato-on & Dahlstrom (2003)

	Balliet		Dato-on
	published studies	all studies	published studies
Beggs's correlation ($\tau(p)$)	.21 (.01)	.23 (<.001)	.15 (.07)
Eggers regression ($b_1(p)$)	1.28 (.06)	1.59 (.02)	1.19 (.21)
adjusted r	.23	.16	.20
Trim-and-fill			
estimator L_0 (filled studies ($adj. r$))	0 (.35)	22 (.24)	17 (.15)
estimator R_0 (filled studies ($adj. r$))	23 (.25)	3 (.30)	0 (.32)

Note: Results for Dato-on & Dahlstrom are discussed below (see Figure 5 and section on other meta-analysis in JDM research).

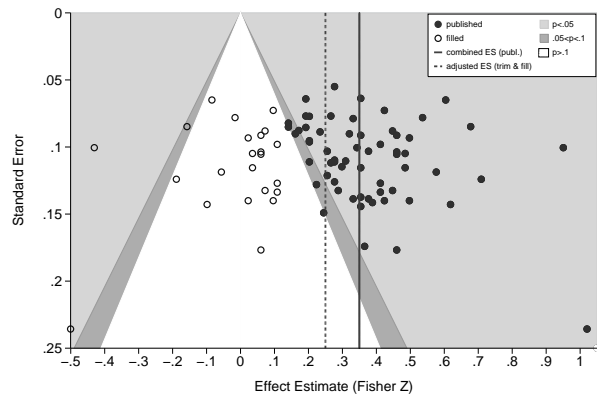
Figure 2: Funnel plot of the effect sizes in the primary studies summarized in the meta-analysis by Balliet and colleagues (2009). The correlational effect sizes were Fisher Z-transformed. The solid lines indicate the combined effect sizes of the published studies and the complete data set. The dashed lines provide the adjusted estimates for these samples of studies, which are based on Egger's regression.



Furthermore, the most striking characteristic of the funnel plot in Figure 2 is that it contains no published

are: $\tau = .11$ ($p = .19$), $b_1 = 1.35$ ($p = .09$) for give-some games (30 studies); and $\tau = .47$ ($p = .01$), $b_1 = 1.76$ ($p = .01$) for take-some games (14 studies). If both moderators are used in conjunction to create homogeneous subsets of studies, the number of effect sizes included in these subsets becomes extremely small. Still, there is a positive association between effect sizes and standard errors in all subsets comprising more than three effects. This association is generally of similar magnitude as in the complete set of published studies. For the 10 studies combining take-some games with outcome-dependent payment, it is even larger ($\tau = .78$, $p = .001$; $b_1 = 2.15$, $p = .002$). A portion of these results are influenced by individual outliers. However, eliminating the outliers does not consistently change the results (i.e. it leads to a larger indication of bias in some cases and a smaller in others).

Figure 3: Funnel plot of the effect sizes in the published studies summarized by Balliet and colleagues (2009). The white dots are studies imputed by the trim-and-fill procedure (estimator R_0). The dashed line indicates the adjusted estimate of the combined effect size resulting from the trim-and-fill procedure.



effect sizes associated with two-tailed p -values $> .10$ (or one-tailed p -values $> .05$). In other words, all published studies were at least “marginally significant”. The exploratory test by Ioannidis and Trikalinos (2007) also indicates a lack of non-significant effect sizes ($p = .001$, binomial test).⁸ In conjunction with the asymmetric distribution of effect sizes, this constitutes strong evidence that the literature on SVO is biased due to an exclusion of non-significant results. This conclusion is further corroborated by a comparison of the results in published and unpublished studies. Approximately one half of the ef-

⁸The power calculations for primary studies needed in this test were based on the combined effect size in the complete data set ($r = .30$). If, instead, the combined effect size from the published studies ($r = .35$) is used, the test result is $p = .05$.

fect sizes in unpublished studies are non-significant. The combined effect size ($r=.21$) of unpublished studies is significantly smaller than the combined effect size of published studies ($r=.35$), $Q(1)=11.81$, $p=.001$. Thus, the evidence suggests that the combined effect size in published studies overestimates the true effect. Based on the published studies alone, Egger's regression and trim-and-fill (estimator R_0) yield adjusted estimates of the combined effect size of $r=.23$ and $r=.25$, respectively (see Table 1). These adjusted estimates do not warrant a final conclusion regarding the magnitude of the population effect.⁹ Rather, they should be regarded as a form of sensitivity analysis—large corrections may indicate a lack of robustness. Still, the difference between the combined effect of $r=.35$ in the published studies and the adjusted combined effects reveals that the exclusion of non-significant studies led to an overestimation of the effect of SVO on cooperation that may be of a theoretically and practically relevant magnitude.

By including unpublished studies, Balliet and colleagues (2009) follow the most prominent advice for preventing biased meta-analytical results. However, the success of this approach depends on the representativeness and size of the sample of unpublished studies included. In this example, there still appear to be non-significant and negative effect sizes missing despite the inclusion of 21 unpublished studies (see Figure 2). The results of the statistical methods confirm that the asymmetry in the funnel plot is not reduced by including unpublished studies (see Table 1). On the contrary, because many of the unpublished studies have relatively small standard errors and report negligible effect sizes, the adjusted combined effect size of Egger's regression is even reduced to $r=.16$ in the complete data set.

An additional, interesting aspect of our reanalysis stems from a test of a third moderator hypothesis reported by Balliet and colleagues (2009). Contrary to their expectations, this test reveals that the combined effect sizes in one-shot ($r=.31$) and iterated ($r=.29$) games are similar in magnitude and not significantly different. Clearly, both sets of published studies are asymmetrical and lack non-significant results (see Figure 4). However, only the combined effect size in one-shot games is corrected by the inclusion of unpublished results, whereas the combined effect in iterated games remains almost constant. Therefore, the failure to find a moderator effect for experimen-

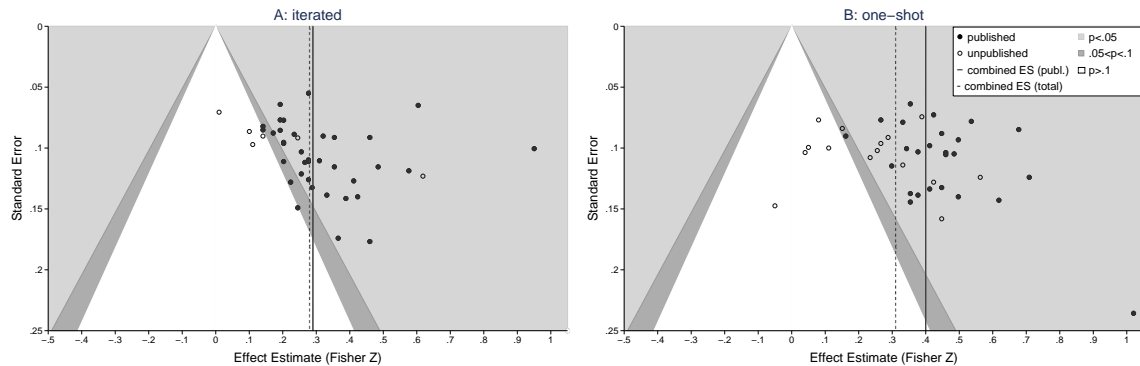
tal game repetition may be interpreted as the result of an unnoticed, selective correction for publication bias in the sample of one-shot games. Indeed, there is a significant moderator effect when one considers only the published studies, $r=.40$ for one shot games and $r=.28$ for iterated games, $Q(1)=20.58$, $p < .001$.

A selective correction for publication bias would only be justified when one sample (in this case one-shot games) is more strongly biased than the other. A closer look at the data, however, indicates the exact opposite. A large proportion of the effects from published studies on iterated games are located in the area of "marginal significance" or directly below the conventional 5% significance criterion (see Figure 4); this is not the case for the one-shot sample. Thus, the plots suggest that researchers had more difficulty achieving significant results when using iterated games. However, the statistical tests of asymmetry are not clear regarding the magnitude of bias in the two data sets. This is primarily due to the fact that, in these reduced data sets, the results of the regression methods and trim-and-fill are strongly influenced by single effect sizes (the largest effects in both sets). Generally, Begg's rank correlation is the most robust against outliers. Indeed, here it is the only method that yields fairly stable results independent of the exclusion of these effects. It indicates a markedly stronger bias among published studies on iterated games ($\tau=.38$, $p < .001$) than among published studies on one-shot games ($\tau=.17$, $p=.12$). Based on this result, one may conclude that the true moderator effect is even larger than it appears based on the published studies.

Although the reanalysis of the moderator data certainly leaves room for interpretation, the key issue with respect to the topic of publication bias is that the available data of 61 published and 21 unpublished studies based on a total of 8,815 participants do not resolve this issue in an unambiguous manner. Balliet and colleagues (2009) point out that their meta-analysis is the first quantitative summary of 40 years of research on the relationship between SVO and cooperation. What have we learned about this relationship from this extensive research effort? Given the results of our reanalysis, it still seems safe to conclude that there is a positive correlation between SVO and cooperative behavior in social dilemmas. All of our adjusted estimates were positive; and the unpublished studies are all, with the exception of a single effect size (of negligible magnitude), positive. However, as our results suggest that studies with non-significant results were excluded from publication and were, consequently, to some degree unavailable, we can be confident that the true combined effect size of studies on SVO is smaller than $r=.35$ resulting from published studies alone and even smaller than $r=.30$ reported in the meta-analysis. However, as long as we do not know exactly which results are missing, it will be dif-

⁹Strictly speaking, neither the combined effect size of the published studies nor the adjusted estimates of Egger's regression and trim-and-fill can be understood as an estimate of an underlying population effect when moderator variables exist. The combined effect size of the published studies simply represents their weighted mean. The adjusted estimates may be understood as the weighted mean of all studies that were actually conducted (under the assumption that the proportion of studies employing different levels of the moderators is equal in published and unpublished studies).

Figure 4: Funnel plots of the effect sizes in the primary studies according to game form (iterated vs. one-shot)



difficult to “guess” the true effect magnitude. The most conservative corrected estimate of the combined effect we calculated was $r=.16$. This discrepancy in effect magnitude would reflect a reduction in variance accounted for from approximately 10 percent to approximately two percent, which certainly appears practically meaningful. In addition, the results of moderator analyses must be viewed with skepticism, as it is to be expected that different samples of studies will be affected differently by publication bias. Making use of available unpublished studies might be helpful in identifying these problems. However, it will fail to solve them unless the available studies are a representative and sufficiently large sample of all unpublished studies.

4 What about other meta-analyses in JDM?

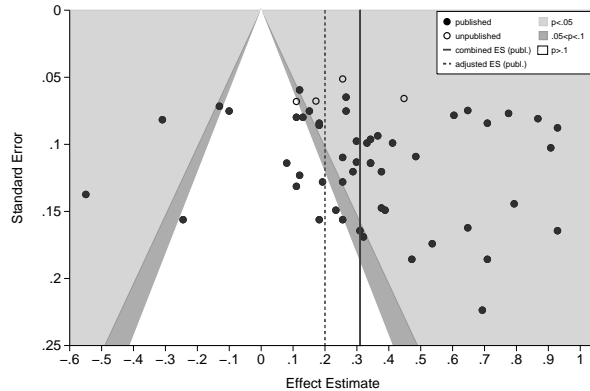
We reanalyzed one additional meta-analysis from the field of JDM (Dato-on & Dahlstrom, 2003) that addresses contrast effects in judgments. The main research question is whether more extreme primes cause more moderate judgments about target stimuli and, subsequently, larger contrast effects. The meta-analysis comprises 55 studies from 27 published articles and three dissertations. It reports a fixed combined effect of $r=.29$. In the funnel plot, this sample of studies is also characterized by substantial asymmetry (see Figure 5). Although several effect sizes are non-significant, there is still a gap in the plot between the positive and negative effects located in approximately the middle of the area of non-significance. Begg’s rank correlation indicates a positive association between effect sizes and their standard errors (see Table 1). Egger’s regression also indicates a strong yet non-significant relationship. Trim-and-fill detects asymmetry with the estimator L_0 . The adjusted combined effects provided by Egger’s regression and trim-and-fill (estima-

tor L_0) are $r=.20$ and $r=.15$, respectively. The exploratory test by Ioannidis and Trikalinos (2007) does not indicate a lack of non-significant studies when the original estimate of the combined effect size ($r=.29$) is used to calculate the power of the primary studies. However, when the power calculation is based on the lower bound of the 95% confidence interval of the combined effect size ($r=.25$), the binomial test results in a $p=.09$. Taken together, the evidence from the statistical methods is less conclusive than in the reanalysis of the meta-analysis by Balliet and colleagues (2009). However, it still raises doubts concerning the validity of the meta-analytical results. The statistical methods suggest that the distribution of effect sizes is asymmetric and that this asymmetry may be caused by an exclusion of non-significant effects. As a result, the combined effect size reported in the meta-analysis may overestimate the true effect size. At the very least, the meta-analytical results should be interpreted with great caution.¹⁰

A further example of publication bias that is widely recognized in the area of JDM was identified by Acker (2008), who conducted a meta-analysis on unconscious thought theory (Dijksterhuis & Nordgren, 2006). The central tenet of this theory is that unconscious thought will lead to better performance than conscious thought in complex decision tasks. Acker collected data from 17 studies that allowed for a comparison of performance following conscious and unconscious thought. Only six of these studies (Dijksterhuis, 2004; Dijksterhuis, Bos, Nordgren, & Van Baaren, 2006) were published when the meta-analysis was conducted. These published studies uniformly found evidence in favor of unconscious

¹⁰Dato-on and Dahlstrom indicated three significant moderators in their meta-analysis. The subsets of studies defined by these moderators generally did not show less heterogeneity than the total data set. Thus, it is unlikely that these moderators can explain the asymmetry present in Figure 4. However, this implies the existence of additional moderator variables that were not identified in the meta-analysis, the influence of which remains unknown.

Figure 5: Funnel plot of the effect sizes in the primary studies summarized in the meta-analysis by Dato-on and Dahlstrom (2003). The adjusted estimate of the combined effect size is based on Egger's regression.



thought theory, which was in some cases statistically significant. In contrast, most of the unpublished studies found smaller effect sizes, six of which were negative. The combined effect sizes (estimated with a random effects model) of the published and unpublished studies are $g=.43$ and $g=.14$, respectively.¹¹ While Acker applied no formal methods for the assessment of publication bias, he additionally noted “that the experiments with fewer participants consistently generated substantially larger effect sizes than the larger studies” (p. 301). Indeed, the negative relationship between the precision of the studies and their effect sizes is strong (Begg: $\tau = .38$, $p=.02$; Egger: $b_1 = 4.25$, $p=.02$), suggesting that even the complete data set produces a biased combined effect ($g=.25$). It seems noteworthy to us that the two studies yielding the least precise estimates yet exceptionally large effect sizes (Dijksterhuis et al., 2006) were published in *Science*.

Although the meta-analyses discussed above strongly suggest that publication biases affect research results in the area of JDM,¹² they represent only a small, certainly non-representative sample of research in this area. Therefore, we searched PsychInfo for additional JDM

¹¹We computed these figures from the effect sizes and standard errors provided in Acker (2008).

¹²The meta-analysis by Acker (2008) differs from those on SVO and contrast effects in that it summarizes research on a fairly new hypothesis. Therefore, it cannot demonstrate that a number of studies have apparently vanished permanently from the literature. Indeed, all of the studies included by Acker were published after the meta-analysis appeared. Still, it illustrates a publication bias in the sense that significant yet less precise studies were published earlier than non-significant studies (see Sutton, 2005, for similar examples in the area of medicine). It may also have contributed to preventing a more enduring publication bias. At the very least, it seems that there are now more non-significant and negative results published on unconscious thought theory than on the relationship between SVO and cooperation or contrast effects in judgment.

meta-analyses using the keywords “judgment” or “decision making” in combination with the methodology “meta analysis”, in order to determine which methods of bias detection were used as well as whether evidence of bias was found. Of the resulting 120 manuscripts, many represented studies that either did not conduct a meta-analysis or were not of interest to the JDM community. Of the remaining studies, we selected 12 meta-analyses that we deemed relevant to core JDM research (references marked with an asterisk).

Eight meta-analyses either completely ignored the problem of publication bias or conducted only a fail-safe N analysis, which, as discussed above, does not represent an appropriate analysis of bias. Two meta-analyses conducted tests of publication bias using moderator analyses. Greenwald, Poehlman, Uhlmann and Banaji (2009) compared published and unpublished effects and found no significant difference. Spengler and colleagues (2009) compared effects published in APA journals with those published elsewhere and found that those in APA journals were significantly larger. Finally, only two meta-analyses used a portion of the methods discussed in this paper. Thornton and Dumke (2005) found no indication of bias using a funnel plot and a correlational analysis. Karelaia and Hogarth (2008) assessed bias with Begg’s rank correlation, trim-and-fill analysis and funnel plots; the former two showed evidence of bias in some subgroups.

Overall, we were able to find only four additional meta-analyses that allow for inferences regarding the presence of publication bias in JDM research. Two of these reported evidence of publication bias.

5 Discussion

The results of our reanalyses strongly suggest that publication biases also occur in the field of JDM. Both of the data-sets we reanalyzed showed evidence of bias. Also, a third example of a bias was previously demonstrated in a meta-analysis (Acker, 2008) on unconscious thought theory (Dijksterhuis, 2004). In every case, statistically non-significant results were underrepresented in the literature—at least at the time when data for the meta-analyses were collected. This bias against non-significant results is certain to yield an inflated estimate of the underlying effect when published effects are aggregated.

Our selection of meta-analyses was more or less arbitrary and guided mainly by practical considerations (our main criterion was whether a meta-analysis seemed to provide all necessary information from the primary studies). Thus, the question remains how many effects in JDM are affected by publication biases and, consequently, appear more stable and relevant in the literature than they truly are. Our survey of meta-analyses from the

field shows that there is currently no empirical answer to this question. Most meta-analyses ignored the problem of publication bias or assessed it with unsound methods. Of the four meta-analyses we located that allow for any assertion, two found indications of bias. Any claim regarding the prevalence of publication biases in JDM must remain speculative, as we lack data collections on effects in JDM that were investigated for publication bias with scrutiny. However, given our results and the rareness of non-significant effects in the literature, we suspect that biases occur quite often.

One obvious conclusion from our findings is the need for a greater awareness of this problem in JDM. “Established” effects may turn out to be less relevant once they are tested for publication bias. Meta-analyses should generally perform a thorough and methodologically sound assessment of publication bias and address the issue when discussing their results. Funnel plots should be displayed in all meta-analytical reports, as they provide information not considered by any of the statistical methods for the assessment of funnel plot asymmetry (most notably, whether studies are missing in areas of non-significance) but simultaneously allow for some degree of subjective interpretation. Obviously, the results of meta-analyses that do not present an appropriate investigation of publication bias must be interpreted very carefully.

Another implication of our findings is that publication decisions, at least in some areas of JDM, rely heavily on the results of significance tests. It is this reliance as well as the focus on the question “Is there an effect?” that leads to a body of empirical findings that provides a distorted impression concerning the stability and size of the effect in question. Interestingly, evidence from psychology (Cooper, DeNeve, & Charlton, 1997) and medicine (Dickersin, 1997) suggests that publication biases are mainly caused by the reluctance of researchers to submit non-significant results rather than by the rejection of non-significant results during the peer-review process. The reliance on null hypothesis significance testing is particularly worrisome, as the error rate of the significance test may be very large (Ioannidis, 2005). This high error rate in some scientific fields reflects the simple fact that individual studies with limited sample size are often not capable of yielding conclusive evidence in favor of a research hypothesis. Given that there currently appears to be a preference for the publication of “positive” findings in many scientific fields, it may be advisable to evaluate such findings using statistical methods that do not provide the premature impression of clear-cut results but rather more explicitly illustrate the uncertainty inherent in the statistical inference. In this respect, it might be helpful to focus more strongly on effect sizes and their confidence intervals even though confidence intervals imply the same inference about the null hypothesis as significance tests

(Cummings & Finch, 2001). A more sophisticated alternative that, in our view, evaluates the available evidence more appropriately than significance testing and that can also lead to different conclusions regarding the null hypothesis is Bayesian statistics (e.g., Rouder, Speckman, Sun & Morey, 2009).

However, to overcome the problem of publication biases, the choice of suitable statistical methods will be less important than a broad recognition of the fact that publication decisions should not depend on the question of whether the data favor a specific hypothesis. Publication decisions should be based primarily on theoretical relevance, hypothesis plausibility and methodological quality—and not on significant or, more generally, positive findings. The only characteristic of study results that should be relevant for publication is study precision. After all, a study of sound methodological quality that yields a precise estimate of an effect is informative even if the confidence interval includes zero; and it is always more informative than a study yielding a significant, but imprecise effect estimate with a huge confidence interval.

So, what could be done to alleviate the problem of publication biases? Most effective measures will involve a change in publication policies and incentive schemes in science. For instance, independent and exact replication studies should be easier to publish and more highly valued. Such replication studies are the best possibility to support or refute previous findings; and even a small number of replication studies will allow for a much more reliable assessment of the true effect size in a meta-analysis if there is no selective reporting (Palmer, 2000). With regard to a preference for positive findings in the peer review process, an interesting solution might be the introduction of “result blind reviews”. Such a procedure would ensure that the publication decision is based solely on theoretical relevance, methodological quality of the design and appropriateness of the suggested statistical analysis. While it may be unrealistic to propose that all research in JDM (or even psychology) should be evaluated without regard to its results, studies that undergo a result blind review are likely to produce more objective and reliable results and, thus, should be more highly esteemed. Finally, given the reluctance of researchers to submit non-significant results, it seems safe to conclude that “exploratory testing” is at least one of the driving forces behind publication bias: Hypotheses are tested several times—by using several statistical methods, adding covariates and factors, including several but interchangeable dependent variables, forming subgroups, excluding (extreme) data points, screening data transformations, or simply running multiple studies—but only significant results are reported. Thus, researchers should be encouraged to publicly document their hypotheses and methods in detail *before* an experiment

is done. Schooler (2011) recently proposed an open-access repository for all research findings for this purpose. Again, this might not be a viable option for all JDM research. But a study that is fully described in advance yields more compelling evidence and, therefore, should be easier to publish in more prestigious and widely recognized journals.

In general, any measure that advances the publication and availability of negative results will finally lead to more reliable and trustworthy research findings—and will thus improve the quality of our research field.

References

- Acker, F. (2008). New findings on unconscious versus conscious thought in decision making: Additional empirical data and meta-analysis. *Judgment and Decision Making, 3*, 292–303.
- Balliet, D., Parks, C., & Joireman, J. (2009). Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes & Intergroup Relations, 12*, 533–547.
- Becker, B. (2005). Failsafe N or file-drawer number. In H. Rothstein, A. Sutton & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments* (pp. 111–126). Chichester: Wiley.
- Begg, C., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*, 1088–1101.
- Borenstein, M. (2005). Software for publication bias. In H. Rothstein, A. Sutton & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments* (pp. 193–220). Chichester: Wiley.
- *Christensen-Szalanski, J., & Fobian Willham, C. (1991). The hindsight bias: A meta-analysis. *Organizational Behavior and Human Decision Processes, 48*, 147–168.
- Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods, 2*, 447–452.
- *Covey, J. (2007). A meta-analysis of the effects of presenting treatment benefits in different formats. *Medical Decision Making, 27*, 638–654.
- Cumming, G., & Finch, S. (2001). A primer on the understanding, use, and calculation of confidence intervals that are based on central and noncentral distributions. *Educational and Psychological Measurement, 61*, 530–572.
- Dato-on, M., & Dahlstrom, R. (2003). A meta-analytic investigation of contrast effects in decision making. *Psychology and Marketing, 20*, 707–731.
- Dickersin, K. (1997). How important is publication bias? A synthesis of available data. *AIDS Education and Prevention, 9*, 15–21.
- Dijksterhuis, A. (2004). Think different: The merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology, 87*, 586–598.
- Dijksterhuis, A., Bos, M., Nordgren, L., & Van Baaren, R. (2006). On making the right choice: The deliberation-without-attention effect. *Science, 311*, 1005–1007.
- Dijksterhuis, A., & Nordgren, L. (2006). A theory of unconscious thought. *Perspectives on Psychological Science, 1*, 95–109.
- Duval, S. (2005). The trim-and-fill method. In H. Rothstein, A. Sutton & M. Borenstein (Eds.), *Publication bias in meta analysis: Prevention, assessment and adjustments* (pp. 127–144). Chichester: Wiley.
- Duval, S., & Tweedie, R. (2000a). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*, 89–98.
- Duval, S., & Tweedie, R. (2000b). Trim and fill: a simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*, 455–463.
- Egger, M., Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British medical journal, 315*, 629–634.
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160.
- Fiedler, K. (2011). Voodoo correlations are everywhere—not only in neuroscience. *Perspectives on Psychological Science, 6*, 163–171.
- *Greenwald, A., Poehlman, T., Uhlmann, E., & Banaji, M. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of Personality and Social Psychology, 97*, 17–41.
- Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*, 1–12.
- Hedges, L., & Olkin, I. (1985). *Statistical methods for meta-analysis* Academic Press New York.
- Higgins, J., & Green, S. (2009). *Cochrane handbook for systematic reviews of interventions* New York: Wiley.
- Hilbig, B. E. (2010). Reconsidering “evidence” for fast-and-frugal heuristics. *Psychonomic Bulletin & Review, 17*, 923–930.
- *Hogarth, R., & Karelaia, N. (2007). Heuristic and linear models of judgment: Matching rules and environments. *Psychological Review, 114*, 733–758.

- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine* 2, e124.
- Ioannidis, J. P. A., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical Trials*, 4, 245–253.
- Iyengar, S., & Greenhouse, J. (1988). Selection models and the file drawer problem. *Statistical Science*, 3, 109–117.
- *Karelaia, N., & Hogarth, R. (2008). Determinants of linear judgment: A meta-analysis of lens model studies. *Psychological Bulletin*, 134, 404–426.
- *Kaufmann, E., & Athanasou, J. (2009). A Meta-Analysis of Judgment Achievement as Defined by the Lens Model Equation. *Swiss Journal of Psychology*, 68, 99–112.
- *Kühberger, A. (1998). The Influence of Framing on Risky Decisions: A Meta-analysis. *Organizational Behavior and Human Decision Processes*, 75, 23–55.
- Kuhlman, D. M., & Marshello, A. F. (1975). Individual differences in game motivation as moderators of preprogrammed strategy effects in prisoner's dilemma. *Journal of Personality and Social Psychology*, 32, 922–931.
- Liebrand, W. B. G., & McClintock, C. G. (1988). The ring measure of social values: A computerized procedure for assessing individual differences in information processing and social value orientation. *European Journal of Personality*, 2, 217–230.
- Light, R., & Pillemer, D. (1984). *Summing up: the science of reviewing research*. Cambridge: Harvard University Press.
- Macaskill, P., Walter, S., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta analysis. *Statistics in Medicine*, 20, 641–654.
- Messick, D., & McClintock, C. G. (1968). Motivational Bases of Choice in Experimental Games. *Journal of Experimental Social Psychology*, 4, 1–25.
- Moreno, S., Sutton, A., Ades, A., Stanley, T., Abrams, K., Peters, J., & Cooper, N. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC medical research methodology*, 9, 2–20.
- Orwin, R. (1983). A fail-safe N for effect size in meta-analysis. *Journal of Educational Statistics*, 8, 157–159.
- Palmer, A. (2000). Quasireplication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. *Annual Review of Ecology and Systematics*, 31, 441–480.
- Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, 61, 991–996.
- Rosenthal, R. (1979). The “file drawer problem” and tolerance for null results. *Psychological bulletin*, 86, 638–641.
- Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, 16, 225–237.
- Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, 470, 437.
- *Schwenk, C. (1990). Effects of devil's advocacy and dialectical inquiry on decision making: A meta-analysis. *Organizational Behavior and Human Decision Processes*, 47, 161–176.
- *Spengler, P., White, M., Ægisdóttir, S., Maugherman, A., Anderson, L., Cook, R., et al. (2009). The meta-analysis of clinical judgment project. *The Counseling Psychologist*, 37, 350–382.
- Sterne, J., Becker, B., & Egger, M. (2005). The funnel plot. In H. Rothstein, A. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 75–98). Chichester: Wiley.
- Sterne, J., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, 54, 1046–1055.
- Sterne, J., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. Sutton & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). Chichester: Wiley.
- Sterne, J., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology*, 53, 1119–1129.
- Sutton, A. (2005). Evidence concerning the consequences of publication and related biases. In H. Rothstein, A. Sutton & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis-Prevention, Assessment and Adjustments*. (pp. 175–192). Chichester: Wiley.
- Sutton, A., Duval, S., Tweedie, R., Abrams, K., & Jones, D. (2000). Empirical assessment of effect of publication bias on meta-analyses. *British Medical Journal*, 320, 1574–1577.
- *Thompson, L., & Hrebec, D. (1996). Lose-lose agreements in interdependent decision making. *Psychological Bulletin*, 120, 396–409.
- *Thornton, W., & Dumke, H. (2005). Age differences in everyday problem-solving and decision-making effectiveness: A meta-analytic review. *Psychology and Aging*, 20, 85–99.
- Turner, E. H., Matthews, A. M., Linardatos, E., Tell, R. A. & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent ef-

ficacy. *New England Journal of Medicine*, 358, 252–260.

Van Lange, P., Otten, W., De Bruin, E. M. N., & Joireman, J. A. (1997). Development of prosocial, individualistic, and competitive orientations: Theory and preliminary evidence. *Journal of Personality and Social Psychology*, 73, 733–746.

Appendix

Begg’s rank correlation (Begg & Mazumdar, 1994) uses Kendall’s tau to measure the correlation between standardized effect sizes and their variances. The standardization is necessary to stabilize the variances (Begg & Mazumdar, 1994). The standardized effect size of study i (T_i^*) is defined as:

$$T_i^* = \frac{T_i - \bar{T}_\bullet}{\sqrt{\hat{v}_i^*}}$$

where $\bar{T}_\bullet = \frac{\sum_{i=1}^k T_i}{\sum_{i=1}^k \frac{1}{SE_i^2}}$, T_i is the observed effect size of study i , SE_i the standard error of the observed effect size, and $\hat{v}_i^* = SE_i^2 - \left(\sum_{i=1}^k \frac{1}{SE_i^2}\right)^{-1}$.

In Egger’s regression (Sterne & Egger, 2005), effect sizes (weighted by their inverse variances) are regressed on their standard errors as follows:

$$\hat{T}_i = b_0 + b_1 \times SE_i \text{ weighted by } \frac{1}{SE_i^2}$$

The regression slope b_1 indicates bias and is expected to be zero if bias is absent. Additionally, the intercept b_0 has been suggested as an estimate of the combined effect size adjusted for publication bias (Moreno et al., 2009). The rationale for this is that the intercept gives the predicted effect size for a hypothetical study with a standard error of zero (i.e. infinitely large sample size). If there is no bias, the intercept is equal to the combined effect size of the included studies. With the correlational effect size r , Egger’s regression may yield incorrect results, as the estimated standard error of r depends on the observed effect size (Macaskill, Walter, & Irwig, 2001; Sterne, Becker, & Egger, 2005). For this reason, correlations are transformed into Fisher-Z values throughout this paper.

Another method that not only indicates the presence of bias but also yields an adjusted estimate is the iterative trim-and-fill procedure (Duval & Tweedie, 2000a, 2000b). Trim-and-fill estimates and adjusts for the number of missing effects. In a first step, this method excludes “asymmetric” studies on the right side of the funnel plot for which no counterparts are present on the opposite side. A new pooled estimate is then computed from this reduced data set and the number of missing studies

is re-estimated. When no additional missing studies can be found, all trimmed effect sizes are reinstated. Additionally, their symmetric counterparts are imputed for the missing effects. The resulting, more symmetrical plot is then used to compute the adjusted effect estimate and its variance.

To determine the number of trimmed studies, two different estimators (R_0 and L_0) can be used that are both based on signed ranks of the absolute differences between the effect sizes and the combined effect. In a symmetric funnel plot, the most extreme deviations from the combined effect size will have similar ranks on both sides of the plot. If this is not the case, the size of the estimator R_0 will indicate asymmetry. R_0 depends on the rightmost run of ranks associated with effect sizes located above the pooled estimate. This implicates that a single outlier effect on the left hand side will cause R_0 to be zero (Duval & Tweedie, 2000a; Duval, 2005). In general, R_0 will not properly assess asymmetry when missing studies are accompanied by more extreme effects on the left hand side—a situation that appears as a gap in the funnel plot (and that is present in our reanalyses of the data set of Dato-on and Dahlstrom (2003), see Figure 5, as well as the total data set of Balliet and colleagues (2009), see Figure 2). The estimator L_0 stems from the assumption that, in the absence of bias, the sums of the ranks for the effect sizes on both sides of the pooled estimate will be similar. L_0 depends on the sum of the ranks on the right side of the funnel plot (the Wilcoxon statistic for the given set of data) and indicates asymmetry when this sum is larger than the expected value. Thus, L_0 is more robust against outliers and may detect a gap in the funnel plot. However, it will not necessarily indicate asymmetry if several of the most extreme effects are located on the right side of the plot. In general, both estimators can yield markedly different results, as they assess different characteristics of the distribution of effect sizes. However, if the funnel plot is symmetric, both estimators should indicate that no studies are missing. Therefore, following a recommendation by Duval (2005), we use both estimators in all analyses.

In simulation studies (Begg & Mazumdar, 1994; Duval & Tweedie, 2000a; Macaskill, Walter, & Irwig, 2001; Sterne, Gavaghan, & Egger, 2000), all of the above mentioned methods have been shown to achieve only limited power, especially when the number of studies is low and only a moderate publication bias is present (i.e. only a small proportion of studies are missing). Therefore, the use of a more liberal significance level (i.e. $\alpha = .10$) has been suggested (e.g., Egger, Smith, Schneider, & Minder, 1997). We follow this suggestion in this paper.

In addition to the statistical methods for the assessment of funnel plot asymmetry, we apply an exploratory procedure that provides a formal evaluation of the num-

ber of significant and non-significant studies in a meta-analysis (Ioannidis & Trikalinos, 2007). This procedure tests whether the observed number of significant findings differs from the number expected in the absence of bias. The expected number of “positive” findings results from the power of the primary studies. Power is calculated based on a standard Wald Z-Test (which also provides the contours indicating the different significance levels in the funnel plots displayed in this paper) under the assumption of a fixed α -level. The difference between the observed and expected number of “positive” findings can be tested for significance using either a χ^2 or binomial distribution. A significant result indicates an excess of ‘positive’ findings, and thus a lack of non-significant findings, among the primary studies. Due to power considerations, Ioannidis & Trikalinos (2007) recommend using a significance level of $\alpha=.10$.

In its simplest form, this procedure uses the combined effect size in the meta-analysis to calculate the power of the primary studies. However, in the presence of bias, the combined effect size is certain to be an overestimate of the underlying true effect size. Therefore, the power of the primary studies may be overestimated, as well, and the expected number of “positive” findings thus inflated. Therefore, Ioannidis and Trikalinos (2007) suggest using reduced estimates of the underlying effect in addition to the combined effect size for exploratory purposes. More specifically, they interpret significant test results as an indication of publication bias if the effect estimate used in the power calculation lies within the 95% confidence interval of the original combined effect size.

The analyses reported in this article were performed with Stata 11. While Stata itself does not include statistical packages for the analysis of publication bias, well-functioning macros are available on the Internet. The command used for generating the funnel plots is *confunnel*. Trim-and-fill analyses were performed with the command *metatrim*. Finally, Egger’s regression and Begg’s rank correlation were computed with the command *metabias*. A Stata macro for assessing a possible lack of non-significant studies in a meta-analysis is provided by Ioannidis (www.dhe.med.uoi.gr). A useful, but slightly outdated description of various computer programs to address publication bias is provided by Borenstein (2005). This description is also available online: (<http://www.metaanalysis.com/downloads/PBSoftware.pdf>).