Psychological Methods

1082-989X/14/$12.00 http://dx.doi.org/10.1037/met0000025

Meta-Analysis Using Effect Size Distributions of Only Statistically Significant Studies

Marcel A. L. M. van Assen, Robbie C. M. van Aert, and Jelte M. Wicherts
Tilburg University

Publication bias threatens the validity of meta-analytic results and leads to overestimation of the effect size in traditional meta-analysis. This particularly applies to meta-analyses that feature small studies, which are ubiquitous in psychology. Here we develop a new method for meta-analysis that deals with publication bias. This method, *p*-uniform, enables (a) testing of publication bias, (b) effect size estimation, and (c) testing of the null-hypothesis of no effect. No current method for meta-analysis possesses all 3 qualities. Application of *p*-uniform is straightforward because no additional data on missing studies are needed and no sophisticated assumptions or choices need to be made before applying it. Simulations show that *p*-uniform generally outperforms the trim-and-fill method and the test of excess significance (TES; Ioannidis & Trikalinos, 2007b) if publication bias exists and population effect size is homogenous or heterogeneity is slight. For illustration, *p*-uniform and other publication bias analyses are applied to the meta-analysis of McCall and Carriger (1993) examining the association between infants' habituation to a stimulus and their later cognitive ability (IQ). We conclude that *p*-uniform is a valuable technique for examining publication bias and estimating population effects in fixed-effect meta-analyses, and as sensitivity analysis to draw inferences about publication bias.

Keywords: meta-analysis, publication bias, the trim-and-fill method, test of excess significance, sensitivity analysis

Supplemental materials: http://dx.doi.org/10.1037/met0000025.supp

When more studies are conducted on a particular topic the need to synthesize the results of these studies grows. Meta-analysis has become a standard method to synthesize results; it is the statistical synthesis of the data from separate but similar, that is, comparable studies, leading to a quantitative summary of the pooled results (Last, 2001). In meta-analysis, one effect size measure (e.g., Cohen's *d*) is commonly extracted from each study together with study characteristics. These data are used to estimate a common underlying effect, and sometimes the effect and its heterogeneity are modeled as a function of the studies' characteristics. Applications of meta-analysis are numerous and their number continuous to grow. For instance, according to a search in PsycINFO (using the string AB "meta-analysis"), the number of peer-reviewed articles concerning meta-analysis went up from 67 in 1985 (0.2% of the total number of articles) to 1,265 in 2012 (0.9% of the articles; cf. Kisamore & Brannick, 2008). The number of citations of meta-analyses grows as well (Aytug, Rothstein, Zhou, & Kern, 2012). These trends suggest that meta-analysis is or is becoming an influential methodological tool in psychology and related fields.[1]

One of the greatest threats to the validity of meta-analytic results is publication bias (Banks, Kepes, & Banks, 2012; Rothstein, Sutton, & Borenstein, 2005). We narrowly define publication bias here as "the selective publication of studies with a statistically significant outcome;" that is, the overrepresentation in the literature of studies with a significant outcome compared to studies with so-called null results. The evidence of publication bias is overwhelming (e.g., van Assen, van Aert, Nuijten, & Wicherts, 2014). For instance, Fritz, Scherndl, and Küberger (2013) examined 1,000 randomly drawn psychological studies in 2007 and observed three times as many outcomes just reaching significance than outcomes just failing significance. Furthermore, in psychology about 95% of published articles contain statistically significant outcomes, and this percentage has been increasing over the years (Fanelli, 2012). Neither the high percentage nor its increase can be explained by the studies' statistical power since power is generally low (Ellis, 2010) and there is no evidence that it has grown over the years (Fanelli, 2012). Explanations of publication bias include researchers' reluctance to submit studies with nonsignificant results (Cooper, DeNeve, & Charlton, 1997; Coursol & Wagner, 1986), and

Marcel A. L. M. van Assen, Robbie C. M. van Aert, and Jelte M. Wicherts, Department of Methodology and Statistics, Tilburg University.

The preparation of this article was supported by Grants 406-13-050 and 016-125-385 from the Netherlands Organization for Scientific Research (NWO). We thank Perke Jacobs and Gregory Francis for their valuable comments on a draft of this article.

Correspondence concerning this article should be addressed to Marcel A. L. M. van Assen, Department of Methodology and Statistics, Tilburg University, 5000 LE Tilburg, The Netherlands. E-mail: m.a.l.m.vanassen@tilburguniversity.edu

[1] Aguinis et al. (2011) conclude that meta-analysis is one of the most influential methodological tools in management and related fields after observing that meta-analyses were cited three times as much as other empirical articles from 1963 to 2007 in the *Academy of Management Journal*, one of the most influential management journals.

lower appraisal of these studies by reviewers (Coursol & Wagner, 1986; Mahoney, 1977) and editors (Coursol & Wagner, 1986).

We continue our introduction on publication bias by first briefly considering three harmful consequences of publication bias. Then we relate how often publication bias is addressed in meta-analytic studies. Thereafter, we describe different goals and problems of current publication bias methods, and end with the goals and an overview of our study.

Three harmful consequences of publication bias are that researchers may exploit degrees of freedom (*df*) in the analysis of data (Simmons, Nelson, & Simonsohn, 2011), uncertainty of the existence of a true effect underlying a published statistically significant effect, and more generally, overestimation of the population effect (e.g., Asendorpf et al., 2013). Researcher *df*, or researchers' behavior directed at obtaining statistically significant results (Simonsohn, Nelson, & Simmons, 2013), which is also known as *p*-hacking or questionable research practices in the context of null hypothesis significance testing (e.g., O'Boyle, Banks, & Gonzalez-Mulé, 2014), results in a higher frequency of studies with false positives (Simmons et al., 2011) and inflates genuine effects (Bakker et al., 2012). Additionally, even in the absence of researcher *df*, systematic investigations demonstrate that publication bias leads to overestimation of effects, which can be dramatic if sample sizes are small (Bakker, van Dijk, & Wicherts, 2012; Francis, 2012; Gerber, Green, & Nickerson, 2001; Kraemer, Gardner, Brooks, & Yesavage, 1998). Consider extreme publication bias, that is, only statistically significant effects are published, and a population effect that is of medium or small size. A study's published effect size is then hardly informative of the underlying population effect and merely reflects sample size (Francis, 2012; Gerber et al., 2001; Kraemer et al., 1998). Moreover, a replication of a small study will generally obtain a smaller effect than the original study. For example, Gerber, Green, and Nickerson (2001, p. 388) show that in two-group studies with a total sample size of 50, the probability is about .95 that the observed effect in the replication study is smaller than in the original study. This property may at least partly explain why many replication studies fail to confirm results of original studies (Begley & Ellis, 2012; Prinz, Schlange, & Asadullah, 2011; Sarewitz, 2012). Obviously, if individual published studies obtain biased effect size estimates, meta-analyses mainly using these individual studies will yield biased estimates as well, and may falsely give the impression of a consistent research finding (Francis, 2012).

Because of the harmful consequences of publication bias it will not come as a surprise that meta-analysis experts note that publication bias analyses should be included in meta-analytic studies (e.g., Aytug et al., 2012; Banks, Kepes, & McDaniel, 2012; Field & Gillett, 2010; Sterne, Gavaghan, & Egger, 2000; Sutton, 2006). However, publication bias is unfortunately often not adequately addressed in meta-analytic studies. For example, reviews showed that publication bias was assessed in less than 10% of meta-analytic studies in industrial organization psychology studies (Sutton, 2006), less than 10% in management sciences (Aguinis, Dalton, Bosco, Pierce, & Dalton, 2011), 18% in organizational sciences (Aytug et al., 2012), 56% in education research (Banks, Kepes, & Banks, 2012), 31% in management and industrial/organizational psychology (Banks, Kepes, & McDaniel, 2012), 70% in journals published by the American Psychological Association and the Association for Psychological Science (Ferguson & Brannick, 2012), and 33% in judgment and decision-making research (Renkewitz, Fuchs, & Fiedler, 2011). To conclude, the failure to address publication bias is omnipresent, although there is considerable variation across disciplines.

Many tests of publication bias have been developed over the years. Most of these tests address the question whether any publication bias exists. A problem of latter tests lies in their limited power to detect publication bias, particularly if the number of studies in the meta-analysis is low (Borenstein, Hedges, Higgins, & Rothstein, 2009; Sterne & Egger, 2006). Because of limited power, one may falsely conclude that no publication bias exists in a meta-analysis, while the population effect size is still overestimated. Hence, rather than tests of publication bias, more interesting questions would be how much bias there is, and to what degree it affects the conclusions drawn from meta-analyses (Borenstein et al., 2009, p. 284). Preferably, publication bias methods should yield an accurate estimate of the population effect size after taking publication bias into account. Only a few methods analyzing publication bias generate such estimates, but the general consensus is that these methods should be considered as sensitivity analyses rather than yielding accurate estimates (Duval, 2005; Duval & Tweedie, 2000b). In the present article we develop a new fixed-effect meta-analysis method that should, unlike existing methods, yield an accurate estimate of the population effect size, even when publication bias is extreme. More specifically, the proposed method allows for (a) testing of publication bias, (b) estimating effect size, and (c) testing of the null-hypothesis of no effect. No current meta-analysis method possesses all three qualities.

We continue with an overview of methods analyzing publication bias. The overview is short for two reasons. First, other sources already present similar overviews (e.g., Banks, Kepes, & Banks, 2012; Kepes, Banks, & Oh, 2012; Rothstein et al., 2005). And second, we examine the performance of methods in a challenging meta-analytic context in which only two of these methods, the trim-and-fill method (Duval & Tweedie, 2000a, 2000b) and the test for excess significance (TES; Ioannidis & Trikalinos, 2007b), can be applied. Next, we explain our own method. Subsequently, we present the results of a simulation study to examine the performance of the new method to test publication bias, estimate population effect size, and test the null-hypothesis of no effect. We compare the performance of the new method with the performance of traditional fixed-effect meta-analyses, the trim-and-fill method, and TES, and apply all methods to a meta-analysis on the relation between infant habituation performance and later IQ (McCall & Carriger, 1993).

Methods for Assessing Publication Bias

We briefly discuss the following methods for assessing publication bias, along with their most important properties: failsafe *N* (Rosenthal, 1979), funnel plot (Light & Pillemer, 1984), Begg and Mazumdar's (1994) rank correlation test, Egger's test (Egger, Davey Smith, Schneider, & Minder, 1997), the trim-and-fill method (Duval & Tweedie, 2000b), the TES, and selection models (Hedges and Vevea, 2005). The oldest and most popular (e.g., Banks, Kepes, & McDaniel, 2012; Ferguson & Brannick, 2012) method is failsafe *N* (Rosenthal, 1979), which provides the number of studies needed to render a statistically significant effect of a meta-analysis insignificant. Because of its problematic assump-

tions and typically overly optimistic results, experts recommend abandoning failsafe *N* (e.g., Becker, 2005).

The funnel plot (Light & Pillemer, 1984) typically displays studies' effect sizes on the *x*-axis and their standard error or their precision (the inverse of a study's standard error) on the *y*-axis (Sterne & Egger, 2001). Figure 1 shows the (contour-enhanced) funnel plot of the meta-analysis of McCall and Carriger (1993; cf. Bakker et al., 2012). Funnel plot asymmetry, with a lower frequency of studies in the lower center of the plot corresponding to studies with a small and statistically insignificant effect size and a small sample size, is interpreted as an indication of publication bias. Hence, the funnel plot in Figure 1 indicates that publication bias may have affected the results. However, funnel plots can also be asymmetric for other reasons (Sterne, Becker, & Egger, 2005). To overcome this interpretation problem, Peters, Sutton, Jones, Abrams, and Rushton (2008) developed the contour-enhanced funnel plot, which explicitly links the presence of studies in the funnel plot to their statistical (in)significance. The contour-enhanced funnel plot in Figure 1 suggests publication bias, because the asymmetry of the plot is linked to the statistical significance of the studies. Nonetheless, funnel plot methods are subjective, and many errors are made when identifying publication bias using the funnel plot (Terrin, Schmid, Lau, & Olkin, 2003). Even experienced meta-analysts only correctly identified 52.5% of the cases in which a funnel plot was or was not affected by publication bias (Terrin, Schmid, & Lau, 2005). Two methods, Begg and Mazumdar's (1994) rank correlation and Egger's regression method (Egger et al., 1997; Sterne & Egger, 2006), formally test funnel plot asymmetry. Both methods test the association between studies' effect size and corresponding standard error, where a significant (typically positive) association signals publication bias. Because these methods have low statistical power (e.g., Borenstein et al., 2009, p. 291; Sterne & Egger, 2006), both tests are usually applied using a significance level of .10. Due to their low power, their application is only recommended for meta-analyses based on at least 10 (Banks, Kepes, & Banks, 2012; Sterne & Egger, 2006) or even 15 effect sizes (Kepes et al., 2012). Rothstein and Bushman (2012) also argued that the results of both tests are not meaningful if between-study heterogeneity in effect size is substantial. Finally, a clear limitation of both methods is that they can only be applied if there is reasonable variation in studies' sample size, with preferably at least a few samples with medium or large sample sizes (Borenstein et al., 2009, p. 284).

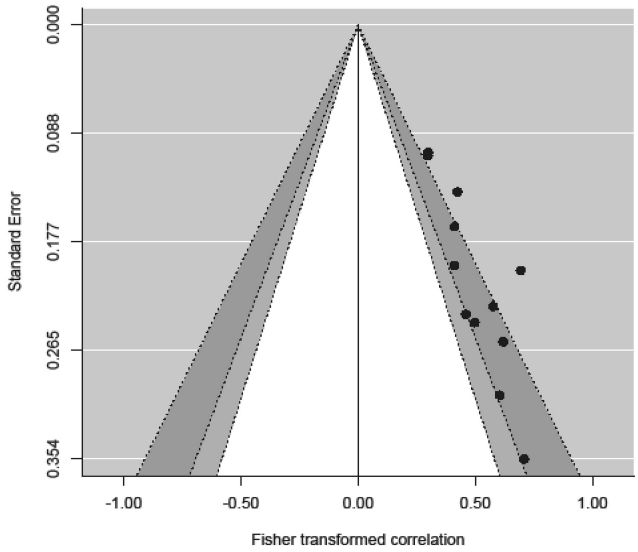

Figure 1. Contour-enhanced funnel plot of the meta-analysis of McCall and Carriger (1993). Areas represent studies with two-tailed *p*-values larger than .10 (white), smaller than .05 (light gray), smaller than .01 (dark gray), and smaller than .001 (light gray outside large triangle).

The trim-and-fill method developed by Duval and Tweedie (2000a, 2000b) is another method for assessing publication bias on the basis of the funnel plot. It entails an iterative procedure that fills in missing studies that are needed to restore funnel plot symmetry, and provides an estimate of both the number of such missing studies and the effect size. Duval and Tweedie (2000a, 2000b) developed three estimators (R_0, L_0, Q_0) for the number of missing studies. Estimators R_0 and L_0 perform better than Q_0, and L_0 is more robust than R_0 against the occurrence of a few aberrant studies (Duval & Tweedie, 2000a, 2000b). L_0 is also used in most applications of the trim-and-fill method. Stated advantages of the trim-and-fill method are that it is relatively simple and provides an estimate of the effect size corrected for publication bias. However, the consensus is that the method should not be regarded as a way of yielding a more "valid" estimate of the overall effect size, but rather as a sensitivity analysis (Duval, 2005; Duval & Tweedie, 2000b; Viechtbauer, 2010). Results on the performance of the trim-and-fill method are mixed; some suggest the method is quite powerful and yields close to unbiased effect size estimates (Duval & Tweedie, 2000b), whereas others suggest it has low power to test the null-hypothesis of no effect (Ferguson & Brannick, 2012). Agreement exists, however, that the method should not be used when population effect sizes are heterogeneous, because then it is likely to add nonexisting studies (Rothstein & Bushman, 2012; Terrin et al., 2003).

Ioannidis and Trikalinos (2007b) developed a test for publication bias based on a comparison between the observed (O) and expected (E) number of statistically significant studies in a meta-analysis. The expected number E is calculated as the sum of the studies' observed power, based on the effect size as estimated in the meta-analysis: $E = \sum_{i=1}^{K} (1 - \beta_i)$. The test for excess significance (TES) for publication bias is the common χ^2-test, with degrees of freedom equal to 1:

$$\frac{(O-E)^2}{E} + \frac{(O-E)^2}{K-E}.$$

If the p value of the test statistic is significant at .10, the test is interpreted as a signal of publication bias for a given meta-analysis. However, Francis (2012, 2013) showed that a statistically significant test outcome may also be the result of researcher *df* such as data peeking. Any process (publication bias or researcher *df*) leading to an abundance of statistically significant studies may be picked up by the TES. The TES has low power when only a limited number of studies is included in a meta-analysis (Francis, 2012, 2013; Ioannidis & Trikalinos, 2007a), and has particularly low power when population effects are heterogenous (Francis, 2013). Ioannidis and Trikalinos (2007b) also recommend not using the test if between-study heterogeneity exists, but to first create homogenous subgroups of effect sizes before applying the test.

Finally, the TES neither provides an answer to the question whether the population effect differs from zero, nor does it provide a (corrected) estimate of the effect.

In selection models, the probability of observing an effect depends on its value. Several versions of selection models exist (Hedges & Vevea, 2005; Terrin et al., 2003). Some versions estimate both the meta-analytic effect and the so-called weight function representing the probabilities of observing an effect as a function of their value. These versions are quite technical and have typically been effective only with meta-analyses containing relatively large numbers of studies (more than 100; Field & Gillett, 2010). The requirement of at least 100 studies severely limits the usefulness of selection models to estimate effect size in actual meta-analyses. However, other versions have been developed that do not estimate the weight function but allow the user to specify the weight function in advance (Vevea & Woods, 2005). Hedges and Vevea (2005) argued that these a priori specified selection models provide a means for sensitivity analyses. Terrin et al. (2003) examined the performance of a selection model with a step weight function with one cutpoint at $p = .05$ in meta-analyses of either 10 or 25 studies. Estimation failed to converge most of the time when the population effect size was homogenous or when it was heterogeneous with 10 studies. Convergence was better (58%–98%) for heterogeneous effect sizes with 25 studies, and the selection model outperformed the trim-and-fill method. When studies' population effects are heterogeneous, Hedges and Vevea (2005) recommend selection models as sensitivity analysis, because more simple methods such as the trim-and-fill method and the TES provide misleading results in that case. However, Borenstein, Hedges, Higgins, and Rothstein (2009, p. 281) concluded that "selection models have rarely been used in actual research because they are difficult to implement and also because they require the user to make some relatively sophisticated assumptions and choices." Although it should be noted that R routines are available (e.g., Vevea & Woods, 2005), it is unlikely that selection models will be used routinely in meta-analysis (Hedges & Vevea, 2005).

The *p*-Uniform Method

p-uniform is a new method for conducting meta-analyses that allows for testing publication bias and estimating a fixed effect size under publication bias, or that can be used as a sensitivity analysis to address and examine publication bias in meta-analyses. The method only considers studies with a statistically significant effect, and hence discards those with an insignificant effect. Hedges (1984) also suggested a method to estimate effect size using only statistically significant studies, based on maximum likelihood. And currently Simonsohn, Nelson, and Simmons (2014) are also working on a method to estimate effect size only using statistically significant studies.

p-uniform makes two assumptions. First, like in other methods, the population effect size is taken to be fixed rather than heterogeneous. Although the assumption of a fixed effect will not be tenable for all psychological meta-analyses, Klein et al.'s (2014) "Many Labs Replication Project" provides evidence that it holds for lab studies on many psychological phenomena; 36 scientific groups in 12 different countries directly replicated 16 effects, with no evidence of a heterogeneous effect size in eight of 16 effects (50%). Heterogeneity may be more common in observational studies. Second, *p*-uniform assumes that all studies with statistically significant findings are equally likely to be published and included in the meta-analysis The second assumption is formalized as $f(p_i) = C$ for $p_i \leq \alpha$, indicating that there is no association between an effect size's significant *p* value and the probability that the study containing this *p* value will get published. *p*-uniform does not make assumptions about the magnitude of the publication probability (the value of *C*), or the probability that statistically insignificant studies get published ($f(p_i)$ for $p_i > \alpha$). An example of a violation of the second assumption is if highly significant findings, for example, in combination with a large sample size, have a higher probability of getting published and being included in the meta-analysis. A violation will probably have minor consequences on the performance of *p*-uniform, because most statistically significant findings will get published. In principle, *p*-uniform allows α to be specific for a study or researcher, which is relevant if studies or researchers vary in their chosen significance-level (e.g., some use .01 whereas others use .05) or in the direction of the test (one-tailed and two-tailed tests correspond to one-tailed significance-levels of α and $\alpha/2$, respectively).

The basic idea of *p*-uniform is that the distribution of *p* values conditional on the population effect size is uniform. This assumption is equivalent to the assumption underlying standard null hypothesis testing, with the important distinction that we now focus on the (conditional) p^{μ} value distribution, which is the *p* value distribution under the alternative hypothesis that the population effect size equals μ. *p*-uniform's effect size estimate will equal the effect size μ yielding a p^{μ} value distribution that is fitted best by a uniform distribution. *p*-uniform's test of the hypothesis of no effect is based on the deviation of the p^{O} value distribution from the uniform distribution, where the p^{O} value distribution corresponds to the distribution of original *p* values (i.e., corresponding to *p* values of the test of no effect, or $\mu = 0$). *p*-uniform's test of publication bias is based on the deviation of the $p^{\hat{\mu}}$ value distribution from the uniform distribution, where $\hat{\mu}$ equals the effect size estimate of traditional fixed-effect meta-analysis.

We will explain effect size estimation and the two tests using an artificial example. The example is based on testing the hypothesis of no effect ($\mu = 0$) against the alternative of a positive effect ($\mu > 0$) with $\alpha = .05$. However, *p*-uniform can estimate and test any effect size measure. In the example, 80 studies with sample size 25 are generated using a fixed-effect model with $\mu = .33$ and $\sigma = 1$, where all statistically significant studies and 25% of insignificant studies are published. If each study tests the hypothesis of no effect ($\mu = 0$) against the alternative of a positive effect ($\mu > 0$) with $\alpha = .05$, then each study has a power of .5. Figure 2a shows the distribution of transformed *p* values (p^{O} value distribution, or the distribution of *p* values $\times 1/\alpha$) of the $K = 40$ statistically significant studies of one simulation of the traditional example. Traditional fixed-effect meta-analysis carried out on all 50 published studies using the Metafor package (Viechtbauer, 2010) yields a biased effect size estimate of 0.43 ($SE = .063$, $p < .001$).

Test of $\mu = 0$. If $\mu = 0$ then the p^{O} value distribution in Figure 2a should be close to the uniform distribution. Hence, *p*-uniform tests the hypothesis $\mu = 0$ by testing whether the observed p^{O} value distribution deviates from the uniform distribution. Fisher's (1932) method has been used before to test devia-

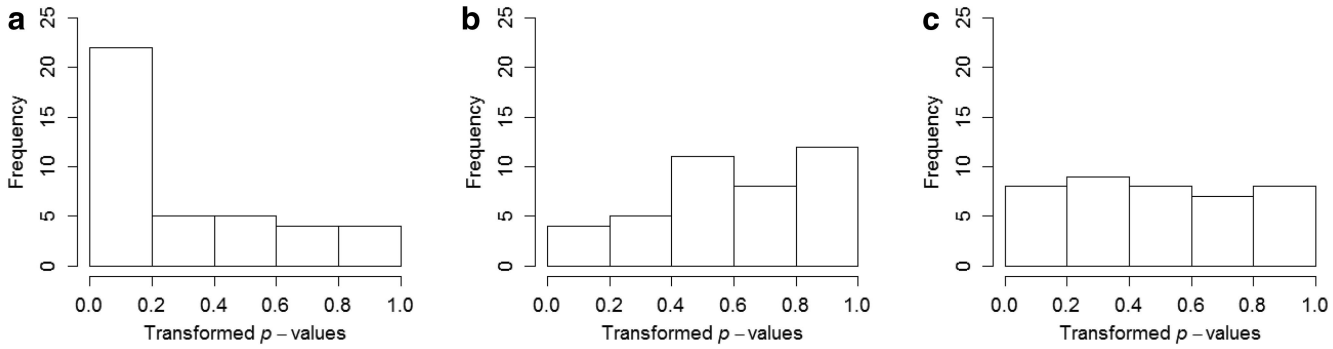

Figure 2. *p*-value distribution for (a) $\mu = 0$, (b) $\mu = \hat{\mu}$, and (c) $\mu = \hat{\mu}^*$ as a function of the transformed significant *p*-values on the *x*-axis and its frequency on the *y*-axis.

tions from the uniform distribution. Notably, independently of us, Simonsohn, Nelson, and Simmons (2013) have applied exactly the same test of $\mu = 0$ as we did. The first step of Fisher's method is to convert each *p* value into numbers in the interval from 0 to 1 by computing the conditional probability of the *p* value given its significance ($\alpha = .05$). The probability that a *p* value is statistically significant is .05 if $\mu = 0$, hence all *p* values are multiplied by 20 in the first step. Applying Fisher's (1932) method, if $\mu = 0$ then the test statistic $L^0 = -\sum_{i=1}^{K} \ln(20p_i)$ is gamma distributed with *K* and 1 degrees of freedom, here denoted by $\Gamma(K, 1)$. If the studies' *p* values are generally small, as in Figure 2a, L^O will be high. *p*-uniform rejects $\mu = 0$ whenever the value of L^O is larger than the 95th percentile of the gamma distribution, denoted by $\Gamma_{.95}(K, 1)$. In the example with $K = 40$, $\Gamma_{.95}(40, 1) = 50.94$. The null-hypothesis is rejected since $L^O = 82.26$ ($p < .001$); the population effect is larger than zero.

Test of publication bias. The test for publication bias by *p*-uniform amounts to a one-tailed test of the null-hypothesis $\mu = \hat{\mu}$, that is, whether the population effect size equals the effect size estimate of a traditional fixed-effect meta-analysis. The basic idea is that the null-hypothesis is rejected if the $p^{\hat{\mu}}$ value distribution deviates from the uniform distribution. The $p^{\hat{\mu}}$ value distribution is a conditional distribution. More generally, we will assume a test of $\mu = \mu^*$ for defining this conditional distribution. The definition uses the sampling distributions of effect size $M_i^{\mu^*}$ of all studies *i*, assuming $\mu_i = \mu^*$. The conditional *p* value distribution $p_i^{\mu^*}$ is then defined as:

$$p_i^{\mu^*} = \frac{p\left(M_i^{\mu^*} \geq \hat{\mu}_i\right)}{p\left(M_i^{\mu^*} \geq M_i^{CV}\right)}.$$

M_i^{CV} denotes the critical value of $M_i^{\mu^*}$ for which $p(M_i^0 \geq M_i^{CV}) = \alpha$, and $\hat{\mu}_i$ denotes the estimated effect size in study *i*. The probabilities in the numerator and denominator are calculated under the assumption that $M_i^{\mu^*}$ is normally distributed. In words, $p_i^{\mu^*}$ represents the probability of observing effect $\hat{\mu}_i$ or larger, conditional on both a population effect μ^* and a significant *p* value (when tested against the null hypothesis of no effect). It is important to note that in each study *i* can be based on a different sample size N_i, and that $p_i^{\mu^*}$'s dependence on μ^* is stronger for larger N_i.

Figure 2b depicts the distribution of $p_i^{\hat{\mu}}$, that is, the $p^{\hat{\mu}}$ value distribution. The distribution is not uniform but skewed to the left with many high *p* values, suggesting publication bias. The hypothesis of no publication bias is rejected if $L^{\hat{\mu}} < \Gamma_{.05}(40, 1) = 30.2$, with $L^{\hat{\mu}} = -\sum_{i=1}^{K} \ln(p_i^{\hat{\mu}})$. Applying Fisher's test to the distribution of $p_i^{\hat{\mu}}$ yields $L^{\hat{\mu}} = 28.11$ ($p = .020$), indeed suggesting publication bias; the population effect is smaller than its value estimated by the traditional fixed-effect meta-analysis.

Interval and point estimation of μ. The $100(1 - \alpha)\%$ confidence interval $\hat{\mu}_L^* \leq \mu \leq \hat{\mu}_U^*$ is obtained by $L^{\hat{\mu}_L^*} = \Gamma_{1-0.5\alpha}(K, 1)$ and $L^{\hat{\mu}_U^*} = \Gamma_{0.5\alpha}(K, 1)$. That is, each border of the interval is a value of μ for which the null-hypothesis is only just accepted in a two-tailed test at significance level α. The probability that the null-hypothesis is rejected that effect size equals μ is exactly .05, because (only) for μ is the *p* value distribution exactly uniform. Consequently, this proofs that the interval estimate of *p*-uniform is unbiased: 95% of all confidence intervals (CI) contain μ, or the coverage probability of *p*-uniform is exactly .95.

The borders of the confidence interval are easily obtained, because $L^{\hat{\mu}^*}$ decreases monotonically in $\hat{\mu}^*$.[2] The confidence interval in the example is $0.21 \leq \mu \leq 0.43$. *p*-uniform's point estimate $\hat{\mu}^*$ equals the effect size yielding a $p^{\hat{\mu}^*}$ value distribution that is fitted best by a uniform distribution. The point estimate is defined as the value of $\hat{\mu}^*$ for which L^{μ} equals *K*, which is the expected value of $\Gamma(K, 1)$. In the example, $\hat{\mu}^* = .32$. Figure 2c depicts the distribution of $p_i^{0.32}$. Note that 0.32 closely corresponds to $\mu = .33$ used to generate the studies in this hypothetical example.

Alternative estimators in *p*-uniform: 1 − *p*. The basic idea of *p*-uniform is that the *p* value distribution conditional on the population effect size is uniform. However, the distribution of some transformation of *p* values are then also uniform. For instance, if *p* is uniformly distributed, then so is $1 - p$. Consequently, we can also (a) test $\mu = 0$, (b) test publication bias, and (c) estimate $\hat{\mu}^*$, using $1 - p_i^{\mu^*}$ rather than $p_i^{\mu^*}$. The two estimators are differently sensitive to outliers, that is, studies with extreme

[2] *p*-uniform's point estimates and the bounds of its confidence interval are obtained by the R CRAN function uniroot. The input of uniroot is a function and an interval. It searches for a value in the interval for which the function equals zero. The functions were $L^{\hat{\mu}_L^*} - \Gamma_{1-0.5a}(K, 1)$, $L^{\hat{\mu}^*} - K$, $L^{\hat{\mu}_U^*} - \Gamma_{0.5a}(K, 1)$ for lower bound, point estimate, and upper bound of the confidence interval, respectively. The interval was $[-1, 1]$ in all cases.

effect size estimates, where the estimator based on $p_i^{\mu^*}$ is very much affected by outliers, whereas the other is not. A very large effect size will yield a small $p_i^{\mu^*}$, hence a large $-\ln(p_i^{\mu^*})$, resulting in a large positive effect of that effect size on estimate $\hat{\mu}^*$. However, one very large effect size hardly affects $\hat{\mu}^*$ whenever the estimator based on $1 - p_i^{\mu^*}$ is used, because then $-\ln(1 - p_i^{\mu^*})$ approaches 0. To conclude, we expect the estimator based on $1 - p_i^{\mu^*}$ to be more robust to outliers and a violation of the homogeneity assumption than the estimator based on $p_i^{\mu^*}$. Properties of both estimators are examined in this study.

Characteristics of *p*-uniform. *p*-uniform allows for testing the null-hypothesis of no effect, testing publication bias, and estimating point and interval effect sizes. Other methods do not meet these three goals simultaneously. The trim-and-fill method also estimates effect size after imputing some studies that may have been missing, but statistical properties (e.g., bias) of that trim-and-fill estimate remain unclear. Because *p*-uniform is derived from solid statistical theory, *p*-uniform yields unbiased interval estimation (i.e., coverage probability equal to $1 - \alpha$) if its assumptions are met.

One assumption of *p*-uniform is that no questionable research practices were used in the studies, or, as Simonshon et al. (2013) put it, "*p*-hacking" did not occur. *p*-hacking will typically result in *p* values just below .05 (Simonsohn et al., 2013). Because *p* values close to .05 provide evidence for a low or even negative population effect size in *p*-uniform, *p*-hacking will in general result in an *under*estimation of the population effect size whenever *p*-uniform is applied. We consider this conservatism to be a positive quality of *p*-uniform; it will give estimates on the safe side, rather than traditional meta-analysis methods that *over*estimate population effect size because of *p*-hacking.

Another important assumption of *p*-uniform is that the population effect size is fixed. Our simulation study of *p*-uniform includes a test on the robustness of *p*-uniform to a violation of the homogeneity assumption. We expected that both point and interval estimation of the effect size would no longer be accurate in the case of between-study heterogeneity, and that estimation would be more biased whenever estimation is based on $p_i^{\mu^*}$ rather than $1 - p_i^{\mu^*}$. Note, however, that the performance of other methods assessing publication bias is also negatively affected by between-study heterogeneity (Moreno et al., 2009; Peters et al., 2007; Terrin et al., 2003). Moreover, it is often possible to select homogeneous subsets of studies on the basis of methodological or substantive characteristics, and apply *p*-uniform to these subsets. This is also the recommended approach for the other methods for assessing publication bias whenever there is heterogeneity (e.g., Ioannidis & Trikalinos, 2007b; Kepes et al., 2012).

A final assumption of *p*-uniform is that there is no association between the probability of statistically significant studies being in the meta-analysis and their *p* value. This is a weaker assumption than the (typically untenable) assumption underlying traditional meta-analysis, namely that all studies, statistically significant or not, have an equal chance to be included in the meta-analysis. Selection models either make a stronger assumption than *p*-uniform on this function for the whole range of *p* values, or estimate the probability of a study to be selected in the meta-analysis as a function of its *p* value. Estimation of particularly this function is problematic in selection models, requiring a very large number of studies (100 or more), and often leading to convergence problems (Terrin et al., 2003) and biased (Hedges & Vevea, 1996) or unrealistic functions (Hedges & Vevea, 2005).

A disadvantage of *p*-uniform seems that it discards all information from statistically insignificant studies. If there is no publication bias, using information from all studies will certainly yield a more precise estimate of population effect size. However, retrieval of unpublished studies is often hard and possibly biased (Ferguson & Brannick, 2012), for instance because such studies are typically not even documented properly (Cooper et al., 1997). Moreover, it is impossible (without study or trial registers) to be aware of how many unpublished studies there actually are. However, it is likely that the percentage of statistically insignificant studies is higher among the unpublished studies. To conclude, although meta-analysts often recommend researchers to search extensively for both published and unpublished studies when conducting traditional meta-analysis (e.g., Rothstein & Bushman, 2012), this search and its outcomes may introduce bias as well (Ferguson & Brannick, 2012). Most importantly, although omitting statistically insignificant studies may seem rather restrictive, the majority of published studies report statistically significant results, with a prevalence estimate of around 95% in psychology (e.g., Fanelli, 2010, 2012; Sterling, 1959; Sterling, Rosenbaum, & Weinkam, 1995). Hence, not many available studies need to be omitted by *p*-uniform anyway. Finally, statistically insignificant studies *must* be omitted in *p*-uniform; only by omission of insignificant studies will *p*-uniform yield accurate estimates.

Method

All methods for assessing publication bias work for any effect size measure (cf. Borenstein et al., 2009, Chapter 34). For illustrative purposes, we compare the methods in the most simple research situation. Effect sizes of studies were generated with a fixed population mean μ and standard deviation $\sigma = 1$ in all conditions, and a right-tailed test of the null-hypothesis H_0: $\mu = 0$ was conducted in each individual study with $\alpha = .05$. The performance of *p*-uniform and other techniques for assessing publication bias were examined by means of Monte-Carlo simulations. In these simulations, equal sample sizes of 25 were imposed for each study in the meta-analysis. A sample size of 25 resembles the median cell size of 24 in both between- and within-subjects designs in experimental psychology observed by Wetzels et al. (2011).

Due to using equal sample sizes, not all available techniques for assessing publication bias can be applied. Neither the rank-correlation test, nor Egger's test can deal with equal sample sizes (e.g., Ioannidis & Trikalinos, 2007a), and were therefore excluded from the simulation study. The fixed-effect model was applied because studies' effect sizes were generated from the same population with one fixed mean. The trim-and-fill method was imposed to impute only studies in the left-hand side of the funnel plot because studies were tested for being significantly larger than zero. Two-tailed tests ($\alpha = .05$) were conducted for testing the effect size estimates obtained by the fixed-effect model, the trim-and-fill method, and *p*-uniform. The TES and *p*-uniform's publication bias test were also conducted two-tailed with an alpha level of 0.05; a 0.05 significance level rather than the more common 0.10 was

selected to be consistent with the tests of effect size and its 95% CI.

For each condition, *p*-uniform was applied and compared with other existing methods for three purposes. First, we evaluated *p*-uniform's performance in estimating the population's effect size. *p*-uniform's effect size estimates, standard deviations of the effect size estimates, 95% CI, and coverage probabilities were compared with estimates obtained by the traditional fixed-effect model and the trim-and-fill method. We calculated the coverage probability as the proportion of runs with μ in the calculated 95% CI. Hence, an accurate method yields a coverage probability of .95 in all conditions. Second, in each replication we tested whether the population effect size is different from 0 (H_0: $\mu = 0$). For this test, Type-I error rates and statistical power were used to compare *p*-uniform, the fixed-effect model, and the trim-and-fill method. Third, we tested whether *p*-uniform can detect the presence of publication bias (H_0: $\mu = \hat{\mu}$). Type-I error rates and statistical power were also used to compare *p*-uniform's publication bias test with the TES.

Three parameters were varied in the main simulation study: the number of studies (K), the population effect size μ, and the proportion of statistically nonsignificant studies selected in the meta-analysis (p_p). Simultaneous with selecting values for K, levels for statistical power were chosen in such a way that the expected number of studies with an observed mean significantly larger than zero was eight in each condition. Recall that eight is a very small number of studies, because some publication assessment methods such as Begg and Mazumdar's (1994) rank correlation and Egger's regression method are only recommended when the number of effect sizes is at least 10 or 15. We particularly selected a small value of K to show that *p*-uniform may work well in meta-analyses based on a small number of studies that are common in the literature. The following values for K and statistical power ($1 - \beta$) were selected: $K = 160$ ($1 - \beta = \alpha = .05$); $K = 40$ ($1 - \beta = 0.2$); $K = 16$ ($1 - \beta = 0.5$); and $K = 10$ ($1 - \beta = 0.8$). Six different levels of publication bias were selected: $p_p = (0; 0.025; 0.05; 0.25; 0.5;$ and 1), where p_p denotes the proportion of statistically insignificant studies getting published. In case of extreme publication bias ($p_p = 0$), meta-analyses only consisted of on average eight published studies. The conditions $p_p = 0.025$ and $p_p = 0.05$ were chosen based on the probability of finding a statistically significant effect in the literature.[3] Proportions $p_p = 0.25$ and $p_p = 0.5$ were selected to reflect situations with less severe publication bias. A condition without publication bias ($p_p = 1$) was also included in order to compare the performance of *p*-uniform to the traditional fixed-effect model. This is the situation where the traditional fixed-effect model yields an unbiased estimate based on all studies. For each condition in the simulation study, 10,000 replications were conducted.

We also ran an additional simulation study to examine the robustness of *p*-uniform to violations of the homogeneity assumption. Four cells of the design of the main simulation study were selected ($K/\mu = (0; 0.33) \times p_p = (0; 0.25)$), and heterogeneity was manipulated using three levels ($\tau^2 = (0.013333; 0.04; 0.12)$ in each of these four cells. Parameter τ^2 represents the variance of true study means. The levels of τ^2 correspond to low ($I^2 = .25$), moderate ($I^2 = .50$), and high ($I^2 = .75$) heterogeneity (Borenstein et al., 2009, p. 119). The main dependent variables in the simulation were the point and interval estimates of traditional meta-analysis, the trim-and-fill method, and *p*-uniform.

To summarize, the main simulation study consisted of $K/\mu \times p_p = 4 \times 6 = 24$ conditions, whereas the additional simulation study had $K/\mu \times p_p \times \tau^2 = 2 \times 2 \times 3 = 12$ conditions. Simulations and *p*-uniform were programmed in R (R Core Team, 2012). The Metafor package (Viechtbauer, 2010) was used for conducting the trim-and-fill method and fixed-effect (main simulation) and random-effects (small simulation) meta-analyses. See the supplementary information for the R code of our simulations.

Results

Estimation of Effects When Effects Are Homogenous

Convergence rates for the effect size estimates with *p*-uniform were above 98.9% and 96.3% across conditions for the $p_i^{\mu^*}$ and $1 - p_i^{\mu^*}$ estimator, respectively.[4] Table 1 shows average effect size estimates, standard errors or standard deviations of the effect size estimates, confidence intervals, and coverage probabilities of the fixed-effect model, the trim-and-fill method, and *p*-uniform. The performance of *p*-uniform was only evaluated as a function of the population effect size μ because the method does not take statistically nonsignificant studies into account. Coverage probabilities of *p*-uniform were 95% in all conditions for both estimators (see last

[3] Assume that the probability that an effect truly exists ($P(\text{H}_1)$) is 0.5. Ioannidis (2005) used this value as starting point in his article and argued that this value may be lower in fields with less confirmatory research. Also assume that the statistical power accompanied with the applied statistical test is 0.5 (using $\alpha = 0.05$). Statistical power is often lower than the convention of 0.8. Bakker et al. (2012) even suggested that the typical power in psychological research is 0.35. These findings suggest that assuming a statistical power of 0.5 may even be liberal. The proportion of statistically significant studies in the literature ($P(\text{'H}_1\text{'} \mid \text{lit})$) can then be found after entering values for p in the following equation:

$$P(\text{'}H_1\text{'} \mid lit) = \frac{P(\text{'}H_1\text{'} \cap lit)}{P(\text{'}H_0\text{'} \cap lit) + P(\text{'}H_1\text{'} \cap lit)}$$

$$= \frac{(1-\beta)\cdot P(H_1) + \alpha \cdot P(H_0)}{p[\beta \cdot P(H_1) + (1-\alpha)\cdot P(H_0)] + (1-\beta)\cdot P(H_1) + \alpha \cdot P(H_0)}$$

where $P(\text{'H}_0\text{'})$ is the proportion of statistically non-significant findings in the literature. For instance, the proportion of statistically significant findings in the literature if $p = 0.025$ is:

$$\frac{0.5 \cdot 0.5 + 0.05 \cdot 0.5}{0.025[0.5 \cdot 0.5 + 0.95 \cdot 0.5] + 0.5 \cdot 0.5 + 0.05 \cdot 0.5} = 0.94.$$

If $p = 0.05$, the proportion of statistically significant studies in the literature is 0.88. These results are in line with research by Fanelli (2012) who showed that the proportion of studies reporting a positive result is approximately 85.9% in a variety of research fields, and in line with psychological research where 96%–97% of the studies report statistically significant results (Sterling, Rosenbaum, & Weinkam, 1995).

[4] Lack of convergence primarily occurred when there was no effect ($\mu = 0$). Averages for the lower and upper bound of *p*-uniform's confidence interval and effect size estimates were computed after exclusion of nonconverging replications. Coverage probabilities were based on all replications because lower and upper bounds of the confidence interval in case of nonconvergence were below -1 or above 1. As a result, if the estimate of one bound did not converge, the other bound's estimate could always be used to determine if μ was within the confidence interval.

Table 1
Average Effect Size Estimates and Corresponding Standard Errors/Standard Deviations, Confidence Intervals, and Coverage Probabilities for the Fixed-Effect Model, the Trim-and-Fill Method, and p-Uniform Based on Monte-Carlo Simulations of Homogenous Effects (10,000 Replications)

	μ (*K*)							
p_p	0 (160)		0.16 (40)		0.33 (16)		0.5 (10)	
0								
Fixed-effect model	0.412 (0.028) [0.267, 0.557]	CP: <.0001	0.440 (0.035) [0.295, 0.585]	CP: .009	0.489 (0.045) [0.347, 0.631]	CP: .359	0.569 (0.054) [0.429, 0.708]	CP: .901
Trim-and-fill	0.411 (0.028) [0.267, 0.556]	CP: <.0001	0.439 (0.035) [0.295, 0.583]	CP: .009	0.487 (0.044) [0.346, 0.629]	CP: .362	0.566 (0.054) [0.428, 0.705]	CP: .906
1/40								
Fixed-effect model	0.274 (0.071) [0.157, 0.392]	CP: .026	0.408 (0.050) [0.271, 0.546]	CP: .037	0.481 (0.047) [0.341, 0.621]	CP: .402	0.567 (0.055) [0.427, 0.706]	CP: .903
Trim-and-fill	0.247 (0.066) [0.135, 0.358]	CP: .035	0.391 (0.067) [0.259, 0.524]	CP: .080	0.477 (0.051) [0.339, 0.616]	CP: .415	0.564 (0.055) [0.426, 0.702]	CP: .908
1/20								
Fixed-effect model	0.202 (0.066) [0.101, 0.304]	CP: .074	0.382 (0.056) [0.251, 0.512]	CP: .074	0.474 (0.049) [0.336, 0.612]	CP: .443	0.564 (0.056) [0.426, 0.703]	CP: .907
Trim-and-fill	0.187 (0.062) [0.088, 0.286]	CP: .092	0.358 (0.071) [0.233, 0.482]	CP: .142	0.468 (0.057) [0.331, 0.604]	CP: .464	0.561 (0.056) [0.424, 0.699]	CP: .910
1/4								
Fixed-effect model	0.054 (0.035) [−0.004, 0.112]	CP: .542	0.266 (0.056) [0.166, 0.365]	CP: .434	0.426 (0.056) [0.300, 0.551]	CP: .696	0.549 (0.059) [0.414, 0.684]	CP: .928
Trim-and-fill	0.049 (0.037) [−0.009, 0.107]	CP: .577	0.250 (0.054) [0.153, 0.347]	CP: .533	0.412 (0.066) [0.289, 0.534]	CP: .726	0.543 (0.062) [0.410, 0.677]	CP: .926
1/2								
Fixed-effect model	0.020 (0.024) [−0.023, 0.063]	CP: .833	0.207 (0.044) [0.127, 0.288]	CP: .772	0.383 (0.056) [0.270, 0.497]	CP: .859	0.531 (0.061) [0.400, 0.662]	CP: .946
Trim-and-fill	0.011 (0.032) [−0.031, 0.053]	CP: .791	0.196 (0.046) [0.117, 0.275]	CP: .821	0.371 (0.061) [0.259, 0.482]	CP: .875	0.523 (0.065) [0.394, 0.652]	CP: .937
1								
Fixed-effect model	0.000 (0.016) [−0.031, .031]	CP: .949	0.161 (0.032) [0.099, 0.223]	CP: .948	0.330 (0.050) [0.232, 0.428]	CP: .952	0.500 (0.063) [0.376, 0.624]	CP: .951
Trim-and-fill	−0.020 (0.030) [−0.050, 0.011]	CP: .634	0.149 (0.039) [0.088, 0.209]	CP: .869	0.322 (0.053) [0.225, 0.418]	CP: .926	0.492 (0.066) [0.370, 0.615]	CP: .932
p-uniform								
Estimator *p*	−0.059 (0.224) [−0.427, 0.313]	CP: .952	0.112 (0.187) [−0.224, 0.418]	CP: .950	0.298 (0.142) [0.031, 0.539]	CP: .951	0.481 (0.103) [0.282, 0.677]	CP: .952
Estimator 1-*p*	−0.011 (0.271) [−0.468, 0.367]	CP: .948	0.140 (0.240) [−0.307, 0.459]	CP: .949	0.313 (0.188) [−0.094, 0.564]	CP: .952	0.493 (0.137) [−0.183, 0.686]	CP: .947

Note. K = total number of studies; p_p = proportion of nonsignificant studies included in a meta-analysis; () = average standard error or, in case of *p*-uniform, standard deviation of all 10,000 estimates; [] = average bounds of 95% confidence interval; CP = coverage probability.

row of Table 1), so exactly equal to the nominal coverage rates, confirming that *p*-uniform performs very well when its assumptions are satisfied. Figure 3 presents the average effect size estimates with the proportion of statistically nonsignificant studies included in the meta-analysis (p_p) on the *x*-axis, and on the *y*-axis the population effect size μ (horizontal dotted lines) and the average effect size estimates ($\hat{\mu}$ and $\hat{\mu}^*$). *p*-uniform's average effect size estimates are indicated by an asterisk (estimator $p_i^{\mu^*}$) and a cross (estimator $1 - p_i^{\mu^*}$) on the *y*-axis. *p*-uniform's average effect size estimate ($\hat{\mu}^*$) had a slight negative bias for both estimators, which was significantly different from zero in some conditions. That is, bias of estimator $p_i^{\mu^*}$ for $\mu = 0$, $\mu = 0.16$, $\mu = 0.33$, $\mu = 0.5$ was −0.059 ($-z = 5.9, p < .001$), −0.048 ($-z = 4.8, p < .001$), −0.035 ($-z = 3.5, p < .001$), −0.018 ($-z = 1.8, p = .065$), respectively, and of estimator $1 - p_i^{\mu^*}$ it was −0.011 ($-z = 1.1, p = .27$), −0.020 ($-z = 2.0, p = .046$), −0.020 ($-z = 2.0, p = .046$), −0.007 ($-z = 0.7, p = .48$), respectively.[5] Apparently, for the conditions in the simulations, the estimator $1 - p_i^{\mu^*}$ slightly outperformed the estimator $p_i^{\mu^*}$.

Average effect size estimates of the fixed-effect model and the trim-and-fill method are presented as a function of p_p and population effect size μ using lines in Figure 3. Unsurprisingly, the fixed-effect model and the trim-and-fill method yielded accurate average effect size estimates in cases of no publication bias ($p_p = 1$). In particular, average effect size estimates obtained by the fixed-effect model (open bullets) fell exactly on the dotted lines reflecting the population effect size μ. Without publication bias ($p_p = 1$), the average effect size estimates of the trim-and-fill

[5] $z = \frac{\mu - \hat{\mu}}{1/\sqrt{10{,}000}}$, where μ is the population value, $\hat{\mu}$ is the effect size estimate, $1/\sqrt{10{,}000}$ the standard error of $\hat{\mu}$.

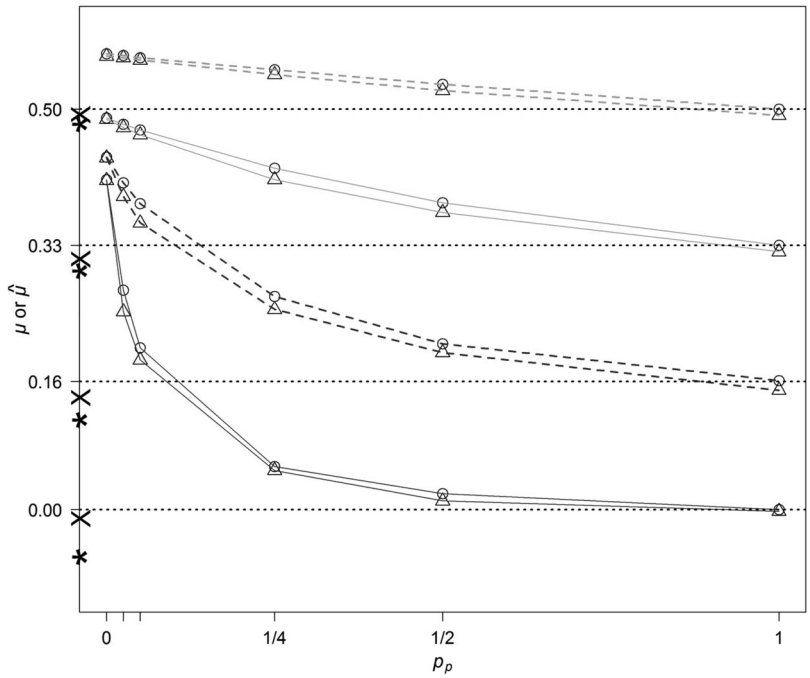

Figure 3. Average effect size estimates of the fixed-effect model, the trim-and-fill method, and *p*-uniform as a function of the proportion *p* of non-significant studies included in the meta-analysis and the population effect size μ. Average effect size estimates are indicated by open bullets (traditional fixed-effect model), triangles (trim-and-fill), asterisks (*p*-uniform estimator *p*), and crosses (*p*-uniform estimator 1–*p*). Dotted black lines illustrate the population effect size μ. Solid black lines refer to $\mu = 0$, dashed black lines refer to $\mu = 0.16$, solid gray lines refer to $\mu = 0.33$, and dashed gray lines refer to $\mu = 0.5$.

method (triangles in Figure 3) slightly underestimated the population effect size μ ($\hat{\mu} = 0.49$). This underestimation of the average effect size was caused by the imputation of studies while no studies were missing. Table 2 shows the average number of studies imputed by the trim-and-fill method in each condition. The first row of the last column indicates that, on average, nine studies were imputed when there was no effect ($\mu = 0$) and no publication bias ($p_p = 1$), resulting in an underestimated effect. The other rows in Table 2 also illustrate the poor performance of the trim-and-fill method. If the proportion of statistically nonsignificant studies included in the meta-analysis (p_p) decreases, more studies are omitted from the meta-analysis and the trim-and-fill method should impute more studies. However, this is not the case because the trim-and-fill method hardly ever imputed studies if there was extreme publication bias (cf. $p_p = 0$, third column in Table 2).

In conditions with publication bias ($p_p < 1$), the fixed-effect model and the trim-and-fill method severely overestimated effect sizes. This is shown in Figure 3: As publication bias increased, the lines representing the fixed-effect model and the trim-and-fill method deviated more strongly from the population effect size μ. These average effect size estimates deviated more from the population effect size μ when there was at the same time no effect ($\mu = 0$) and extreme publication bias ($p_p = 0$), with $\hat{\mu} = 0.41$ for both the fixed-effect model and the trim-and-fill method (see the first two rows of the first column in Table 1). If there was actually an effect in the population ($\mu > 0$), the overestimation in average

Table 2
Average Number of Imputed Studies by the Trim-and-Fill Method Based on Monte-Carlo Simulations (10,000 Replications)

	p_p					
μ (*K*)	0	0.025	0.05	0.25	0.5	1
0 (160)	0.06 (0.25)	1.41 (1.45)	0.82 (1.18)	0.59 (1.84)	2.01 (4.84)	9.00 (12.06)
0.16 (40)	0.05 (0.25)	0.82 (1.36)	1.17 (1.48)	0.90 (1.33)	0.81 (1.52)	1.49 (2.74)
0.33 (16)	0.07 (0.28)	0.17 (0.59)	0.27 (0.74)	0.61 (1.07)	0.61 (1.09)	0.49 (1.08)
0.5 (10)	0.10 (0.36)	0.11 (0.39)	0.12 (0.50)	0.21 (0.59)	0.27 (0.68)	0.30 (0.72)

Note. Studies were imputed on the left-hand side of the funnel plot. p_p = proportion of nonsignificant studies included in the meta-analysis; μ = the effect size estimate used for simulating data; (*K*) = total number of studies; () = standard deviation.

effect sizes of both the fixed-effect model and the trim-and-fill method decreased in μ. The lines belonging to the fixed-effect model and the trim-and-fill method for $\mu = 0.5$ (dashed gray lines in Figure 3) diverged less from its population effect size μ than the lines belonging to both methods for $\mu = 0$ (solid black lines).

Coverage probabilities of both the fixed-effect model and the trim-and-fill method were far below the nominal 95% rate for $\mu < 0.5$ and whenever publication bias was present ($p_p < 1$). For conditions without an effect ($\mu = 0$) and extreme publication bias ($p_p = 0$), coverage probabilities were even close to 0 (see the first two rows of the first column in Table 1). Coverage probabilities became closer to the nominal rate as the effect increased and the amount of publication bias decreased. However, the coverage probability was still unsatisfactory in condition $\mu = 0.33$ and $p_p = .5$ for the fixed-effect model (0.88) and the trim-and-fill method (0.86). Coverage probabilities of both methods approached 95% when $\mu = 0.5$ and in conditions without publication bias ($p_p = 1$).

To conclude, coverage probabilities of *p*-uniform were 95% in all conditions, which did not apply to the fixed-effect model and the trim-and-fill method. Average effect size estimates of *p*-uniform were accurate, albeit slightly underestimated. Average effect size estimates of the fixed-effect model and the trim-and-fill method substantially deviated from the population effect size μ except for a medium size population effect ($\mu = 0.5$) and no publication bias ($p_p = 1$). At the same time, the standard deviations of *p*-uniform's effect size estimates were substantially larger than those of the fixed-effect model and the trim-and-fill method. As a consequence, average effect size estimates of *p*-uniform were accurate but more uncertain. In contrast, the results of the fixed-effect model and the trim-and-fill method provided false certainty. These estimates were precise but highly inaccurate if the population effect size μ was smaller than medium ($\mu < 0.5$) and publication bias was present ($p_p < 1$).

Test of an Effect When Effects Are Homogenous

In Table 3, Type-I error rates and statistical power of the fixed-effect model, the trim-and-fill method, and estimator $p_t^{\mu^*}$ of *p*-uniform are presented for testing whether the population effect size equals 0. *p*-uniform's Type-I error rates were exactly equal to the nominal rate in all conditions (see third row of the last column in Table 3). Statistical power of *p*-uniform increased in μ from 0.26 for $\mu = 0.16$ to 0.98 for $\mu = 0.5$. Consequently, *p*-uniform already has very high power to detect a medium effect size ($\mu = d = 0.5$) when only eight studies with $n = 25$ are statistically significant.

If there was no effect ($\mu = 0$) and no publication bias ($p_p = 1$), the Type-I error rate of the trim-and-fill method was lower than the nominal rate ($\alpha = .035$). This was caused by the imputation of studies while no publication bias was present (see Table 2). If there was publication bias ($p_p < 1$) the Type-I error rates were grossly overestimated by the fixed-effect model and the trim-and-fill method. The Type-I error rates increased as publication bias became more severe. If there was no effect ($\mu = 0$) and extreme publication bias ($p_p = 0$), Type-I error rates of the fixed-effect model and the trim-and-fill method equaled 1 (see first two rows of the first column in Table 3) meaning that both methods always yielded a Type-I error in this condition. This Type-I error rate was severely inflated due to overestimated average effect size estimates by both methods as explained in the previous section (see also Table 1 and Figure 3).

The fixed-effect model and the trim-and-fill method were powerful in detecting an effect when it truly existed ($\mu > 0$) and no publication bias was present ($p_p = 1$). The levels of statistical power rapidly approached one for these conditions (see last column in Table 3). If there was an effect ($\mu > 0$) and publication bias was present ($p_p < 1$), the statistical power of the fixed-effect

Table 3
Results of Monte-Carlo Simulations (10,000 Replications) on Type-I Error Rates and Statistical Power for Testing Whether the Effect Size Is Significantly Different From Zero

	p_p					
	0	1/40	1/20	1/4	1/2	1
μ (K)						
0 (160)						
Fixed-effect model	1.000	0.985	0.952	0.566	0.249	0.053
Trim-and-fill	1.000	0.978	0.939	0.524	0.208	0.035
p-uniform (estimator *p*)						0.050
0.16 (40)						
Fixed-effect model	1.000	1.000	1.000	0.998	0.999	0.999
Trim-and-fill	1.000	1.000	0.999	0.996	0.996	0.990
p-uniform (estimator *p*)						0.259
0.33 (16)						
Fixed-effect model	1.000	1.000	1.000	1.000	1.000	1.000
Trim-and-fill	1.000	1.000	1.000	1.000	1.000	1.000
p-uniform (estimator *p*)						0.722
0.5 (10)						
Fixed-effect model	1.000	1.000	1.000	1.000	1.000	1.000
Trim-and-fill	1.000	1.000	1.000	1.000	1.000	1.000
p-uniform (estimator *p*)						0.980

Note. p_p = proportion of nonsignificant studies included in a meta-analysis; μ = the effect size estimate for simulating data; (*K*) = total number of studies.

model and the trim-and-fill method was close to 1 or equaled 1 in every condition. However, these results reflect false certainty because effect size estimates of both methods were overestimated due to the presence of publication bias (see previous section).

To summarize, the accurate proportion of Type-I errors was made for testing whether the population effect size equals 0 based on *p*-uniform and *p*-uniform's statistical power was high to detect a population effect of medium size ($\mu = 0.5$) with only eight small statistically significant studies. The fixed-effect model and the trim-and-fill method overestimated the effect size in case of publication bias and therefore yielded many Type-I errors and false certainty with respect to the presence of population effects.

Publication Bias Test When Population Effects Are Homogenous

Table 4 shows Type-I error rates and statistical power of two publication bias tests: estimator $p_i^{\mu^*}$ of *p*-uniform and the TES. Type-I error rates of *p*-uniform were close to 5% in the conditions $\mu < 0.5$ without publication bias ($p_p = 1$; see last column in Table 4). With $\mu = 0.5$ and without publication bias ($p_p = 1$), Type-I error rates obtained by *p*-uniform were lower than the nominal rate ($\alpha = .012$). *p*-uniform had reasonable statistical power when a considerable number of studies had been excluded from the meta-analysis. For example, statistical power of the method was 0.75 for $\mu = 0.16$ and extreme publication bias ($p_p = 0$; see fourth row of first column in Table 4).

The last column in Table 4 illustrates that in conditions without publication bias ($p_p = 1$) the TES was more conservative than *p*-uniform. Type-I error rates of the TES ranged from 0.022 for no effect ($\mu = 0$) to 0.003 for $\mu = 0.5$. With one exception, the TES was less powerful than *p*-uniform in detecting publication bias. This exception was that the TES had more power if no effect existed ($\mu = 0$) and at least some statistically nonsignificant studies were published ($p_p > 0$). *p*-uniform had more statistical power to detect publication bias if there was no effect ($\mu = 0$) and extreme publication bias ($p_p = 0$), and if an effect indeed existed ($\mu > 0$).

The statistical power of the TES and *p*-uniform was low for the two largest population effect sizes ($\mu = 0.33$ and $\mu = 0.5$). For example, for $\mu = 0.5$ the statistical power was not higher than 0.03 for *p*-uniform and 0.001 for the TES. The statistical power of *p*-uniform was low for two reasons. First, few studies were statistically significant (eight on average) resulting in a wide confidence interval for the average effect size estimate. Second, few studies were not statistically significant (on average two for $\mu = 0.5$ or eight for $\mu = 0.33$), such that the average effect size estimate based on all studies was close to the average effect size estimate based on only the statistically significant studies. In conditions where only few studies were omitted from the meta-analysis, which occurred when the population effect size or a study's power is high, publication bias was hard to detect.

To conclude, both publication bias tests were too conservative, but this conservatism was higher for the TES. *p*-uniform had higher statistical power than the TES when there was an effect ($\mu > 0$). *p*-uniform was especially powerful compared with the TES when no or only a limited amount of statistically nonsignificant studies were included in the meta-analysis. This is a common situation in meta-analytical reviews, particularly in psychology (Fanelli, 2012).

Estimation of Effects When Population Effects Are Heterogeneous

Here we study the performance of the methods under violations of a homogeneous population effect. Convergence rates for the effect size estimates with *p*-uniform were above 98.3% and 99.2% across conditions for the $p_i^{\mu^*}$ and $1 - p_i^{\mu^*}$ estimator, respectively. Table 5 shows average effect size estimates, standard errors or standard deviations of the effect size estimates, and coverage probabilities of the random-effects model (with the most frequently used DerSimonian Laird procedure), the trim-and-fill method, and *p*-uniform. We compare the results of the three methods to each other, but also compare them with the results of these methods when effects are homogenous (see Table 1). First, note how introducing heterogeneity increases the number of significant studies from eight when effect size is homogenous or $\mu = .33$, to 32.8 when heterogeneity is

Table 4
Results of Monte-Carlo Simulations (10,000 Replications) on Type-I Error Rates and Statistical Power for Publication Bias Tests

	p_p					
	0	1/40	1/20	1/4	1/2	1
μ (*K*)						
0 (160)						
p-uniform (est. *p*)	0.902	0.519	0.340	0.090	0.063	0.051
TES	0.555	0.570	0.644	0.565	0.239	0.022
0.16 (40)						
p-uniform (est. *p*)	0.748	0.620	0.520	0.184	0.092	0.050
TES	0.338	0.245	0.185	0.065	0.029	0.006
0.33 (16)						
p-uniform (est. *p*)	0.365	0.342	0.319	0.182	0.100	0.043
TES	0.074	0.068	0.061	0.023	0.005	0.002
0.5 (10)						
p-uniform (est. *p*)	0.033	0.032	0.031	0.024	0.019	0.012
TES	0.001	0.001	0.001	0.001	0.002	0.003

Note. p_p = proportion of nonsignificant studies included in a meta-analysis; μ = the effect size estimate for simulating data; (*K*) = total number of studies; TES = test for excess significance.

Table 5
Average Effect Size Estimates and Corresponding Standard Errors/Standard Deviations, and Coverage Probabilities for the Random-Effects Model, the Trim-and-Fill Method, and p-Uniform Based on Monte-Carlo Simulations of Heterogeneous Population Effects (10,000 Replications)

		$\mu = 0, K = 160$			$\mu = 0.33, K = 16$		
	Heterogeneity (τ^2)	Low (0.0133)	Mod. (0.04)	High (0.12)	Low (0.0133)	Mod. (0.04)	High (0.12)
	No. of significant studies	12.33 (3.35)	19.62 (4.11)	32.80 (5.16)	8.04 (2.01)	8.02 (2.00)	8.02 (2.00)
$p_p = 0$	Random-effects model	0.433 (.059)	0.469 (.046)	0.554 (.036)	0.514 (.073)	0.554 (.075)	0.648 (.075)
		CP < .0001	CP ≤ .0001	CP ≤ .0001	CP = .225	CP = .107	CP = .035
	Trim-and-fill	0.433 (.059)	0.469 (.046)	0.554 (.036)	0.512 (.073)	0.553 (.075)	0.645 (.089)
		CP < .0001	CP < .0001	CP < .0001	CP = .229	CP = .112	CP = .039
$p_p = 0.25$	Random-effects model	0.081 (.039)	0.126 (.044)	0.211 (.054)	0.447 (.072)	0.473 (.081)	0.532 (.097)
		CP = .441	CP = .185	CP = .026	CP = .604	CP = .535	CP = .505
	Trim-and-fill	0.071 (.039)	0.099 (.045)	0.145 (.055)	0.428 (.071)	0.449 (.081)	0.497 (.111)
		CP = .539	CP = .404	CP = .268	CP = .655	CP = .594	CP = .570
p-uniform	Estimator *p*	0.091 (.160)	0.262 (.102)	0.503 (.075)	0.367 (.137)	0.464 (.137)	0.641 (.153)
		CP = .827	CP = .223	CP < .0001	CP = .887	CP = .644	CP = .206
	Estimator 1-*p*	0.086 (.219)	0.228 (.141)	0.406 (.080)	0.357 (.187)	0.428 (.173)	0.535 (.163)
		CP = .895	CP = .593	CP = .045	CP = .926	CP = .840	CP = .610

Note. K = total number of studies; p_p = proportion of non-significant studies included in a meta-analysis; () = average standard error or, in case of *p*-uniform, standard deviation of all 10,000 estimates; CP = coverage probability.

high and $\mu = 0$ (second row of Table 5). Consequently, *p*-uniform uses relatively more than 5% (up to about 20%) of the studies if no effect exists and effects are heterogeneous.

From the results of Table 5 and comparing its results with those in Table 1, it follows that random-effects meta-analysis and the trim-and-fill method perform worse as heterogeneity increases; both bias increases and the coverage probability decreases in heterogeneity. Moreover, the estimate of heterogeneity (τ^2) is biased in random-effects meta-analysis as well; for example, τ^2 is severely underestimated if only statistically significant studies are published, whereas τ^2 is grossly overestimated if 25% of the statistically insignificant studies are published (not shown in Table 5). The trim-and-fill method on average imputed less than .1 studies if only statistically significant studies are published (also when about 130 or more studies were omitted), and up to 6.3 studies when 25% of the statistically insignificant studies are published and no effect exists (when on average about 95 studies were omitted; not shown in Table 5). To conclude, the performance of random-effects meta-analysis and the trim-and-fill method is bad in case of publication bias and worsens when heterogeneity increases.

Whereas the performance of *p*-uniform is excellent when effects are homogenous (see Table 1), performance worsens when heterogeneity increases; both bias increased and the coverage probability decreased in heterogeneity (see Table 5). As expected, estimator $1 - p_i^{\mu^*}$ is more robust to heterogeneity than estimator $p_i^{\mu^*}$. However, in our opinion the performance of $1 - p_i^{\mu^*}$ is only acceptable when heterogeneity is low, with coverage probabilities of .895 and .926 and bias of .086 and .047 for $\mu = 0$ and $\mu = .33$, respectively. Both *p*-uniform estimators outperformed traditional random-effects meta-analysis and the trim-and-fill method under conditions of heterogeneity when statistically insignificant studies are not published ($p_p = 0$), but not when $p_p = 0.25$. This suggests that if effects are heterogeneous, *p*-uniform only outperforms the other methods when publication bias is extreme (with p_p close to 0). To conclude, *p*-uniform is generally not robust to heterogeneous effects, only provides acceptable estimates if heterogeneity is low, and outperforms other methods only if publication bias is extreme under conditions of heterogeneity.

Application to Meta-Analysis of McCall and Carriger (1993)

McCall and Carriger (1993) carried out a meta-analysis on studies examining the association between infants' habituation to a give stimulus and their later cognitive ability (IQ). Their meta-analysis used 12 studies with sample sizes varying from 11 to 96 reporting a correlation between children's habituation during their first year of life and their IQ as measured between 1 and 8 years of age (see also: Bakker et al., 2012). Of these 12 correlations, 11 were statistically significant, and one was not, $r = .43$, $p = .052$. Because there was no indication of heterogeneity in the studies' effect sizes ($\chi^2 = 6.74, p = .82, I^2 = 0$), a fixed-effect meta-analysis was performed on the 12 studies. This resulted in a Fisher-transformed correlation of .41 ($p < .001$), corresponding to an estimated correlation of .39 (CI 95%: [.31, .47]).

The apparent negative association between effect size and standard error in the contour-enhanced funnel plot (see Figure 1) suggests publication bias. This is confirmed by both Begg and Mazumdar's rank-correlation test ($\tau = 0.636, p = .003$) and Egger's test ($z = 2.24$, $p = .025$). The TES also provides evidence for the presence of publication bias ($\chi^2 = 6.22$, $p = .013$). The funnel plot after application of the trim-and-fill technique using statistic L_0 is presented in Figure 4. Six studies were imputed to the left. Trim-and-fill's estimate of the Fisher-transformed correlation was .35 ($p < .001$), corresponding to an estimated correlation of .34 (CI 95% [.26, .41]). Based on the R_0 statistic, nine studies were imputed reducing the Fisher-transformed correlation to 0.31 ($p < .001$). The untransformed correlation coefficient based on the R_0 statistic became .30 (CI 95% [.23, .37]). Hence, the trim-and-fill method reduced the estimated correlation somewhat for both statistics (from .39 to

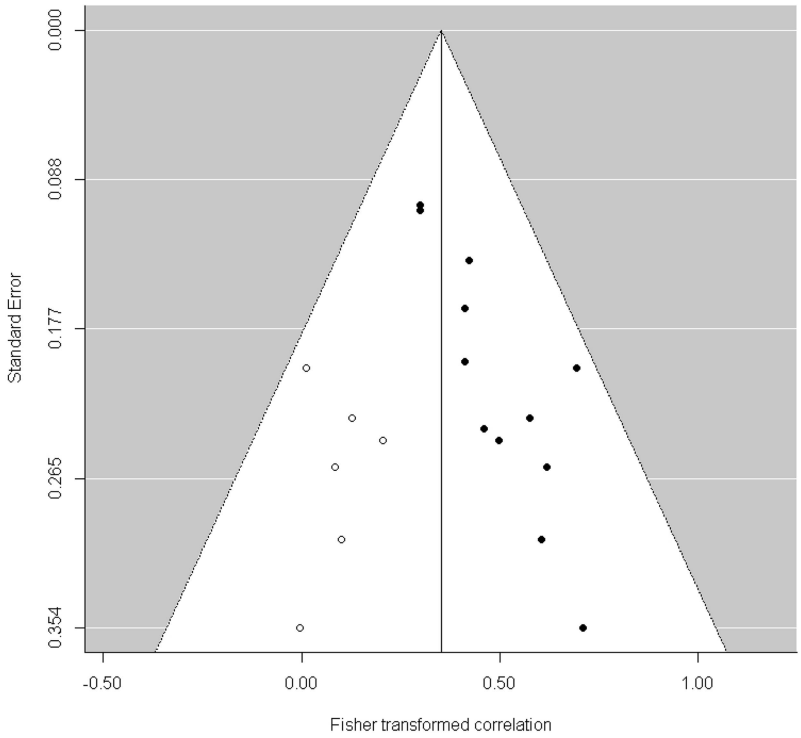

Figure 4. Funnel plot of the studies of McCall and Carriger's (1993) meta-analysis after the trim-and-fill method imputed six studies (open circles) based on the L_0 statistic. The vertical line corresponds to trim-and-fill's effect size of 0.352.

.34 for L_0 and .30 for R_0), but still suggested a significant and medium correlation.

The $p_i^{\mu^*}$ estimator of *p*-uniform was performed on the 11 statistically significant studies. The publication bias test indicated publication bias ($L^{\hat{\mu}} = 4.07$, $p = .003$).[6] Its estimated Fisher-transformed correlation was .175, corresponding to an estimated correlation of .17 (95% CI [−.027, .35]), which did not differ significantly from 0 ($L^O = 17.35$, $p = .083$, two-tailed test). To conclude, the effect size estimate obtained by *p*-uniform is remarkably lower than the fixed-effect estimate, and suggests that the evidence in favor of a positive association between infants' habituation and their later cognitive ability (IQ) is not conclusive.

Discussion

Publication bias is a major threat to meta-analytical reviews (Banks, Kepes, & McDaniel, 2012; Rothstein et al., 2005), and is omnipresent in many fields of scientific research. Hence, publication bias analyses should be routinely included in meta-analysis (e.g., Borenstein et al., 2009, p. 291; Rothstein et al., 2005). Current techniques cannot provide accurate average effect size estimates and should be interpreted as sensitivity analyses, and tests for publication bias often suffer from a lack of power (e.g., Begg & Mazumdar, 1994; Borenstein et al., 2009, p. 291; Sterne et al., 2000) or are overly conservative (Francis, 2012; Ioannidis & Trikalinos, 2007b). Due to overestimated average effect sizes in case of publication bias, Type-I error rates of statistical tests for testing whether the population effect size is zero become strongly inflated. The objective of this article was to introduce a new method (*p*-uniform) that can (a) accurately estimate average effect size in case of publication bias, (b) test whether the population effect size is zero, and (c) test for publication bias. *p*-uniform is counterintuitive for meta-analysts because the method only takes the *p* values of statistically significant studies into account. The basic idea of *p*-uniform is that the distribution of the statistically significant *p* values conditional on the population effect size is uniform. Our simulation study compared the performance of *p*-uniform with the TES, the fixed-effect model, and the trim-and-fill method by means of a simulation study. Stringent conditions for examining the performance of *p*-uniform were selected, with small numbers of studies included in the meta-analysis and small sample sizes for each individual study.

Results of the main simulation study on homogenous population effect sizes showed good statistical properties of *p*-uniform in comparison with the trim-and-fill method, TES, and standard fixed-effects meta-analysis. Coverage probabilities of *p*-uniform were always 95%, whereas *p*-uniform's slightly underestimated the population effect. Our results and those of others (Moreno et al., 2009; Peters et al., 2007; Terrin et al., 2003) clearly show that the fixed-effect model and the trim-and-fill method cannot be trusted when there is publication bias. The average effect size estimates and coverage probabilities of existing methods were only

[6] All test statistics of *p*-uniform are compared with a Gamma distribution with $df_1 = 1$ and $df_2 = 11$. F *z*

acceptable in the absence of publication bias ($p_p = 1$) or sufficient power in the primary studies (.80 for $\mu = 0.5$). For testing whether the population effect is zero, the Type-I error rate of *p*-uniform was exactly equal to the nominal rate, and *p*-uniform's statistical power was high to detect a population effect of medium size. The fixed-effect model and the trim-and-fill method yielded too many Type-I errors if publication bias was present. Both *p*-uniform and the TES for the presence of publication bias were too conservative. However, *p*-uniform's publication bias test outperformed the TES in most conditions of homogenous population effects. An additional simulation study on heterogeneous population effects revealed that both *p*-uniform and other fixed-effects techniques performed poorly under increasing heterogeneity. Our transformed estimator $1 - p_i^{\mu^*}$ was more robust to heterogeneity than estimator $p_i^{\mu^*}$, but its performance was only acceptable if heterogeneity was low. However, the transformed estimator did outperform other fixed-effect techniques when publication bias was extreme.

p-uniform did not converge to an effect size estimate in a small percentage of the simulations (<2%) when no effect existed. The reason of the nonconvergence is the small number of studies in combination with the distribution of *p* values under the null-hypothesis of no effect; *p*-uniform sometimes cannot estimate μ if all *p* values are higher than .025 and close to .05. Because this is unlikely as *K* increases, the nonconvergence problem quickly disappears if *K* increases. For instance, *p*-uniform's convergence rates were all above 99.9% if the number of studies was twice as large as in the conditions with homogeneous population effects, with 16 rather than eight expected statistically significant studies.

The effect size estimates of both estimators *p*-uniform based on Fisher's method were slightly negatively biased. The negative bias is a consequence of the estimate $\hat{\mu}$ being a nonlinear function of *p*. We first examined the bias for estimating $\hat{\mu}$ on the basis of one single statistically significant study. The expected value of $\hat{\mu}$ turned out negative because *p* values close to .05 yielded very negative estimates of μ. The negative bias decreases in the study's sample, with factor $\sqrt{N}$, and in population effect size μ. Additional simulations, with on average twice as many statistically significant studies in a meta-analysis (16 instead of eight), suggested that the bias also decreases in the number of statistically significant studies whenever effect size is larger than zero, although the bias did not disappear entirely. Future studies should consider examining systematically the performance of other statistical tests for uniformity than those based on Fisher's method (such as *p*-uniforms $p_i^{\mu^*}$ and $1 - p_i^{\mu^*}$ estimator; e.g., using the fact that the expected value of the uniform distribution equals 0.5, the Kolmogorov–Smirnov test (Massey, 1951), and the Anderson-Darling test (Anderson & Darling, 1954)) decrease this bias in effect size estimates and also provide lower standard errors than we obtain with Fisher's method.

The newly proposed *p*-uniform method has numerous advantages over existing techniques in examining and correcting for publication bias. First of all, it is the first method that can provide an effect size estimate, test whether the population effect is zero, and test for publication bias at the same time. An important second advantage of *p*-uniform is that, even though power may be low for testing publication bias in applications with a small number of studies, the average effect size is accurately estimated by *p*-uniform when its assumptions are satisfied. When there is publication bias and effects are homogenous, *p*-uniform has good statistical properties compared with fixed-effect meta-analyses, the TES, and the trim-and-fill method. Our study did not compare *p*-uniform's performance with that of Egger's and the rank correlation test. However, because other studies (e.g., Moreno et al., 2009) showed that the latter two methods had low power for the conditions with eight studies examined in our simulation study, it is likely *p*-uniform also outperforms them. Third, no sophisticated assumptions or choices have to be made when applying *p*-uniform. No additional (unpublished) data have to be collected and interpretation of the results is straightforward. Hence, in principle, meta-analysts should be able to easily apply *p*-uniform in their research. We are currently working on developing a website that will have R programs enabling researchers to apply *p*-uniform to their research. Finally, *p*-uniform will provide conservative effect size estimates in case of researcher *df*, rather than further overestimating effect size.

We suggest a number of recommendations for the practice of meta-analysis. First, because publication bias is ubiquitous and effects may be small or nonexistent, we follow-up on others (e.g., Aytug et al., 2012; Banks, Kepes, & McDaniel, 2012; Field & Gillett, 2010; Sterne, Gavaghan, & Egger, 2000; Sutton, 2006) by recommending the application of publication bias analysis in all meta-analyses. We recommend applying *p*-uniform to estimate average effect size and to test for publication bias if the population effect is homogenous, or to apply *p*-uniform as a sensitivity analysis to address and examine publication bias in meta-analyses. Although the restriction to homogenous effects may seem to restrict the potential usefulness of *p*-uniform, examinations of results of meta-analyses suggest that there is no evidence of heterogeneity in about half of the meta-analyses in psychology based on lab studies (Klein et al., 2014), and medicine (Borenstein et al., 2009, p. 119). Also, it is often feasible to select on the basis of theoretical and methodological considerations homogeneous subsets of studies that are reasonably expected to have one underlying population effect. Another alternative may be to apply selection models as sensitivity analysis whenever there is strong evidence for heterogeneity, because other techniques provide misleading results when effects are heterogeneous (Hedges & Vevea, 2005).

Future studies should examine how *p*-uniform performs (compared with selection models and other existing methods) if its assumptions are violated, and how *p*-uniform may be adapted to be more robust to violations of heterogeneity. Although our results show that *p*-uniform's $1 - p_i^{\mu^*}$ estimator is more robust than the $p_i^{\mu^*}$ estimator, other estimators can be developed that are even more robust. Notably, methods to incorporate heterogeneity in the estimation could be examined in the future, for example, by specifying a distribution of effects sizes rather than one fixed effect size (as is done in selection models). *p*-uniform's performance also has to be examined in conditions where the probability of publishing depends on the *p* value lower than 0.05. The effect of researcher *df* on *p*-uniform's performance also deserves attention in future studies. Researcher *df* will lead to a lower average effect size estimate obtained by *p*-uniform because studies with *p* values just below .05 are overrepresented. Performance of *p*-uniform should also be evaluated in less restrictive conditions than the selected conditions in the present simulation studies. For instance, in theory, *p*-uniform should perform just as well when

studies vary in sample size; in conditions with studies varying in sample size the performance of *p*-uniform can then also be compared with Egger's test and the rank correlation test. Finally, following others (Banks, Kepes, Banks, 2012; Banks, Kepes, & McDaniel, 2012), we recommend conducting publication bias analyses in both past and future meta-analytic studies. Moreover, following Banks, Kepes, and Banks (2012, p. 193), we encourage journals to publish reevaluations of previous meta-analytic reviews regardless of their results to avoid 'publication bias in publication bias results'.

Publication bias can distort the validity of meta-analyses and may lead to false conclusions with far-reaching consequences. Current meta-analytic techniques perform well in the absence of publication bias. However, it cannot be assumed that there is no publication bias in a particular research field because not all file-drawers can be opened, and relevant studies will be below the radar of meta-analysts. As a consequence, traditional techniques may lead to unreliable results as this study and other studies have shown. *p*-uniform takes a different perspective on analyzing meta-analytical datasets to counteract this problem. In simulations, *p*-uniform showed promising results that were superior to those from existing methods. The method still needs further development, but can become the technique for examining publication bias and estimating population effects in meta-analytic reviews.

References

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. http://dx.doi.org/10.1177/0149206310377113

Aguinis, H., Pierce, C. A., Bosco, F. A., Dalton, D. R., & Dalton, C. M. (2011). Debunking myths and urban legends about meta-analysis. *Organizational Research Methods, 14,* 306–331. http://dx.doi.org/10.1177/1094428110375720

Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association, 49,* 765–769. http://dx.doi.org/10.1080/01621459.1954.10501232

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27,* 108–119. http://dx.doi.org/10.1002/per.1919

Aytug, Z. G., Rothstein, H. R., Zhou, W., & Kern, M. C. (2012). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods, 15,* 103–133. http://dx.doi.org/10.1177/1094428111403495

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7,* 543–554. http://dx.doi.org/10.1177/1745691612459060

Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication bias: The antagonist of meta-analytic reviews and effective policymaking. *Educational Evaluation and Policy Analysis, 34,* 259–277. http://dx.doi.org/10.3102/0162373712446144

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,* 182–197. http://dx.doi.org/10.1111/j.1468-2389.2012.00591.x

Becker, B. J. (2005). Failsafe *N* or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 111–125). Chichester, UK: Wiley.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50,* 1088–1101. http://dx.doi.org/10.2307/2533446

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483,* 531–533. http://dx.doi.org/10.1038/483531a

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: Wiley, Ltd. http://dx.doi.org/10.1002/9780470743386

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. http://dx.doi.org/10.1037/1082-989X.2.4.447

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology: Research and Practice, 17,* 136–137. http://dx.doi.org/10.1037/0735-7028.17.2.136

Duval, S. (2005). The trim and fill method. In H. R. Rothstein, M. Borenstein, & A. J. Sutton (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 127–144). Chichester, UK: 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. http://dx.doi.org/10.1111/j.0006-341X.2000.00455.x

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315,* 629–634. http://dx.doi.org/10.1136/bmj.315.7109.629

Ellis, P. D. (2010). *The essential guide to effect sizes: An introduction to statistical power, meta-analysis and the interpretation of research results*. Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511761676

Fanelli, D. (2010). Do pressures to publish increase scientists' bias? An empirical support from U.S. States data. *PLoS ONE, 5,* e10271. http://dx.doi.org/10.1371/journal.pone.0010271

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90,* 891–904. http://dx.doi.org/10.1007/s11192-011-0494-7

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17,* 120–128. http://dx.doi.org/10.1037/a0024445

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *The British Journal of Mathematical and Statistical Psychology, 63,* 665–694. http://dx.doi.org/10.1348/000711010X502733

Fisher, R. A. (1932). *Statistical methods for research workers*. London, UK: Oliver & Boyd.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19,* 975–991. http://dx.doi.org/10.3758/s13423-012-0322-y

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology, 57,* 153–169. http://dx.doi.org/10.1016/j.jmp.2013.02.003

Fritz, A., Scherndl, T., & Küberger, A. (2013). *Publication bias and the correlation between effect size and sample size in psychological research: Sources, consequences, and remedies*. Manuscript submitted for publication.

Gerber, A. S., Green, D. P., & Nickerson, D. (2001). Testing for publication bias in political science. *Political Analysis, 9,* 385–392. http://dx.doi.org/10.1093/oxfordjournals.pan.a004877

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9,* 61–85. http://dx.doi.org/10.2307/1164832

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21,* 299–332. http://dx.doi.org/10.3102/10769986021004299

Hedges, L. V., & Vevea, J. L. (2005). Selection method approaches. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 145–175). Chichester, UK: Wiley.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine, 2,* e124. http://dx.doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P., & Trikalinos, T. A. (2007a). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal, 176,* 1091–1096. http://dx.doi.org/10.1503/cmaj.060410

Ioannidis, J. P., & Trikalinos, T. A. (2007b). An exploratory test for an excess of significant findings. *Clinical Trials, 4,* 245–253. http://dx.doi.org/10.1177/1740774507079441

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods, 15,* 624–662. http://dx.doi.org/10.1177/1094428112452760

Kepes, S., Banks, G. C., & Oh, I. S. (2012). Avoiding bias in publication bias research: The value of null findings. *Journal of Business and Psychology*. Advance online publication.

Kisamore, J., & Brannick, M. (2008). An illustration of the consequences of meta-analysis model choice. *Organizational Research Methods, 11,* 35–53. http://dx.doi.org/10.1177/1094428106287393

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahnik, S., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45,* 142–152. http://dx.doi.org/10.1027/1864-9335/a000178

Kraemer, H. C., Gardner, C., Brooks, J., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3,* 23–31. http://dx.doi.org/10.1037/1082-989X.3.1.23

Last, J. M. (2001). *A dictionary of epidemiology*. Oxford, UK: Oxford University Press.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1,* 161–175. http://dx.doi.org/10.1007/BF01173636

Massey, F. J., Jr. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association, 46,* 68–78. http://dx.doi.org/10.1080/01621459.1951.10500769

McCall, R. B., & Carriger, M. S. (1993). A meta-analysis of infant habituation and recognition memory performance as predictors of later IQ. *Child Development, 64,* 57–79. http://dx.doi.org/10.2307/1131437

Moreno, S. G., Sutton, A. J., 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,* 2. Retrieved from http://www.biomedcentral.com/1471-2288/9/2 http://dx.doi.org/10.1186/1471-2288-9-2

O'Boyle, E. H., Jr., Banks, G. C., & Gonzalez-Mulé, E. (2014). The Chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management*. Advance online publication. http://dx.doi.org/10.1177/0149206314527133

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine, 26,* 4544–4562. http://dx.doi.org/10.1002/sim.2889

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. http://dx.doi.org/10.1016/j.jclinepi.2007.11.010

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10,* 712. http://dx.doi.org/10.1038/nrd3439-c1

R Core Team. (2012). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Renkewitz, F., Fuchs, H. M., & Fiedler, S. (2011). Is there evidence of publication biases in JDM research? *Judgment and Decision Making, 6,* 870–881.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86,* 638–641. http://dx.doi.org/10.1037/0033-2909.86.3.638

Rothstein, H. R., & Bushman, B. J. (2012). Publication bias in psychological science: Comment on Ferguson and Brannick (2012). *Psychological Methods, 17,* 129–136. http://dx.doi.org/10.1037/a0027128

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 1–7). Chichester, UK: Wiley. http://dx.doi.org/10.1002/0470870168

Sarewitz, D. (2012). Beware the creeping cracks of bias. *Nature, 485,* 149–150. http://dx.doi.org/10.1038/485149a

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. http://dx.doi.org/10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2013). *P*-curve: A key to the file drawer. *Journal of Experimental Psychology*. Retrieved from http://ssrn.com/abstract=2256237

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve and effect size: Correcting for publication bias using only significant results. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2377290

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—Or vice versa. *Journal of the American Statistical Association, 54,* 30–34.

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

Sterne, J. A. C., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 73–98). Chichester, UK: Wiley.

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology, 54,* 1046–1055. http://dx.doi.org/10.1016/S0895-4356(01)00377-8

Sterne, J. A. C., & Egger, M. (2006). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99–110). Chichester, UK: Wiley. http://dx.doi.org/10.1002/0470870168.ch6

Sterne, J. A. C., 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. http://dx.doi.org/10.1016/S0895-4356(00)00242-0

Sutton, A. J. (2006). Evidence concerning the consequences of publication

and related biases. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 175–192). Chichester, UK: Wiley. http://dx.doi.org/10.1002/0470870168.ch10

Terrin, N., Schmid, C. H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology, 58,* 894–901. http://dx.doi.org/10.1016/j.jclinepi.2005.01.006

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22,* 2113–2126. http://dx.doi.org/10.1002/sim.1461

Van Assen, M. A. L. M., Van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M. (2014). Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS ONE,* e84896. http://dx.doi.org/10.1371/journal.pone.0084896

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods, 10,* 428–443. http://dx.doi.org/10.1037/1082-989X.10.4.428

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36,* 1–48.

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6,* 291–298. http://dx.doi.org/10.1177/1745691611406923

Received December 16, 2013
Revision received July 15, 2014
Accepted July 17, 2014 ■