The effect of publication bias on the Q test and assessment of heterogeneity

Augusteijn, Hilde E M; van Aert, Robbie C M; van Assen, M.A.L.M.

Published in:
Psychological Methods

DOI:
10.1037/met0000197

Publication date:
2019

Document Version
Early version, also known as pre-print

Citation for published version (APA):
Augusteijn, H. E. M., van Aert, R. C. M., & van Assen, M. A. L. M. (2019). The effect of publication bias on the Q test and assessment of heterogeneity. *Psychological Methods*, *24*(1), 116-134. https://doi.org/10.1037/met0000197

The following manuscript is a pre-print. This article has been peer reviewed, and has been accepted for publication at Psychological Methods:

Augusteijn, H.E.M., van Aert, R.C.M., & van Assen, M.A.L.M. (in press). The effect of publication bias on the assessment of heterogeneity. *Psychological Methods.*

Feedback, suggestion, comments and remarks are more than welcome. They can be send to h.e.m.augusteijn@tilburguniversity.edu.

Thank you,

Hilde, Robbie and Marcel

THE EFFECT OF PUBLICATION BIAS ON THE ASSESSMENT OF HETEROGENEITY

Hilde E.M. Augusteijn*	Robbie C.M. van Aert	Marcel A.L.M. van Assen
Tilburg University	Tilburg University	Tilburg University
The Netherlands	The Netherlands	The Netherlands

One of the main goals of meta-analysis is to test for and estimate the heterogeneity of effect sizes. We examined the effect of publication bias on the Q-test and assessments of heterogeneity as a function of true heterogeneity, publication bias, true effect size, number of studies, and variation of sample sizes. The present study has two main contributions, and is relevant to all researchers conducting meta-analysis. First, we show when and how publication bias affects the assessment of heterogeneity. The expected values of heterogeneity measures H2 and I2 were analytically derived, and the power and the type I error rate of the Q-test were examined in a Monte-Carlo simulation study. Our results show that the effect of publication bias on the Q-test and assessment of heterogeneity is large, complex, and non-linear. Publication bias can both dramatically decrease and increase heterogeneity in true effect size, particularly if the number of studies is large and population effect size is small. We therefore conclude that the Q-test of homogeneity and heterogeneity measures H2 and I2 are generally not valid when publication bias is present. Our second contribution is that we introduce a web application, Q-sense, which can be used to determine the impact of publication bias on the assessment of heterogeneity within a certain meta-analysis and to assess the robustness of the meta-analytic estimate to publication bias. Furthermore, we apply Q-sense to two published meta-analyses, showing how publication bias can result in invalid estimates of effect size and heterogeneity.

Manuscript accepted for publication, 2018-07-16.

Keywords: Meta-analysis, Publication bias, Heterogeneity, Q-test, Extreme homogeneity

Introduction

Meta-analysis has become the most important tool for researchers to gain an overview of the existing literature within a specific field (e.g., Aguinis, Gottfredson, & Wright, 2011; Head, Holman, Lanfear, Kahn, & Jennions, 2015). Meta-analyses are a form of systematic review that statistically combine the results from similar studies to synthesize available evidence in a specific research area (Rhodes, Turner, & Higgins, 2015). These quantitative systematic reviews aim to combine data across many studies or data sets to obtain a summary estimate of the effects (Ioannidis, 2008). In contrast to narrative reviews, meta-analyses make use of the *effect size* of a study (Borenstein, Hedges, Higgins & Rothstein, 2009). Meta-analyses are increasingly popular within many research areas, including psychology. The annual number of meta-analytic publications in PsycINFO has considerably increased over the years. As of 2018 2,100 meta-analyses are published every year and take up 1.3 per cent of all PsycINFO articles.

Meta-analyses are used both to estimate the true population effect size (i.e. the average true effect size) and to explain *heterogeneity* in this effect. Primary studies, will differ in their designs and may not estimate the same true effect. Hence, differences between the estimated effect sizes of these included studies in a meta-analysis are inevitable (Higgins, Thompson, Deeks, & Altman, 2002). If the variation between study results is greater than that expected by chance alone (sampling error), this is called statistical heterogeneity (e.g., Higgins & Green, 2011). Statistical heterogeneity indicates that true study effects are influenced by clinical factors (e.g., differences in the studied population), substantive heterogeneity factors (e.g., differences in the studied population or the compared (e.g., differences in experimental treatments interventions), or methodological heterogeneity factors (e.g., different study designs or measurement procedures). Even when the mean of the distribution of the true effect size is positive, it is quite possible for heterogeneity to indicate that, in some situations, the effect size is zero or even negative. This provides important qualifications of the average effect. If relevant characteristics are not coded as moderators, a researcher may conclude that the treatment does have the desired effect when it may only work for particular subgroups of participants. In order to gain valid results, every meta-analysis should therefore examine statistical heterogeneity (Hardy & Thompson, 1998; American Psychological Association, 2008). Consensus seems to be growing that random-effects meta-analysis, which incorporates and estimates statistical heterogeneity, should be preferred over fixed-effect meta-analysis that assumes homogenous effect sizes (Viechtbauer, 2005), since the assumption of homogeneity often does not hold (Schmidt, Oh, & Hayes, 2009).

Different tests and measures are available to assess and test for (statistical) heterogeneity. In the random-effects model, heterogeneity is represented by parameter τ^2, which is the variance of true effect sizes. Since τ^2 is not comparable across meta-analyses, it is not suitable for describing the impact of heterogeneity on meta-analyses (Higgins & Thompson, 2002). Therefore, other heterogeneity statistics have been proposed such as the I^2 statistic, the H^2 index (Higgins & Thompson, 2002), and the *Q*-statistic that is most commonly used to test the null-hypothesis of no statistical heterogeneity (Cochran, 1954) (i.e., H_0: $\tau^2 = 0$). The present paper examines how statistical properties of these heterogeneity statistics are affected by publication bias. We define *publication bias* as the selective publication of studies with a statistically significant outcome (e.g., van Assen, van Aert & Wicherts, 2015), resulting in the overrepresentation in the literature of studies with significant outcomes compared to studies with null results.

The evidence for publication bias is overwhelming, particularly in psychology. Fanelli (2012) showed that about 95% of published articles in psychiatry and psychology contain statistically significant outcomes, and that this percentage has been increasing over the years (cf. Sterling 1959; Sterling et al., 1995). Neither the high percentage nor its increase can be explained by the studies' statistical power, since power is generally low in psychology (e.g., Ellis, 2010; Bakker, van Dijk, & Wicherts, 2012) and has not increased over the years (e.g, Smaldino & McElreath, 2016; Hartgerink, Wicherts & van Assen, 2017). Franco, Malhotra and Simonovits (2014) and Cooper, DeNeve and Charlton (1997) provided direct evidence for publication bias in psychology and related fields. Further evidence of publication bias in psychology was obtained by the Reproducibility Project Psychology, which replicated 100 key effects from articles published in three prominent and high-impact psychology journals (Open Science Collaboration, 2015). Whereas 97% of the original studies reported that the key effect was statistically significant, only 36% of the replicated effects were statistically significant (Open Science Collaboration, 2015), even though the statistical power of the replication studies was generally higher than that of the original studies. This indicates that published effect sizes are often overestimated and that significant results are indeed overrepresented in the literature.

Publication bias affects estimates of effect size and between-study variance in meta-analysis (e.g., Jackson, 2007; Lane & Dunlap, 1978). Since evidence of publication bias is overwhelming in psychology (e.g., Fanelli, 2012; Sterling, Rosenbaum, & Weinkam, 1995), we agree with Jackson that publication bias is "the greatest threat to the validity of a meta-analysis" (2006a, p. 2911). It is well-known that publication bias results in overestimated effect sizes, and that more overestimation is associated with smaller true effect sizes, smaller study sample sizes, and stronger publication bias (e.g., Borenstein, et al., 2009; Nuijten, van Assen, Veldkamp & Wicherts, 2015; van Assen et al., 2015). Interestingly, overestimation by publication bias is unaffected by the number of studies in the meta-analysis (e.g., Nuijten et al., 2015). Hence, including more studies in a meta-analysis does not remedy overestimation when publication bias exists, and may even provide a false sense of confidence since the precision of the (over)estimated effect size increases (cf. van Assen et al., 2015).

While the literature discussed above clearly identifies the effect of publication bias on the mean effect size in a meta-analysis, the effect of publication bias on estimates of the between-study variance and tests of homogeneity are less clear-cut (Thorlund et al., 2012). Sterne and Egger (2005) argued that it is implausible that underdispersion (underestimated between-study variance) will arise other than by chance. On the other hand, Jackson (2006a; 2007) demonstrated that the estimate of heterogeneity depends on the amount of publication bias and the true effect size, and can be either smaller or larger than the true amount of heterogeneity.

As an example of how publication bias can affect heterogeneity, consider Figure 1 showing the sampling distribution of effect sizes of studies, with the same sample size, and an average true effect size equal to 0 (δ=0). The vertical 'CV' line indicates the critical value of statistical significance. When there is full publication bias (i.e. only statistically significant studies get published), only the grey area on the right remains. The variance of this grey area is only 15% of that of the original distribution, showing that when only statistically significant studies are

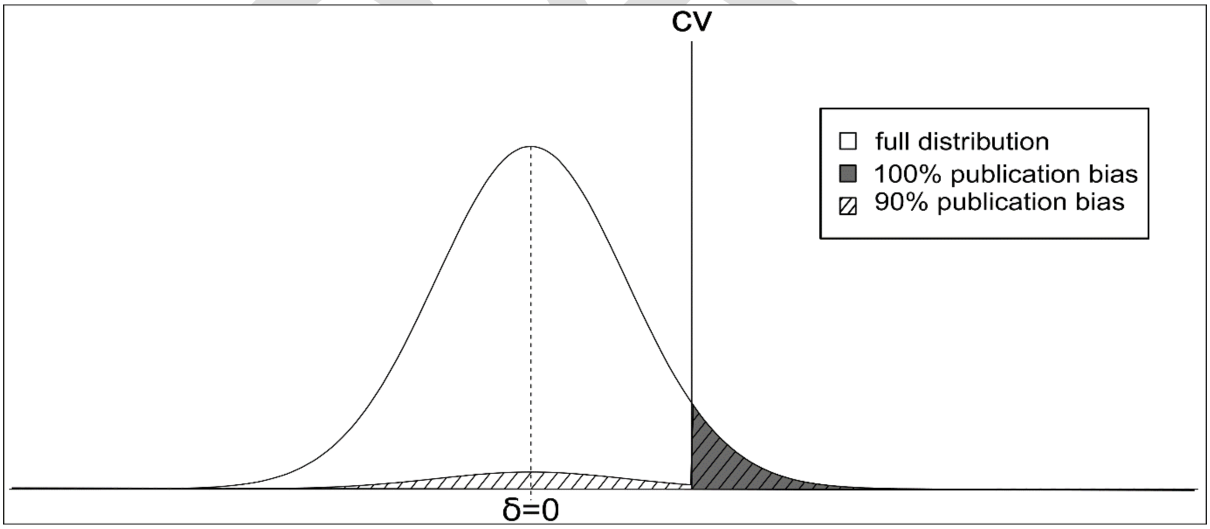

Figure 1. Sampling distribution of effect size with mean true effect size δ = 0 and the critical value (CV) of testing the null hypothesis of no effect. The striped areas correspond to the sampling distribution of published studies when only 10% of insignificant studies get published (bias 90%), and the grey area corresponds to the sampling distribution of statistically significant studies (bias 100%).

published, we underestimate the amount of variation. The striped area shows a different scenario. Here all statistically significant studies are published, as well as 10% of the non-significant studies (90% publication bias). Here, the variance is 1.7 times larger than the variance of the original (full) distribution, due to the relatively large amount of outliers in the right (grey) area.

Because publication bias can lead to either overestimation or underestimation of the between-study variance, Jackson (2006a; 2007) concluded that it is impossible to make generalizations about the implications of publication bias for estimating the between study variance. In his analyses, however, Jackson (2006a; 2007) assumed that studies estimating an effect size larger than a cut-off point are published. Empirical evidence, however, suggests that publication bias based on studies' statistical significance is more realistic, since statistically significant results are over-represented in the literature. Moreover, whereas Jackson only considered meta-analyses with studies with identical sample sizes in his analyses, it is important to study the effects under varying study sample sizes. Similarly, McShane, Böckenholt, and Hansen (2016) recently used simulations to examine the effects of publication bias on heterogeneity assessments for a very limited number of conditions, using two levels of publication bias, five levels of the effect size, five levels of the included number of studies (K) and one level of heterogeneity (τ). These conditions do not, and cannot, capture the complexity of the effects of publication bias on heterogeneity assessment; as they examined the effects of three continuous factors (publication bias, effect size and number of studies) with just 2×5×5 = 50 conditions, and ignored three other important factors (true heterogeneity, study sample sizes, and heterogeneity of study sample sizes), the complex and non-linear effects on heterogeneity assessment as well as the precise conditions under which heterogeneity estimates are (un)biased remain unclear. Therefore, their conclusion "The standard meta-analytic approach tends to underestimate heterogeneity on average when there is selection; however, this result is not uniform, and indeed it sometimes demonstrates an upward bias."(p. 742) is too general and imprecise.

The present study has two main contributions and is relevant to all researchers conducting meta-analyses. First, we show when and how publication bias affects the assessment of heterogeneity. More specifically, we analyze the expected value of the statistics Q, H^2, and I^2, as a function of heterogeneity in true effect size, publication bias, average true effect size, study sample size and variation of sample size. Compared to previous literature (Jackson 2006a, 2007; McShane et al., 2016), our study provides results for more realistic (based on significance rather than effect size), more complex (varying sample sizes within a meta-analysis), and more diverse conditions, on more outcome variables (not only bias of the estimate of the between-study variance, but type I error and statistical power as well). Our results on meta-analyses based on studies with the same sample size are complete, that is, they incorporate any true effect size, heterogeneity, publication bias, and generalize to any sample size and number of studies. We also examine the effects of publication bias in conditions with different sample size variability, both in analytic and Monte Carlo simulation studies. As opposed to previous studies, we also investigate the effect of publication bias on the type I error rate and statistical power of the Q-test in many conditions.

In our analyses, we assume that within-study variances are known, which is a common assumption in meta-analysis (e.g., Raudenbush, 2009). In our main analyses we also assume that average true effect size is known, which allows us to analytically derive our results on heterogeneity, thereby bypassing the problem of estimating heterogeneity as is needed in simulations. The problem of estimation is that more than a dozen of estimators of heterogeneity exist, while none of them performs well in all conditions, and there is limited information which estimator performs best in which condition (Veroniki et al., 2015; Langan, Higgins, & Simmonds, 2016). Bypassing the estimation problem therefore enables us to explore the "pure" effect of publication bias on the assessment of heterogeneity in a published set of studies. The trade-off of this analytic approach is that it requires stricter assumptions that may affect the results. Therefore, we also carried out four simulation studies with less strict assumptions that do estimate the true effect size, instead of assuming it to be known. As the analytic results are more precise (i.e., there is no sampling error in analytic results) and the four simulation studies provide results similar to our main analyses, we only briefly describe the results of our four simulation studies in our paper (details can be found in Supplementary Materials A).

Our second main contribution is to introduce a web application, Q-sense, which can be used by meta-analysts to determine the sensitivity of the *Q*-test and assessment of heterogeneity to publication bias. Q-sense will increase researchers' understanding of their meta-analytic results, since the effect of publication bias is complex and non-linear. Q-sense allows researchers to determine the degree to which their meta-analytic estimate of heterogeneity may be affected by publication bias, thereby examining the robustness of their meta-analytic estimates to publication bias. We illustrate the value of Q-sense in checking robustness and sensitivity to publication bias by applying it to two published meta-analyses.

In the next section, we describe the random-effects meta-analysis model, define the Q, H^2, and I^2 statistics, show how they are related, and derive their distributions when all studies in the meta-analysis have the same sample size. We then discuss the assumptions of our analyses and the conditions that we examine, followed by four sections presenting the effects of publication bias on heterogeneity. The first section presents analytic results on the effect of publication bias on the expected value of statistics Q, H^2, and I^2. The second section focusses on *extreme homogeneity,* when there is less variation in study results than would be expected by chance. We discuss the circumstance where publication bias can create extreme homogeneity and the minimal amount of true heterogeneity needed to be able to detect heterogeneity in those conditions. The third section presents the effect of publication bias on the statistical properties (Type I error rate and power) of the *Q*-test. The fourth section shows the impact of publication bias in additional conditions using sample sizes from the field of psychology. After these four results sections, we describe the web-application Q-sense that allows researchers to determine the sensitivity of the meta-analytic estimate of heterogeneity to publication bias, and apply Q-sense to two published meta-analyses from psychology.

Assessing heterogeneity in a random-effects meta-analysis

The random-effects model assumes that the observed effect size of a study (Y_i) is the result of the average true effect size (μ), the deviation of the study's true effect size from the grand mean (ζ_i) and the study's sampling error (ε_i):

$$Y_i = \mu + \zeta_i + \varepsilon_i \tag{1}$$

Both ζ_i and ε_i are assumed to be normally distributed variables with variances τ^2 and σ_i^2, respectively. σ_i^2 is a function of population variance σ_y^2 and study sample size, for instance σ_y^2 / N_i in case of estimating one group mean. σ_i^2 is estimated in practice and then assumed to be known in meta-analysis. Below we also use σ^2 for σ_i^2 if all studies are based on the same sample size, thereby assuming that they have a common within-study variance.

The variance of true effect sizes τ^2 is independent of the number of studies in the meta-analysis (*K*) and the studies' precision (which is the inverse of σ_i^2). However, τ^2 is not suitable for describing the impact of heterogeneity on the meta-analysis, its conclusions and its interpretation (Higgins & Thompson, 2002). Table 1 shows the statistical properties of τ^2 and the other heterogeneity measures.

Table 1.

Properties of statistics of heterogeneity

Measure	Range	Increasing in K	Decreasing in σ^2	Can assess extreme homogeneity
τ, τ^2	0-∞	No	No	No
Q	0-∞	Yes	Yes*	Yes
I^2	0-1	No	Yes	No
H,H^2	0-∞	No	Yes	Yes

Note. Adapted from Undue reliance of I^2 in assessing heterogeneity may mislead, by G. Rücker, G. Schwarzer, J.R. Carpenter, and M. Schumacher, 2008, BMC Medical Research Methodology, 8, 79.

* "Yes" if the null hypothesis of homogeneity is rejected; if the null hypothesis of homogeneity is not rejected, the estimate does not increase with precision.

The I^2 statistic (Higgins & Thompson, 2002; Higgins, Thompson, Deeks, & Altman, 2003) is often used to interpret the extent of heterogeneity and is defined as

$$I^2 = \frac{\tau^2}{\tau^2+\sigma^2}. \tag{2}$$

The I^2 statistic is one of the most popular heterogeneity statistics because of its ease of interpretation; it represents the proportion of the estimated variance that is due to differences in true effect sizes. Its value is scale invariant (allowing comparison across meta-analyses) and does not depend on the number of studies in a meta-analysis. However, its value increases as the studies' sample size increases (see Table 1). I^2 ranges from 0 (homogeneity; $\tau^2 = 0$) to 1 (heterogeneity and infinite precision; $\tau^2 > 0$ and $\sigma^2 = 0$).

A useful but less popular heterogeneity statistic is H^2 (Higgins & Thompson, 2002):

$$H^2 = \frac{\text{var}(Y_i)}{\sigma^2} = \frac{\tau^2+\sigma^2}{\sigma^2} = \frac{1}{1-I^2} \tag{3}$$

Like I^2, H^2 is independent of the number of studies but dependent on studies' precision. However, H^2 has an advantage over I^2 and τ^2 in that it can detect extreme homogeneity (H^2

<1). Following Ioannidis, Trikalinos, and Zintzaras (2006), evidence of extreme homogeneity is obtained when there is significantly less variance of study effect sizes than would be expected under conditions of homogeneity (i.e., when τ^2=0). At the analytic or population level we have evidence of extreme homogeneity when the expected value of H^2 is smaller than 1.Hence H^2 ranges from 0 to infinity, with values lower than 1 signalling extreme homogeneity, 1 corresponding to homogeneity, and values larger than 1 corresponding to heterogeneity (see Table 1).

The well-known Q-statistic is most commonly used to test the null-hypothesis of no statistical heterogeneity (i.e., H_0: $\tau^2 = 0$). The DerSimonian and Laird estimator for the between-study variance in true effect size is based upon this Q-statistic (DerSimonian & Laird, 1986). It is defined as the squared sum of standardized effect sizes:

$$Q = \sum_{i=1}^{K} \left(\frac{Y_i - \mu}{\sigma_i}\right)^2 \qquad (4)$$

Under the assumptions of the fixed-effects model, the null-hypothesis of homogeneity ($\tau^2 = 0$), and assuming a known true effect size, the Q-statistic follows a central χ^2 distribution with degrees of freedom equal to K and Q has an expected value equal to K and variance $2K$. In practice, the true effect size is unknown, and the distribution has K-1 degrees of freedom. The value of Q increases with K, and when the null-hypothesis is false, Q also increases with increases in studies' precision. Q ranges from 0 to infinity, with expected values of Q between 0 and the degrees of freedom corresponding to extreme homogeneity, expected values of Q equal to the degrees of freedom corresponding to homogeneity, and expected values of Q larger than the degrees of freedom corresponding to heterogeneity (Table 1). The hypothesis of homogeneity is commonly rejected when the observed value of Q exceeds the 90th or 95th percentile of the central χ^2 distribution (Higgins, et al., 2002, 2003). We note that the Q-test has low power when there are few studies in a meta-analysis, and that one should not rely on its result for diagnosing heterogeneity in such situations (e.g., Higgins et al., 2003).

The expected values of statistics *Q*, *H*^{2}, and *I*^{2} are related under statistical heterogeneity and equal study sample sizes. For simplicity, assume all studies assess one population mean μ and have the same population varianceσ_y^2. Then, it can be shown that (Jackson, 2006b):

$$Q = \sum_{i=1}^{K} \left(\frac{Y_i - \mu}{\sigma_i}\right)^2 \sim \left(1 + \frac{N\tau^2}{\sigma_y^2}\right)\chi^2(K) = \frac{1}{1-I^2}\chi^2(K) = H^2\chi^2(K) \tag{5}$$

In words, *Q* follows a central χ^2 distribution with *K* degrees of freedom multiplied by a constant ($\frac{1}{1-I^2}$ or *H*^{2}). This constant equals 1 plus precision × true heterogeneity, which reflects that under heterogeneity *Q* increases with studies' precision and true heterogeneity (Table 1). This equation also reveals that the distribution of *Q* shrinks under extreme homogeneity ($H^2 < 1$) and inflates under heterogeneity ($H^2 > 1$). Using (5), *I*^{2} can be redefined as

$$I^2 = \frac{E(Q) - \mathrm{K}}{E(Q)} \tag{6}$$

whenever E(*Q*) > *K* (otherwise $I^2 = 0$), because

$$I^2 = \frac{E(Q) - \mathrm{K}}{E(\mathrm{Q})} = \frac{K\left(1 + \tau^2/\sigma^2\right) - K}{K(1 + \tau^2/\sigma^2)} = \frac{\tau^2}{\sigma^2 + \tau^2},$$

and *H*^{2} can be redefined as

$$H^2 = \frac{\mathrm{E}(Q)}{K} \tag{7}$$

because

$$H^2 = \frac{E(Q)}{K} = \frac{K\left(1 + \tau^2/\sigma^2\right)}{K} = \frac{\tau^2 + \sigma^2}{\sigma^2}.$$

In our analyses below, we use equations (6) and (7) to derive our results for *I*^{2} and *H*^{2}, under different conditions of true heterogeneity, publication bias, average true effect size, number of studies, study sample size, and variation of sample sizes.

Analyses

Overall design

In our analyses, we assume that all primary studies result from a normal distribution with mean

= μ and known standard deviation σ_y = 1. Our results are derived for effect size measure Cohen's *d*, denoted by δ because we analyze populations. We assume all studies i = 1,…, K examine one population mean with sample size N_i, but our results also hold exactly for a balanced independent two-samples design, comparing two population means, where each group has $2N_i$ observations (i.e., a total sample size of $4N_i$). All code used for generating data and analysis can be found in our Supplementary materials on OSF (https://osf.io/qzt5z/)

Statistical significance of a study is determined using a one-sided test with α = .05. We model publication bias with one parameter *pub* between 0 and 1 representing the relative reduction in the probability of statistically nonsignificant studies getting published compared to statistically significant studies. For example, *pub* = .2 indicates that the probability of getting published is five times higher for significant than for nonsignificant studies. Recently developed meta-analytic methods that attempt to adjust for publication bias employ the same model (e.g., Simonsohn, et al., 2014a; 2014b; van Assen, et al., 2015). Our model is also comparable with selection models determining the weight of a study as a function of the *p*-value, such as the combined probability model (Iyengar and Greenhouse, 1988; Hedges, 1992; Hedges, & Vevea, 2005).

Independent factors

In order to study the effect of publication bias on heterogeneity, we systematically varied the following factors: sample size variation, number of studies, true heterogeneity, true effect size and the amount of publication bias. Differences in sample sizes in our main analyses were manipulated into three different levels; equal sample size (1:1), a small difference in sample sizes (1:3), and a larger difference in sample sizes (1:10). We assumed 20% and 80% of large and small published studies, respectively, since more small studies are published within the psychology literature. The expected value of Q then equals

$$\sum_{i=1}^{K} N_i E\,(Y_i - \mu)^2 = \frac{K}{5}(1 + N_L\tau^2) + \frac{4K}{5}(1 + N_S\tau^2) = K + K\left(\frac{1}{5}N_L + \frac{4}{5}N_S\right)\tau^2,$$

with subscripts S and L referring to small and large sample sizes, respectively. To obtain the same expected value of Q (and H^2 and I^2) for varying sample sizes as for equal sample sizes, the total sample sizes across all studies in the meta-analysis should be equal. We chose a total sample size that was divisible by 5 (for the 1:1 ratio: 4 × 1 + 1 × 1), 7 (for the 1:3 ratio: 4 × 1 + 1 × 3 = 7), and 14 (for the 1:10 ratio: 4 × 1 + 1 × 10 = 14), and would result in realistic sample sizes for each study. We selected a total sample equal to 210, which corresponds to sample sizes of 42 (5 × 42 = 210) in case of equal (1:1) sample sizes, 30 and 90 for slightly varying (1:3) sample sizes (4 × 30 + 1 × 90 = 210), and 15 and 150 (4 × 15 + 1 × 150 = 210). As was noted before, our results are equivalent to a balanced independent two-sample design. When comparing two groups, our results reflect and group sample sizes twice as large (i.e., 84 per group, 60 or 180 per group, and 30 or 300 per group), and thus total sample sizes that are four times as large (840). These sample sizes are somewhat larger than commonly used in the psychological literature (Hartgerink, Wicherts, van Assen, 2017), but are the smallest sample size values that satisfy all our sample size constraints (keeping total sample sizes fixed, given certain ratios of small and large studies). Using slightly larger sample sizes will not affect the patterns of relationships we study, but only compresses or squeezes the *y*-axis of the figures below. More precisely, increasing all sample sizes with factor *C* will squeeze the *y*-axis of figures 3 till 6 with a factor $1/\sqrt{C}$.

While arbitrary and unrealistic, our choice of two different sample sizes in our main analyses enabled us to analytically derive the results of publication bias on heterogeneity. Because the results of the effect of sample size variability may depend on its specific implementation, we carried out four additional Monte Carlo simulation studies with other and more realistic implementations of sample size variability. One study was almost identical to the design of our main analyses, with the same total sample size and same sample size variability, but now with five different sample sizes (6, 30, 42, 54, 78) rather than two (30, 90). Comparing the results of this study with a simulation study corresponding to our main analyses allowed us to directly examine if the distribution of sample size affects the results, given the same mean and variance of this distribution. As the distribution of sample sizes in this additional study may still be considered unrealistic, two other simulation studies used sample size distributions of psychological research. One of them randomly selected sample sizes from studies in the field of social and personality psychology (Fraley & Vazire, 2014), whereas the other study randomly selected sample sizes from studies in one large meta-analysis on the association between brain volume and intelligence (Pietschnig, Penke, Wicherts, Zeiler, & Voracek, 2015). As the trends in the Monte Carlo simulation studies and conclusions based on these studies are similar to those of our main analyses, we only briefly summarize their results later on. All details of the simulation studies are described in Supplementary Materials A.

Two other factors that varied in our analyses are true heterogeneity and the number of studies in the meta-analysis. True heterogeneity, assessed with I^2, was varied from 0 (homogeneity), to small (.25), medium (.5), and large (.75), with values based on the rules of thumb by Higgins and Thompson (2002). Using these I^2 values, we calculated the number of studies (K) for which the power of the Q-test equals .80 in case of equal sample sizes. We obtained values of K yielding a power of .8 by solving

$$P\left(Q \geq \chi^2_{cv}(K)\right) = P\left(\frac{\chi^2(K)}{1 - I^2} \geq \chi^2_{cv}(K)\right) = P\left(\chi^2(K) \geq (1 - I^2)\chi^2_{cv}(K)\right) = .8$$

for K, assuming $I^2 = .25$, .50 and .75, $\alpha = .05$, and that K is a multitude of five published studies. These criteria yielded $K = 145$ and $I^2 = .2504$, $K = 25$ and $I^2 = .4970$, $K = 5$ and $I^2 = .7884$ (see Supplementary Material B). Note that the number of studies is relevant for our analyses of the power of the Q-test and the expected value of Q, but not for analyses of the expected value of I^2 and H^2 (see Table 1).

The true effect size δ varied from 0 to 1, representing a range between null and very large effects. For δ = 1, almost all individual studies are statistically significant and get published, and publication bias has no effect. More specifically, the probability of a significant effect equals .9871, .99994 and .9999993, for sample sizes 15, 30 and 42 respectively.

In our analyses, we varied both δ and *pub* (i.e. the probability of publication of non-significant studies relative to significant studies) in steps of .01, creating a grid of 101×101 = 10,201 combinations. For each of the 4 (heterogeneity of true effect size) × 3 (variation of sample size) = 12 conditions, we computed the expected heterogeneity (I^2 and H^2) in the grids. For 3 (number of studies) × 3 (variation of sample size) = 9 conditions, we computed type I error rate and statistical power of the Q-test in grids. We now describe the dependent variables corresponding to our three research questions in more detail.

Outcome measures

Expected values of I² and H². In these analyses, the dependent variable I^2 is used when E(I^2) ≥ 0, and E(H^2) if E(H^2) < 1 (extreme homogeneity, where I^2 is not defined). These expected values were calculated from the expected value of Q and equations (6) and (7). Working out the expected value of Q assuming varying sample sizes and publication bias yields

$$E(Q) = K\left[P_S \mathrm{H}_S^2 + P_L \mathrm{H}_L^2 + \frac{P_S N_S P_L N_L}{P_S N_S + P_L N_L}(\mu_S - \mu_L)^2\right] = K\left[\mathrm{H}_{avg}^2 + \mathrm{H}_{extra}^2\right] \qquad (8)$$

P_S and P_L refer to the proportion of small and large published studies, which are .2 and .8 respectively. H_S^2 and H_L^2 refer to the expected value of H^2 for small and large studies, respectively, which combine with their corresponding proportions into the weighted average of H^2, i.e., H_{avg}^2. As can be seen in equation (8), H_{extra}^2 is a function of the proportion of small and large published studies, their sample sizes (N_S and N_L) and the squared difference of the means of the populations of published small and large studies. In case of equal sample sizes, $H_{extra}^2 = 0$ because then $\mu_S = \mu_L$. The expected value of a population of published studies μ is calculated as

$$\mu = \frac{pub(1-\pi)\mu_{nonsig}+\pi\mu_{sig}}{pub(1-\pi)+\pi}, \tag{9}$$

with π denoting the statistical power of rejecting the null-hypothesis in one study, μ_{nonsig} denoting the expected effect size of nonsignificant studies and μ_{sig} denoting the expected effect size of significant studies. μ is calculated by integrating the effect size distributions with mean δ and standard deviation $\sqrt{}(\sigma_y^2\ /\ N_i + \tau\text{^}2\ )$ from minus infinity to the critical value (1.645 ×$\sqrt{\sigma_y^2/N_i}$) for nonsignificant studies and from this critical value to infinity for significant studies. H_S^2 and H_L^2 were calculated as

$$H^2 = \mathrm{N}_i\left(\frac{pub(1-\pi)E\left(Y^2\right)_{nonsig}+\pi E\left(Y^2\right)_{sig}}{pub(1-\pi)+\pi} - \mu^2\right), \tag{10}$$

which is the variance of the distribution of published standardized effect sizes, where E(Y^2) is obtained by integrating the effect size distribution from minus infinity to the critical value, and from the critical value to infinity for nonsignificant and significant studies, respectively (see Supplementary Material C) .

True heterogeneity I^2 required to obtain homogeneity. Extreme homogeneity (i.e., E(Q) < K and E(H^2) < 1) is obtained for many combinations of values of true effect size and publication bias. For these combinations, we use equations (8)-(10) to calculate the true heterogeneity in the population of all studies (published and unpublished) required to obtain homogeneity (i.e., E(Q) = K and E(H^2) = 1). We used the iterative bisection method (Adams & Essex, 2013, pp. 85-86) for τ^2 on interval [0, 0.2][1] to obtain E(H^2) = 1, with tolerance 1e-6. From this computed

[1] The bisection method for τ^2 on interval [0, 0.2] is guaranteed to find the only solution for τ^2 such that E(H^2) = 1, if (i) E(H^2) < 1 for τ^2 = 0 and (ii) E(H^2) > 1 for τ^2 = .2. Condition (ii) always holds, because when homogeneity is most extreme (equal sample sizes, N_i = 42), τ^2 = .2 results in E(H^2) = 8.538. Hence, we first checked if condition (i) holds before applying the bisection method (if (i) does not hold, no solution exists).

τ^2 we derived the true heterogeneity I^2 of all studies using equation (6) (see Supplementary Material C).

Type I error rate and power of the Q-test. The statistical properties of the *Q*-test of homogeneity were obtained using Monte-Carlo simulations. Type I error rate and power were estimated for $I^2 = 0$ and the three conditions with $I^2 > 0$ (i.e., $I^2 = .2504$ and $K = 145$, $I^2 = .4970$ and $K = 25$, $I^2 = .7884$ and $K = 5$), respectively. Note that these three heterogeneity conditions were chosen such that the power of the *Q*-test exactly equals .8 when the sample sizes are equal ($N = 42$) and there is no publication bias, at $\alpha = .05$. We also varied the sample size ratio as before (1:1, 3:1, 10:1). To obtain a 95% confidence interval of the type I error rate and power with a width of at most .01, we ran $S = 40{,}000$ iterations for each combination of δ and publication bias. The width of the confidence interval equals $2 \times 1.96 \times \sqrt{p\,(1-p)/S}$. This confidence interval is widest when the type I or power (p) is equal to .5, resulting in a confidence interval with a width of .0098. When p = alpha =.05, the width is equal to .0043, and for $p = .8$ it is equal to .00784. The type I error rate and power were estimated by computing the proportion of statistically significant *Q*-tests across the 40,000 iterations. The effect sizes Y_i of the small and large studies were drawn randomly from the distributions of published studies that were also used to compute equations (8)-(10). In each iteration, Q was calculated using

$$Q = \sum_{i=1}^{K} N_i \times (Y_i - \mu)^2$$

with

$$\mu = \frac{4N_S\mu_S + N_L\mu_L}{4N_S + N_L}$$

Because mean μ is known exactly, we compared Q to the 95th percentile of the χ^2-distribution with K degrees of freedom. R-code used to simulate the statistical properties (type I error rate and power) can be found in Supplementary Material D.

Results

Expected values of I^2 and H^2

Interpreting results, particularly when they are as complex as in our case, is simplified when mechanisms producing these results are grasped at least at an intuitive level. Hence, we start by explaining the non-linear effect of publication bias on heterogeneity in a simple example, before presenting and interpreting our twelve grids with complete results on the expected values of I^2 and H^2.

Figure 2 depicts $E(H^2)$ of published studies as a function of publication bias for $\delta = 0$, equal sample sizes, and four levels of heterogeneity (I^2). Each subfigure shows a horizontal dashed line corresponding to true population values of H^2 and I^2 (i.e., of all studies). When all studies get published, E(H^2) equals H^2, which can also be seen at the complete right of each plot in Figure 2. Figure 2a illustrates the case where there is true homogeneity in the population of all studies ($I^2 = 0$, $H^2 = 1$). When only significant studies get published, variance of effect sizes and E(H^2) is only 0.138, showing a large bias (difference in heterogeneity of published studies and heterogeneity of all studies) in heterogeneity. If the '% of non-significant studies published' is increased, E(H^2) quickly increases until it crosses $H^2 = 1$ at *pub* = .03 (no bias in heterogeneity), achieves its maximum for *pub* =.07 (large bias in heterogeneity), and then slowly decreases until $H^2 = 1$ at *pub* = 1 (again no bias in heterogeneity). An intuitive explanation for the fact that the variance of the distribution of published effect sizes at *pub* = .07 exceeds 1, is that the shape of the distribution is similar to a normal distribution with an excessive amount of outliers at the right tail, similar to the striped area in Figure 1. Curves for other levels of heterogeneity may also show a nonlinear effect of publication bias on heterogeneity of published studies, although patterns may be different (e.g., maximum is achieved at different values of *pub*).

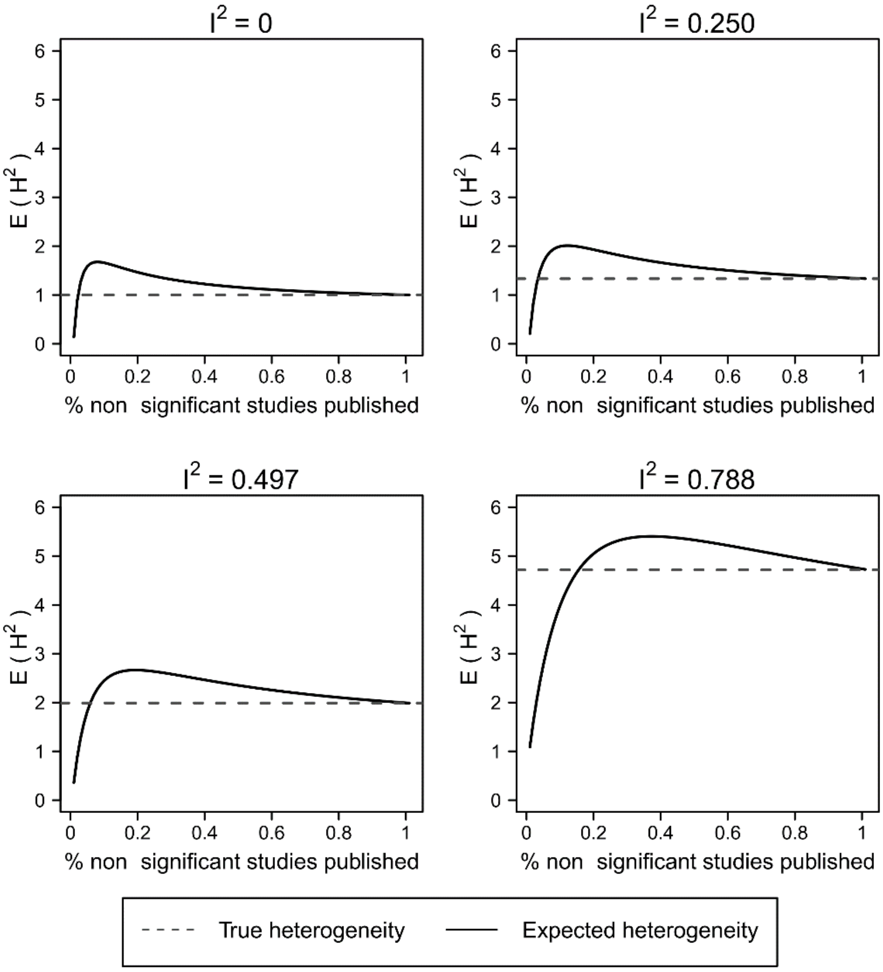

Figure 2. Expected values of H^2 (Y-axis), when δ = 0, as a function of publication bias. (X-axis, 0 and 1 correspond to none and all non-significant studies published, respectively), for four levels of heterogeneity (I^2 is zero (a), small (b), medium (c) and large (d)). The horizontal grey lines show the true heterogeneity population values (in H^2).

Figure 3 presents the expected values of H^2 and I^2 for the 4 (heterogeneity; in columns) x 3 (sample size ratios; rows) conditions. Each plot shows the value of I^2 whenever its larger than 0 (solid iso-contour lines), and E(H^2) otherwise (dotted iso-contour lines). Note that the results on the *x*-axis (i.e. when δ = 0) of the four plots in the first row of Figure 3 were already depicted in Figure 2. All results are independent of the number of studies in the meta-analyses. Importantly, the plots in the first column of Figure 3 can be generalized to any sample sizes with the same ratio (1:1, 1:3 and 1:10) because the same plot is obtained for sample sizes multiplied by X and the y-axis replaced by $\delta/\sqrt{X}$.

The upper left plot corresponds to homogenous true effect size and equal sample sizes. The expected value equals the true value ($I^2 = 0$) (no bias in heterogeneity) when there is no publication bias ($pub = 1$), and at the line running from ($\delta = 0$, $pub = .03$) to ($\delta = .25$, $pub = .99$). Above the $I^2 = 0$ line there is extreme homogeneity. This is the case for the majority of grid values (80%), and the least variation ($E(H^2) = .138$) is observed when $\delta = 0$ and $pub = 0$. Below the I^2=0 line we have heterogeneity, with a maximum of $E(I^2) = 0.403$ ($\delta = 0$, $pub = .07$), which corresponds to small to moderate heterogeneity. To conclude, in case of a fixed population effect size, publication bias can result in a wide range of published effect size distributions that vary from being overly homogenous (an underestimation of heterogeneity) to moderate heterogeneity (an overestimation of heterogeneity).

The other plots on the first row of Figure 3 show $E(H^2)$ and $E(I^2)$ when there is true heterogeneity for equal sample sizes. The patterns of results are the same as for $I^2 = 0$, with underestimation of heterogeneity above and overestimation below the ‘true’ I^2 line, respectively. Noteworthy is that for small heterogeneity ($I^2 = .25$) and medium heterogeneity ($I^2 = .497$) publication bias can still result in extremely homogenous distributions of published effect sizes ($E(H^2) = .208$ and $E(H^2) = .359$, see also the legend below the corresponding plots). Even when there truly is strong heterogeneity ($I^2 = .788$), publication bias may result in a distribution of published effect sizes that has very small heterogeneity ($E(I^2) = .09$). Maximum expected heterogeneity, $E(I^2)$ of published studies equals .625, .788, .815 for true small, medium and large heterogeneity, respectively, corresponding to overestimation of heterogeneity.

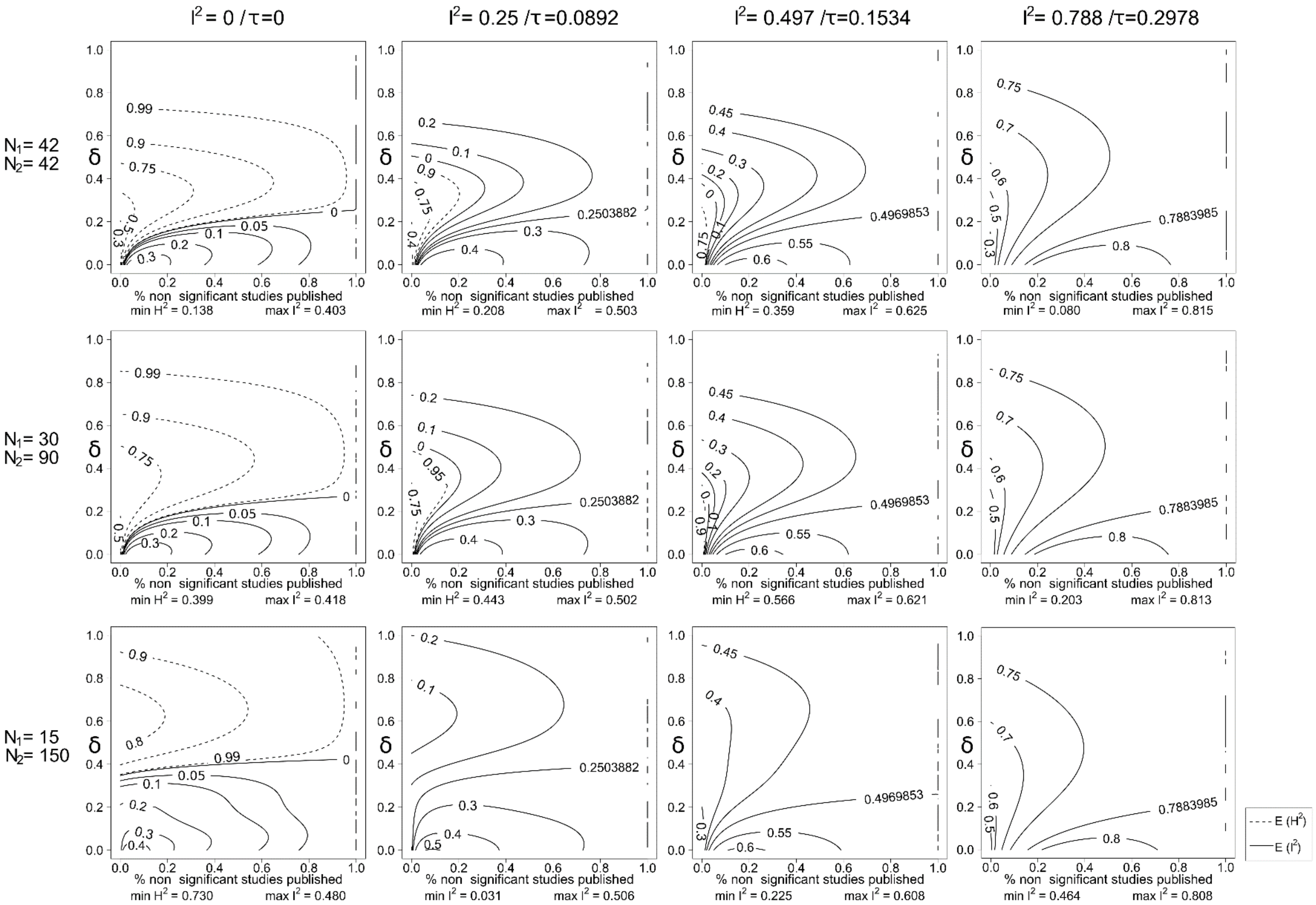
I^2= 0 / τ=0
I^2= 0.25 /τ =0.0892
I^2= 0.497 /τ=0.1534
I^2= 0.788 /τ=0.2978
N_1= 42
N_2= 42
N_1= 30
N_2= 90
N_1= 15
N_2= 150
δ
% non significant studies published
0.0
0.2
0.4
0.6
0.8
1.0
0.99
0.9
0.75
0.5
0.3
0
0.05
0.1
0.2
0.8
0.95
0.4
0.2503882
0.45
0.4969853
0.55
0.6
0.7
0.7883985
min H^2 = 0.138
max I^2 = 0.403
min H^2 = 0.208
max I^2 = 0.503
min H^2 = 0.359
max I^2 = 0.625
min I^2 = 0.080
max I^2 = 0.815
min H^2 = 0.399
max I^2 = 0.418
min H^2 = 0.443
max I^2 = 0.502
min H^2 = 0.566
max I^2 = 0.621
min I^2 = 0.203
max I^2 = 0.813
min H^2 = 0.730
max I^2 = 0.480
min I^2 = 0.031
max I^2 = 0.506
min I^2 = 0.225
max I^2 = 0.608
min I^2 = 0.464
max I^2 = 0.808
E (H^2)
E (I^2)

Figure 3. Contour plots of expected values of H^2 and I^2 for different values of true heterogeneity (columns) and different sample size ratios (rows), as a function publication bias (X-axis: 0 and 1 correspond to none and all non-significant studies published, respectively) and effect size (δ, Y-axis). The text above each column gives the values of true heterogeneity, while the text before each row gives the studies sample sizes. The text below each plot gives the minimum E(I^2) (or E(H^2) if E(I^2) <0) and maximum E(I^2) in the plot

The plots in the first column of Figure 3 show the results for homogeneity and different variations in sample sizes (1:1, 1:3 and 1:10). Similar to the results for equal sample sizes, extreme homogeneity exists above the $I^2 = 0$ line and heterogeneity exists below the $I^2 = 0$ line. The most important difference is that the minimum E(H^2) and maximum E(I^2) increase when increasing in sample size variability. More generally, extreme homogeneity occurs less often, while larger variability in sample sizes across studies resulting in a higher probability of overestimating heterogeneity. However, small differences in sample sizes (1:3) do not result in very different results compared to equal sample sizes. The fact that the effect of publication bias on heterogeneity depends on variation of sample size can be explained by equation (8); the difference in average true effect size of published studies increases when there is greater variability in sample size, resulting in positive values of H^2_{extra}, thereby increasing heterogeneity of published studies. The largest value of expected heterogeneity of published studies under true homogeneity is E(I^2) = .48 for sample size ratio 1:10, $\delta = 0$, and *pub* = .04 (lower left plot). The other plots in the second and third row show the conditions under which both true heterogeneity and varying sample sizes are present. The effect of true heterogeneity in case of varying sample sizes is comparable to its effect in case of equal sample sizes. However, bias in heterogeneity gets lower for higher true heterogeneity and large differences in sample sizes; bias is lowest in the lower right plots (fewest iso-lines). This trend is also apparent when comparing the minimum and maximum heterogeneity across conditions (see legends below the plots). But even in the condition with the least bias (sample ratio 1:10 and $I^2 = .788$), publication bias may result in much lower heterogeneity (minimum E(I^2) = .464, for $\delta = 0.00$ and *pub*=0.00)

True heterogeneity required to obtain homogeneity of published studies

For many combinations of δ and publication bias under true homogeneity, the value of E(H^2) indicated extreme homogeneity (E(H^2) < 1; upper left plot of Figure 3). For these combinations, one may wonder how much true heterogeneity is needed to obtain E(H^2) = 1. For these or lower values of heterogeneity, the hypothesis of $\tau^2 = 0$ will be infrequently rejected for any number of studies, because E(Q) ≤ K < Q_{CV}, with Q_{CV} denoting the critical value of the Q-test. That is, these are the values of true heterogeneity that because of publication bias will very likely go undetected even when increasing the number of studies.

Figure 4 presents the values of true heterogeneity required to observe homogeneity of published studies (E(Q) = K) for the three different sample size ratios. These results are generalizable to any sample sizes with the same ratios (i.e., the same plots are obtained for sample sizes multiplied by X and the y-axis replaced by δ/√X). High values of heterogeneity are required for some values of *pub* and δ, particularly for equal sample sizes and for ratio 1:3. For instance, for (*pub* = 0, δ = 0) under equal sample size (first plot), a distribution of effect sizes of all studies with large under equal sample size (first plot), a distribution of effect sizes of all studies with large heterogeneity I^2 = .774 will still result in a homogenous distribution of published effect

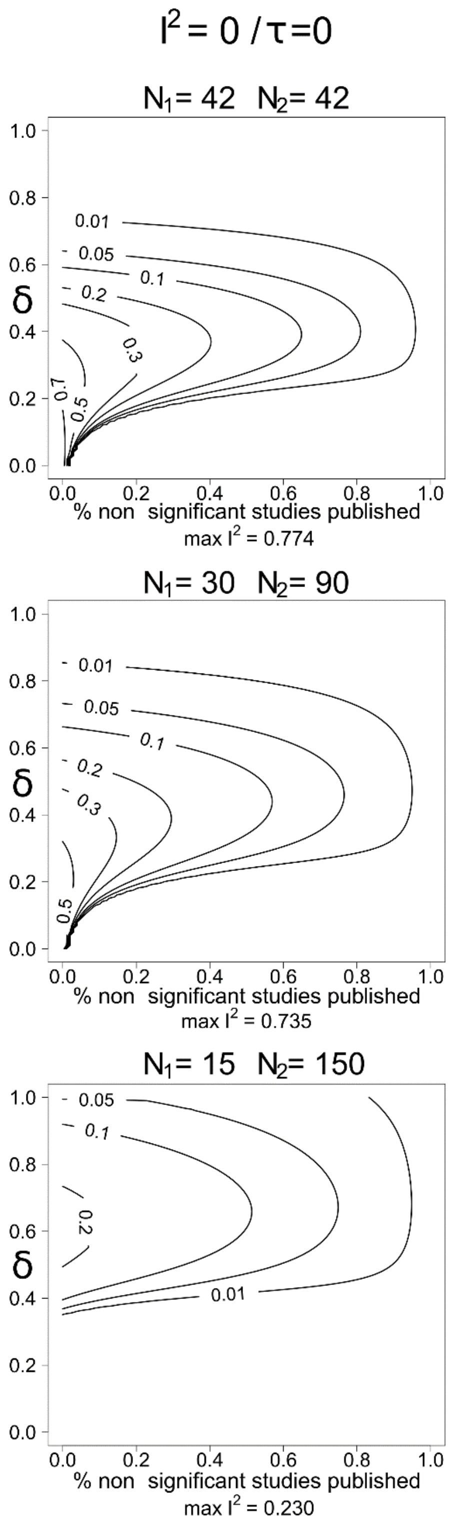

Figure 4. Values of population I^2 needed to obtain homogeneity

sizes. Even if the meta-analysis contained hundreds of studies, large heterogeneity would go undetected by the *Q*-test. For most combinations of effect size and publication bias, the effects of publication bias are less severe.

There are many *pub*-δ combinations, where small true heterogeneity (I^2 = 25%) would go undetected. When sample sizes differ more strongly (1:10), extreme homogeneity decreases (last row Figure 3) and lower values of true heterogeneity are required to obtain homogeneity of published studies (last plot of Figure 4).

Type I error rate and power of the *Q*-test

The first row of Figure 5 presents the Type I error rate as a function of true effect size and publication bias. For equal sample sizes (first row), the distribution of *Q* is known (see equation (5)), and the .05-line corresponds to the E(H^2) = 1-line in the upper left plot of Figure 3. At the top and right edge of each plot, we see Type I error rate clusters of .05 with small deviations in between (deviating from .045 to .055), indicating the Type I error rate approximates .05. The Type I error rate exceeds .05 below the .05 line, while the Type I error rate is lower than .05 above the line. When more studies are included in a meta-analysis, differences in the type I error rate increase. This can be explained by the fact that E(*Q*) and bias in *Q* increase linearly in *K*, whereas the variance of *Q* only increases with $\sqrt{2K}$, resulting in Type I error rates that differ more from .05 as *K* increases. For instance, the maximum Type I error rate equals .217 for (*pub* = .07, δ = 0) and five studies, but increases to .658 and .9999 for 25 and 145 studies, respectively; the minimum Type I error rate equals 0 for all three levels of *K*. For large values of *K*, the Type I error rate quickly converges to 0 or 1 away from the .05-line, and the result of the *Q*-test (rejection of the null-hypothesis of homogenous effect size) is determined by the value of *pub* and δ.

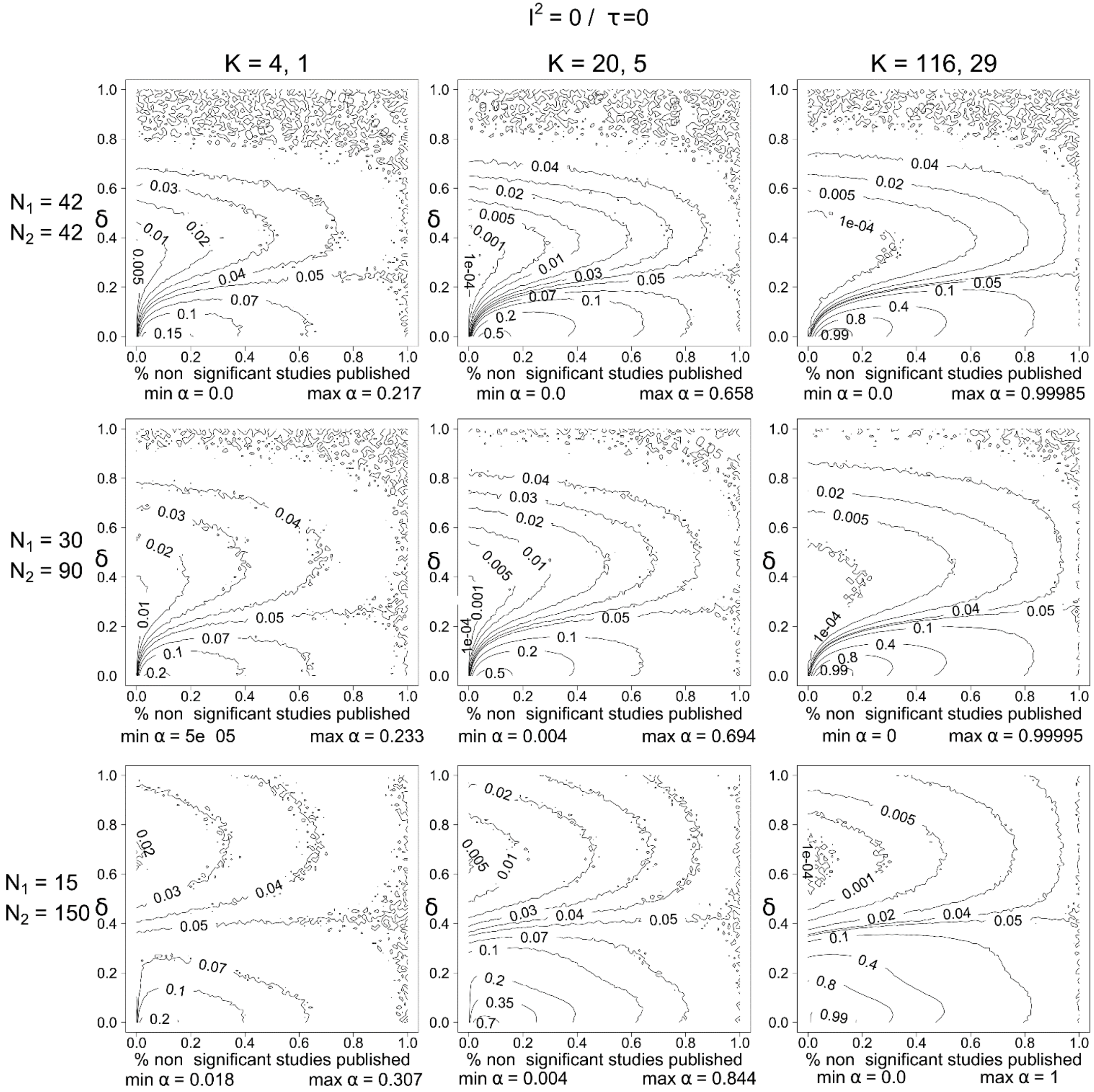
$I^2 = 0$ / τ=0
K = 4, 1
K = 20, 5
K = 116, 29
N_1 = 42
N_2 = 42
N_1 = 30
N_2 = 90
N_1 = 15
N_2 = 150
δ
% non significant studies published
min α = 0.0
max α = 0.217
min α = 0.0
max α = 0.658
min α = 0.0
max α = 0.99985
min α = 5e 05
max α = 0.233
min α = 0.004
max α = 0.694
min α = 0
max α = 0.99995
min α = 0.018
max α = 0.307
min α = 0.004
max α = 0.844
min α = 0.0
max α = 1
0.0
0.2
0.4
0.6
0.8
1.0
0.03
0.01
0.02
0.005
0.04
0.05
0.07
0.1
0.15
0.001
1e-04
0.2
0.5
0.4
0.8
0.99
0.35
0.7

Figure 5. Contour plots of the Type I error rate of the Q-test of homogeneity when $I^2 = 0$, for different sample sizes (columns) and ratios (rows), as a function publication bias (X-axis: 0 and 1 correspond to none and all non-significant studies published, respectively) and effect size, delta (δ, Y-axis). The text below each plot gives the minimum and maximum expected value of α in the plot.

When sample sizes vary, the distribution of the Q-test follows a mixture of two χ^2 distributions, so the .05-line no longer coincides with the E(H^2) = 1 line. Variation in sample sizes causes both the minimum and maximum values of the Type I error rate to increase, compared to equal sample sizes. Otherwise, patterns of results on the Type I error rates are similar to those for equal sample sizes

The results for statistical power of the Q-test are presented in Figure 6. For equal sample sizes, the distribution of Q is known when no publication bias is present. The statistical power of the Q-test is .80, where expected heterogeneity equals the I^2-line in the corresponding plots of Figure 3. Again, increasing K results in larger differences from .8, because bias increases linearly in K, but variance of Q increases only with $\sqrt{2K}$. Note how statistical power can be very low, even when there truly is a large amount of heterogeneity, but only five studies. For a large number of studies, power quickly converges to 0 or 1 away from the .8-line. For instance, when K = 145, power equals 0 at *pub* = 0 and δ = 0 and 1 at *pub* = .07 and δ = 0. To conclude, if heterogeneity is small and the number of studies is large, such that statistical power of the Q-test is .8 in the absence of publication bias, the result of the Q-test is determined by the values of *pub* and δ when there is publication bias.

We mentioned above that Q follows a mixture of two χ^2 distributions when sample sizes vary. This also has consequences for the power of the test: the E(H^2) = 1-line in Figure 3 does not coincide with power = .8-line in the plots in the second and third row of Figure 6. The first column of Figure 6 shows that increasing variation in sample sizes results in three major changes. First, statistical power decreases to approximately .70 when there is no publication bias and sample sizes differ substantially (1:10). Second, when the variation in sample sizes increases, differences in statistical power across different values for *pub* and δ decrease, and third, minimum statistical power increases when there is more variation in sample sizes. These effects of increasing sample size variation can be explained by increases in H^2_{extra} (see equation

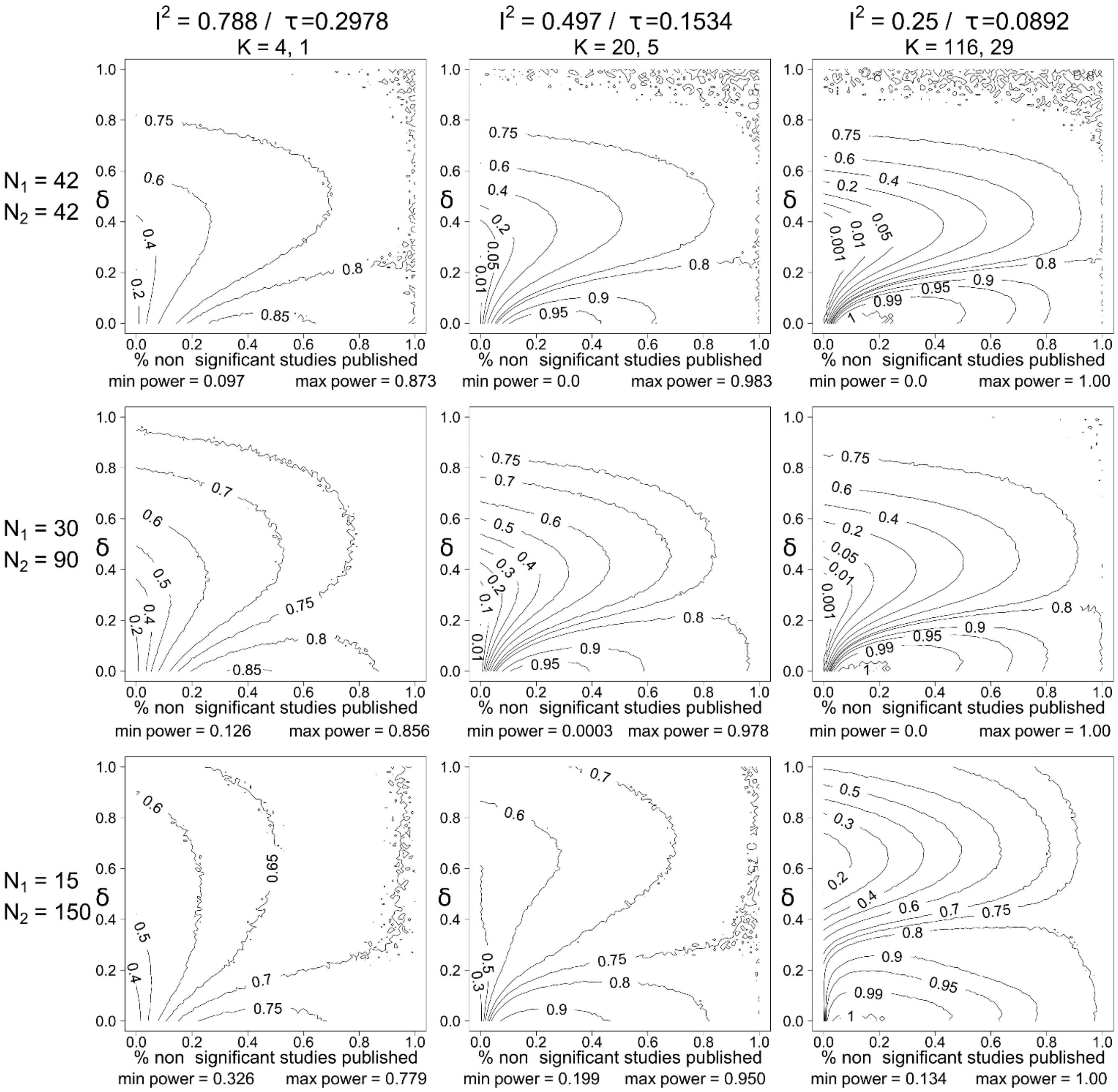
I^2 = 0.788 / τ=0.2978
K = 4, 1
I^2 = 0.497 / τ=0.1534
K = 20, 5
I^2 = 0.25 / τ=0.0892
K = 116, 29
N_1 = 42
N_2 = 42
N_1 = 30
N_2 = 90
N_1 = 15
N_2 = 150
δ
% non significant studies published
0.0
0.2
0.4
0.6
0.8
1.0
min power = 0.097
max power = 0.873
min power = 0.0
max power = 0.983
min power = 0.0
max power = 1.00
min power = 0.126
max power = 0.856
min power = 0.0003
max power = 0.978
min power = 0.0
max power = 1.00
min power = 0.326
max power = 0.779
min power = 0.199
max power = 0.950
min power = 0.134
max power = 1.00
0.75
0.6
0.4
0.2
0.8
0.85
0.05
0.01
0.9
0.95
0.001
0.99
1
0.7
0.5
0.3
0.1
0.65

Figure 6. Contour plots of the power of the Q-test of homogeneity for different amounts of true heterogeneity (columns) and different sample size ratios (rows), as a function publication bias (X-axis: 0 and 1 correspond to none and all nonsignificant studies published, respectively) and effect size (δ, Y-axis). The text below each plot gives the minimum and maximum value of statistical power in the plot

(8)), which, increases the expected value of Q. These three trends resulting from an increasing variation in sample sizes (decrease of power in absence of bias, smaller differences in power due to publication bias and true effect size, larger minimum statistical power) persist through the second and third columns, i.e. for $I^2 = .75$ and K = 5, and $I^2 = .5$ and $K = 25$ and $I^2 = .25$ and $K = 145$.

Results of four simulation studies: Other implementations of sample size variability

We briefly discuss the results on the expected values of H^2 and I^2 based on simulations with 25 studies (K=25). For type I error rates and power we present results for exactly the same conditions as in the previous result section. All the result plots can be found on OSF (Supplementary material A).

The conditions of the first simulation study are identical to the 1:3 sample size ratio condition in the analytic section in the previously described sections. Unsurprisingly, the simulations results are almost identical to the analytic results, with minor differences due to estimation and the different number of degrees of freedom while estimating (K-1 instead of K for the analytic section; see supplemental materials for more details). As the only difference between the first and second additional simulation study is the sample size distribution (30-30-30-30-90 corresponding to those of the main analyses, and 6-30-42-54-78 in the second study, which have same mean and standard deviation of sample sizes), comparing their results allows for a direct and unbiased evaluation of the effect of the sample size distribution on the effects of publication bias. The similarities of results are striking. When there is no publication bias, results are identical. In the presence of publication bias, the same combinations of publication bias and effect size result in over- or underestimations of heterogeneity. Only minor differences occur in the amount of over- or underestimation of heterogeneity, and hence in statistical power and type I error rates. When sample sizes are different (simulation study 2), bias is slightly smaller with slightly less underestimation when true heterogeneity is absent or small, and with type I error rates closer to .05 in these situations, and if power is lower than .80 it is slightly lower in simulation study 2. As differences are very minor, even for K=145, we conclude that our results on the effects of publication bias on heterogeneity assessment are robust to the sample size distribution, given the same mean and variance of this distribution.

The sample sizes of the third and fourth additional simulation studies were taken from published studies from the field of social and personality psychology (Fraley & Vazire, 2014) and from a large meta-analysis focusing on the association between brain volume and intelligence (Pietschnig, et al., 2015). The samples from the third study had the largest variation, from 10 till 10,000. When publication bias is absent, this large variation, along with the lower average sample size, results in smaller values of E(I^2), and lower amounts of power, for the same values of τ (if $\tau > 0$) and number of studies, compared to the other simulation studies. Besides the differences at baseline, the results showed the same trends of sample size variation that we saw in the main analysis. Again, underestimations of between-study variance was less severe when sample sizes vary, and the minimum type I error rate and power increase. The same impact of sample size variation can be seen in the fourth simulation study, where sample sizes range from 4 to 649. The average and variance of the sample size is smaller compared to the third simulation study, resulting in even lower estimates of between-study variance and power when there is no publication bias. In all other aspects, the same trends and patterns are again observed, that we already saw in the main analysis.

The results of our main analyses and the four simulation studies reveal that the assessment of heterogeneity and the type I error rate and statistical power of the Q-test depend on the sample size characteristics, such as the average sample size and the variation in sample size, but not as much the specific distribution of sample sizes. The same pattern and trends of the effect of publication bias can be observed in all studies with possible serious underestimation or overestimation of heterogeneity, and the type I error rate and power of the Q-test can both decrease to 0 or increase to 1 as a consequence of publication bias, depending on complex combinations of conditions.

Application to actual meta-analyses

In this section, we show how analyzing of the effect of publication bias on the assessment of heterogeneity may increase our understanding of meta-analyses. First, we introduce a web-application called Q-sense, that can be used to determine the sensitivity of the heterogeneity assessment and the *Q*-test to publication bias for a meta-analytic data set. More generally, Q-sense enables researchers to determine which values of effect size, heterogeneity, and publication bias are consistent with the observed effect size and heterogeneity. As current meta-analytic tools do not perform well under heterogeneity in the presence of publication bias (e.g., Carter, Schönbrodt, Gervais, & Hilgard, 2017; McShane et al., 2016; van Aert et al., 2016; van Assen et al., 2015). Q-sense is a timely tool to address possible effects of publication bias on meta-analytic results.

Furthermore, we will provide some guidelines for using Q-sense and we will apply *Q*-sense to two different datasets of actual meta-analyses. We first examine a meta-analysis by Der, Batty and Deary (2006) concerning the effect of breastfeeding on intelligence in children, showing a situation where there is no evidence for publication bias, and little impact on the results of the *Q*-test. We follow this with an examination of a meta-analysis concerning the relation between weight and moral judgement by Rabelo, Keller, Pilati and Wicherts (2015), showing how publication bias can explain the observed extreme homogeneity.

Q-sense

Q-sense (https://augusteijn.shinyapps.io/Q-sense/) provides a sensitivity analysis of heterogeneity assessment and the *Q*-test in a meta-analysis. More precisely, it provides the expected values and 95% intervals of values of *Q*, H^2 and I^2 as a function of publication bias for the data in the meta-analysis given specified values of true effect size and heterogeneity. Q-sense allows researchers to determine which presumed population values of effect size, heterogeneity, and publication bias are consistent with the observed effect size and observed heterogeneity. It also enables one to examine how the test of homogeneity is affected by publication bias. Heterogeneity assessment and the *Q*-test in a meta-analysis can be said to be robust if the value and CI of *Q* are relatively unaffected by publication bias for the estimated values of effect size and heterogeneity *and* the expected values of *Q* and I^2 in the sensitivity analysis are close to those observed in the meta-analysis.

Q-sense requires the sample size of the primary studies that are included in the meta-analysis, either as total sample sizes or as two subgroup sample sizes. Furthermore, it requires presumed values of the true effect size (δ), heterogeneity (τ^2), and the observed *Q*-value in the meta-analysis. Sensible presumed values are those observed in the meta-analysis, those corresponding to the null-hypothesis of a zero true effect size or homogeneity of effect sizes, those obtained with meta-analytic methods correcting for publication bias, or those from previous (meta-analytic or large single) studies relevant to the meta-analysis at hand. Using this input, Q-sense will provide the user with a plot of the average *Q* value and 95% CI for different levels of publication bias.

Q-sense varies the amount of publication bias from 0% to 100% in 32 levels (0 to 50% bias in steps of 10%, 55-80% bias in steps of 5%, and 81-100% bias in steps of 1%, since effects of publication bias are strongest in this last interval). For each level of publication bias, 50,000 ÷ *K* iterations are used. In each iteration, the meta-analytic studies' effect sizes are generated using the effect size and level of heterogeneity provided by the researcher, and *Q* and I^2 are estimated (see Supplementary Material E and "Shiny code Q-sense", both on OSF).

The average *Q* value, the 2.5th quantile and the 97.5th quantile of these iterations are used to determine the 95% confidence intervals for *Q* and I^2. These are shown in a figure as a function of publication bias in combination with both the observed value and critical value of *Q*. Furthermore, the results of the 32 publication bias levels (average *Q*-value, 95% confidence interval, average I^2 value, 95% confidence interval and whether the *Q*-value observed by the user fall within the 95% confidence interval) can be downloaded as .csv file.

Recommendations using Q-sense

We recommend applying Q-sense after assessing and testing the effect size and heterogeneity of effects in a meta-analysis. As a first step, we advise entering the estimated values of effect size and heterogeneity (without a correction for publication bias) for the required population values. If Q-sense shows that (i) both the average and CI of the *Q*-statistic are relatively unaffected by publications bias (little variation when publication bias is introduced), and (ii) the observed *Q*-statistic is in its CI for most or all values of publication bias, then one can conclude that the results of the meta-analysis (including the estimates of true effect size and heterogeneity) are robust to publication bias. We recommend reporting Q-sense's results concerning (i) and (ii) in their paper. If (i) and (ii) are both met, the meta-analytic results are robust to publication bias and can be interpreted with more confidence. However, if (i) and (ii) are not both met the meta-analytic results are not robust to publication bias and should be interpreted with caution, and we suggest to proceed with additional analyses with Q-sense.

If the meta-analytic results turn out to be not robust to publication bias, , we recommend further analyses with Q-sense to find values of true effect size, heterogeneity and publication bias that provide a CI of the *Q*-statistic that is consistent with the observed *Q*-statistic. As first follow-up analyses we recommend running Q-sense with a zero true effect size for two reasons. First, the effects of publication bias on estimation are particularly dramatic for zero true effect size. Second, the hypothesis of zero true effect size is especially practically relevant for researchers. In these follow-up analyses different values of true heterogeneity can be entered in attempts to obtain consistent results. If, for true zero effect size, combinations of values of publication bias and true heterogeneity yield both an expected effect size and a *Q*-statistic close to those observed, then the observed meta-analytic results can also be explained by a zero true effect size and publication bias. In that case the researcher must seriously consider the possibility that true effect size is indeed zero. If no combination of values of publication bias and heterogeneity can be found that yield results consistent with those observed in the meta-analysis, the researcher can be more confident that true effect size exceeds zero. Particularly if no results of these first follow-up analyses are consistent with those observed one may proceed with further follow-up analyses.

Finally, if the meta-analytic results are not robust to publication bias, one may conduct second follow-up analyses applying meta-analytic techniques that estimate true effect size and heterogeneity after adjusting for possible publication bias. The estimates obtained from these methods can also be used as input for Q-sense and checked for consistency with the observed meta-analytic findings. Following the Meta-Analysis Reporting Standards (MARS; American Psychological Association, 2008), we recommend adding a separate subsection on the heterogeneity-sensitivity analyses of the meta-analysis at the end of the results section. Many meta-analytic methods have been proposed that attempt to estimate true effect size and heterogeneity while correcting for publication bias. Current methods are still being improved, and new methods continue to be developed. The state of the art knowledge on the performance of current methods is that many of them perform relatively well under homogeneity and extreme publication bias, but fail to perform well under heterogeneity in combination with (almost) only statistically significant studies in the meta-analysis (e.g., see for an overview and discussions of (dis)advantages; Carter et al., 2017; McShane et al., 2016; van Aert et al., 2016;

van Assen et al., 2015). A discussion of all these methods and their performance in different conditions is out of the scope of this paper.

Q-sense applied to Der, Batty, and Deary (2006)

This meta-analysis, examining the effect of breastfeeding on intelligence in children, featured nine effect sizes; five of them were statistically significant, whereas four of them were not (when tested either one- or two-sided). The average sample size of these studies was N = 909.22 (sd = 1,732.25, Ns ranging from 108 to 5475). The characteristics of the meta-analysis can be found in Supplementary material F. A random-effects meta-analysis with the restricted maximum-likelihood estimator resulted in d = 0.138 [95% CI: 0.059:0.217], p = 0.0006, and $Q(8)$ = 21.05, p = .007 (I^2 = 65.76% [95% CI: 14.73%: 96.12%], τ^2 = 0.0071). Following the recommendations, we first applied Q-sense using these estimated values of effect size (d = 0.138) and heterogeneity (τ^2 = 0.0071).

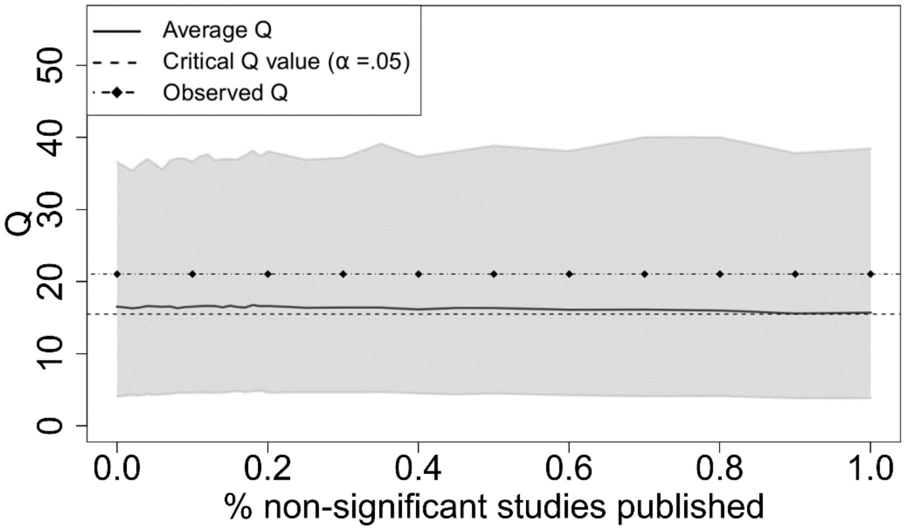

Figure 7. Output of sensitivity analysis of *Q*-sense: *Q*’s average value and CI as a function of publication bias, for the random-effects estimated values of effect size (*d = 0.1383*) and heterogeneity (τ^2 = 0.0071) in Der et al. (2006).

Figure 7 shows the average value of *Q* and its CI as a function of publication bias. This figure shows that (i) *Q*'s average and CI are hardly affected by publication bias, and (ii) the observed value of *Q* in the meta-analysis is always in *Q*'s CI based on the observed values of effect size and heterogeneity. On the basis of the analysis with Q-sense we therefore conclude that Der et al.'s (2006) meta-analytic results are robust to publication bias, thereby increasing our confidence in their meta-analytic findings that there is a small true effect size, a moderate to large heterogeneity, and little or no publication bias.[2].

Q-sense applied to Rabelo et al. (2015)

This meta-analysis contains 25 effect sizes on the effect of experiencing weight on interpersonal judgement, with 23 being statistically significant if tested two-sided, and all of them being statistically significant if tested one-sided. The average sample size is 61.12 (sd = 20.22, N ranging from 30 to 100, https://osf.io/cgmdi/ and Supplementary materials G). The authors used a fixed-effect meta-analysis on these data that resulted in $d = .57$, 95% CI [0.47, 0.67], and $Q(24) = 4.70$ ($p = .999993$), which they indicate as excessive homogeneity and a sign of publication bias. This observed *Q*-value corresponds to $H^2 = 0.196$, which indeed corresponds to extreme homogeneity since $H^2 < 1$ corresponds to $I^2 < 0$.

Following our recommendations we first apply Q-sense to the results of the fixed-effect meta-analysis ($d = .57$, $\tau^2 = 0$). *Q*-sense reveals (see Figure 8) that (i) both the average and CI of the *Q*-statistic are affected by publications bias (the average value of *Q* and the upper and

[2] We note that the conditions of this meta-analysis are similar to those in the two most right plots in the last row of Figure 3 (large variation in sample size, medium to large amount of heterogeneity). If we draw a horizontal line at the effect size $\delta = 0.138$, we are close to the line where heterogeneity is correctly estimated regardless of the level of publication bias.

lower bound of the CI are almost halved when there is full publication bias compared to no publication bias), and (ii) the observed *Q*(24) = 4.7 never falls in the 95% CI of the *Q*-test, regardless of the level of publication bias. Hence we conclude that the results of the meta-analysis are not robust to publication bias, and continue with our first follow-up analyses with Q-sense.

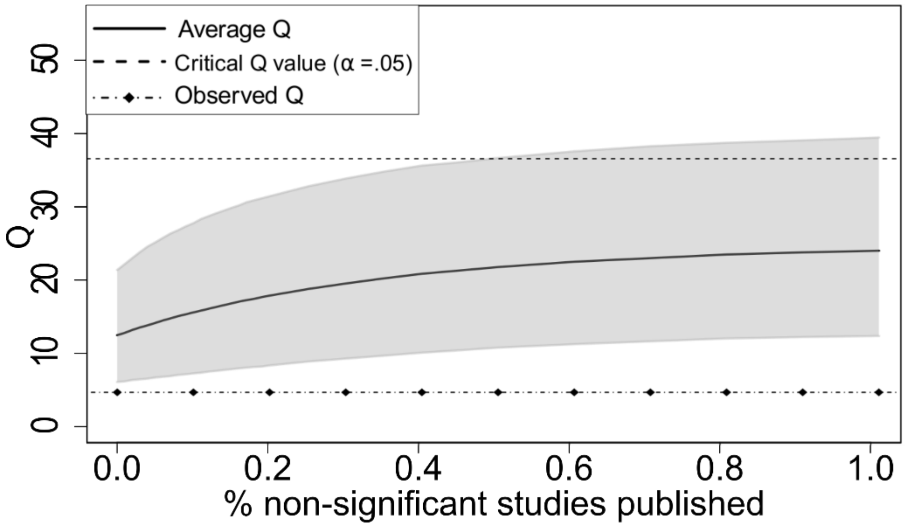

Figure 8. Output of sensitivity analysis of *Q*-sense: *Q*'s average value and CI as a function of publication bias, for the fixed-effects estimated values of effect size (*d = 0.57*) and heterogeneity ($\tau^2 = 0$) in Rabelo et al. (2015).

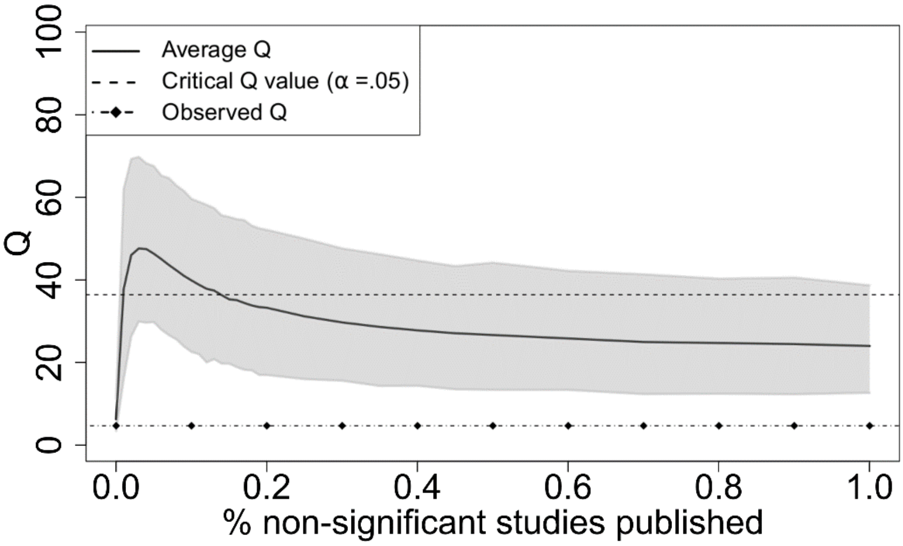

Figure 9. Output of sensitivity analysis of *Q*-sense: *Q*'s average value and CI as a function of publication bias, for the corrected estimated values of effect size (*d = 0*) and heterogeneity ($\tau^2 = 0$) in Rabelo et al. (2015).

Again following our recommendations, we then examined if the meta-analytic findings are consistent with a true effect size equal to zero. We kept heterogeneity at $\tau^2 = 0$, as estimated in the meta-analysis. Figure 9 shows that also in this situation *Q*´s average value and CI are strongly affected by publication bias: When there is no bias (published = 100%), the average observed *Q* value is 24.01 ($p = .461$) [95% CI: 12.65:38.69]. As publication bias increases, the expected value of Q also increases, with a maximum of $Q = 47.63$ ($p = .003$) [95% CI: 29.92:69.79] when published = 3%. In case of full publication bias the average drops to $Q = 6.29$ ($p = .99989$) [95% CI: 3.42:11.16], which is consistent with the observed extreme homogeneity. Figure 10 shows the empirical sampling distribution of *Q* when both the effect size, amount of heterogeneity, and the percentage of non-significant studies published is equal to 0. This figure confirms that both extreme homogeneity and the observed value of *Q* in the meta-analysis are consistent with this scenario. Hence, we conclude on the basis of Q-sense that the observed meta-analytic findings are consistent with a zero true effect size (instead of a medium to large effect size), no heterogeneity, and extreme publication bias.

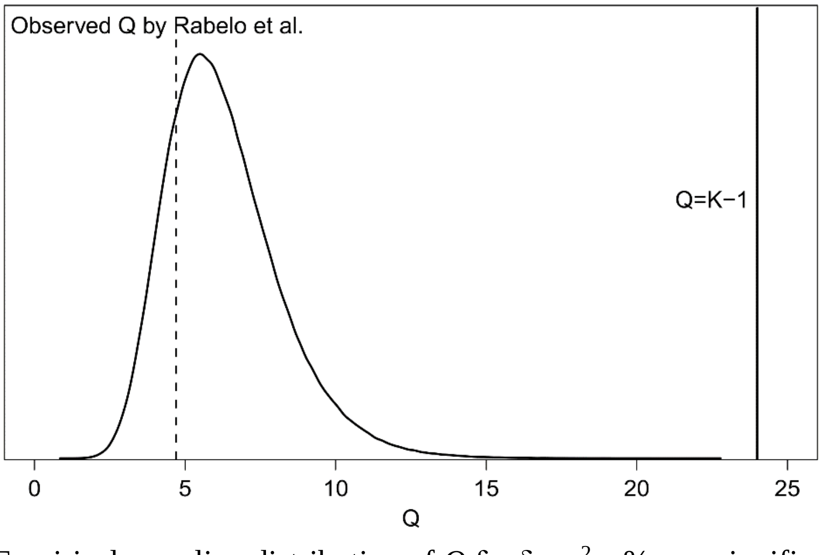

Figure 10: Empirical sampling distribution of *Q* for $\delta = \tau^2 =$ % non-significant studies published = 0, for the meta-analysis of Rabelo, Keller, Pilati and Wicherts (2015). The dashed line shows the observed values by Rabelo et al., (2015). The solid vertical line at $Q = K–1$ shows the expected value when there is no heterogeneity.

We note that Rabelo et al., (2015) also applied *p*-uniform, a meta-analytic method correcting for publication bias that has been shown to work well under homogeneity in combination with many statistically significant studies (van Assen et al., 2015; van Aert, Wicherts, & van Assen, 2016). The publication bias test of *p*-uniform indeed indicated the presence of publication bias (L = 5.1, p < .001), and yielded a corrected estimate of the effect size of d = - 0.179 [-0.676 : 0.159], p = .831, which is in line with the findings of *Q*-sense that the true effect size may equal zero.

Conclusion and discussion

Meta-analyses aim to estimate effect sizes, heterogeneity of effect sizes, and to explain possible heterogeneity using moderators. It is well established that tests and estimates of heterogeneity are influenced by publication bias (Ioannidis, 2008; Jackson, 2006a, 2006b; McShane et al, 2016). However, it remained largely unclear how they are influenced, how severe this influence is, and which factors moderate this influence. The first contribution of this paper was to examine the effect of publication bias on the Q-test and assessments of heterogeneity in a multitude of conditions which, as opposed to previous research on this topic, providing several novel findings on the effect of publication bias. Corroborating the findings of Jackson (2006a, 2007) we found that the effect of publication bias on the assessment of heterogeneity and the performance of the *Q*-test intricately depends on the true effect size, the amount of true heterogeneity, the number of studies, and variation in sample size. The effect of publication bias is non-linear and complex; publication bias can either decrease or increase the expected amount of heterogeneity, depending on the value of the true effect size and severity of publication bias. It is not surprising that the effects of publication bias are larger when the true effect size is smaller. When the true effect size is large, more studies are statistically significant and will be published, and publication bias has less impact because it only applies to non-significant studies. When true heterogeneity is large, heterogeneity is typically (albeit certainly not always) underestimated. For equal sample sizes, extreme homogeneity can occur, especially when the true effect size is small and publication bias is large. When sample sizes vary, extreme homogeneity is expected less frequently, since the minimal expected values of heterogeneity are larger.

Publication bias also causes the Type I error rate and statistical power of the *Q*-test to decrease or increase, again depending on the value of the true effect size and publication bias. The power can even drop to zero, so that the presence of true heterogeneity is impossible to detect. At the same time, the Type I error rate can be as high as 1, meaning that a meta-analysis is guaranteed to find statistically significant heterogeneity, even though the studies are truly homogenous. The power and the Type I error rate of the *Q*-test not only depend on the number of studies and the size of these studies, but also on differences in the studies' sample size, demonstrating the complex effect of publication bias on the assessment of heterogeneity. While it is commonly-stated that the *Q*-test only has sufficient statistical power when the number of studies is large (Hardy & Thompson, 1998; Thompson & Pocock, 1991), our work demonstrates that publication bias, particularly in combination with a small true effect size, may have a large effect on its power (and Type I error rate) as well. To conclude, publication bias has a large effect on assessments of heterogeneity, particularly when publication bias is severe and the true effect size is not large. Consequently, the *Q*-test and assessments of heterogeneity will be biased in these conditions. Furthermore, extreme homogeneity is particularly likely when the amount of true heterogeneity is low, the true effect size is small, and the number of studies is large.

Our second contribution was to develop the web-application *Q*-sense to provide insight in the sensitivity of the results of the *Q*-test to publication bias. The results of a meta-analysis are robust if the observed value and CI of *Q* are relatively unaffected by publication bias for the estimated values of effect size and heterogeneity *and* the value of *Q* in the sensitivity analysis is close to those calculated in the meta-analysis. We advise meta-analysts to report the results of Q-sense in their manuscripts and to investigate whether other combinations of true heterogeneity, effect size, and publication bias could also have resulted in the observed heterogeneity. We applied Q-sense to two published meta-analyses, illustrating how this web-based routine can improve our understanding of meta-analytic results in the presence of publication bias.

Our results also provide new insight into previous research on meta-analyses. For example, Ioannidis et al. (2006), observed that the likelihood of extreme homogeneity appearing in a meta-analysis was unrelated to the number of studies it included. This is surprising, since meta-analyses that include more studies are more likely to have a higher *p*-value on the *Q*-test, provided there is no or small true homogeneity. Ioannidis et al. only infrequently observed extreme homogeneity, suggesting that fields where publication bias is high and true effect sizes are small, we are either dealing with large variations in sample size, or, considerable amounts of true heterogeneity. In their article, Ioannidis, et al. suggest multiple explanations for extreme homogeneity, such as random chance, the metric of the treatment effect, correlated data, stratified or blocked randomization, and fraud. We would like to add publication bias as an additional source of extreme homogeneity that should be considered. The meta-analysis of Rabelo et al. (2015) is an example where a likely explanation of the observed extreme homogeneity is publication bias in combination with a zero true effect size and relatively similar study sample sizes.

In the design of our analyses, we made some assumptions that may limit the generalizability of some of our results. First, we assumed the effects sizes to be normally distributed. We do not feel this is a substantial limitation, because research has shown that when the distribution of effect sizes are non-normal, the Type I error rates and power of the *Q*-

test are still approximately correct when Hedges' *g* is used as effect size measure (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006). Secondly, we interpreted bias in the publication process as a strict division between significant studies (*p*<.05) and nonsignificant studies (*p*>.05). However, other scenarios of publication bias are certainly possible. For instance, it is possible that the probability of publication increases monotonically with a study's *p*-value; the smaller the *p*-value, the higher the likelihood to get published. Furthermore, studies with large samples may be more likely to be published, even if their results are nonsignificant, than small studies with nonsignificant results. Other models of publication bias might offer other results; however, we anticipate that these results too will show that publication bias may strongly affect the assessment of heterogeneity in complex, non-linear ways (e.g. Jackson, 2006a, 2007). Moreover, our main analyses focused on an 80%-20% mixture of small-large studies using certain sample size ratios when investigating the effects of sample size variability. As the implementation of sample size variability may affect overall trends and conclusions, we carried out and discussed the results of four additional simulation studies with other implementations of sample size variation, including one used for comparison, one with a different fixed distribution of sample sizes and two based on sample sizes from the published literature. As trends and conclusions of these four additional simulation studies were similar to those of our main analysis, we conclude that our main results on the effect of sample size variation are generalizable.

Future research could investigate the influences of our choices on the effects of publication bias on the assessment of heterogeneity. First, while we conducted our analyses using Cohen’s *d*, researchers could also examine effects of publication bias using other effect size measures with their own idiosyncrasies. Second, more research on alternative approaches to estimate heterogeneity that corrects for publication bias is clearly needed. Three promising meta-analytic approaches are selection models, Bayesian methods, and methods based on *p*-values.

Selection models explicitly model publication bias, i.e. the probability of findings to get into the literature, and allow for the estimation of both true effect size and heterogeneity (Hedges & Vevea, 2005). Unfortunately, these models are complex, require strong assumptions on the publication bias mechanism, and may require a large number of studies to converge, making them less useful for most meta-analyses (Borenstein et al., 2009; Field & Gillet, 2010). Some Bayesian methods are based on selection models, but incorporate priors for true effect size and true heterogeneity (Kicinski, 2013). Bayesian methods have been developed only recently, and their properties for estimation of heterogeneity are therefore still largely unknown (e.g. Gronau, Duizer, Bakker, & Wagenmakers, 2015; Guan., & Vandekerckhove, 2016). Finally, *p*-uniform (van Assen et al., 2015; van Aert et al., 2016) and *p*-curve (Simonsohn et al., 2014b) are meta-analytic methods that accurately estimate overall effect size in the presence of publication bias, for any number of significant studies, but only when true effect size is homogenous (van Aert et al., 2016). The advantages of the methods based on *p*-values over selection models and Bayesian methods are the weaker assumptions on publication bias and its ability to accurately estimate effect size even when the number of studies is small. However, to become useful for most applications, these methods should be modified in future research such that they can deal with, test for, and assess heterogeneity.

It is known that tests of publication bias provide invalid results in case of true heterogeneity (Ioannidis & Trikalinos, 2007; Peters et al., 2010). It is also important to examine the effect of another aspect of the publication process, *p*-hacking, on the assessment of heterogeneity. *P*-hacking is the result of the researchers' behaviour directed at obtaining statistically significant results (Simmons, Nelson, & Simonsohn, 2011). Examples of *p*-hacking are testing many variables or adding observations up to the point where results are significant, dropping conditions or post-hoc outlier removal (see Wicherts et al., 2016 for an extensive overview). *P*-hacking increases the probability on a Type I error of a study. Research has shown that some methods of *p*-hacking influence the assessment of the effect size in meta-analyses (van Aert et al., 2016; Simonsohn, et al., 2014b), but it is unclear how *p*-hacking influences the assessment of heterogeneity. So where publication bias results in an unrepresentative sample of studies in the meta-analysis, some methods of *p*-hacking may also lead to a distorted sample with inflated effect sizes in the primary studies. Investigating *p*-hacking and enhancing meta-analytic methods in such a way that they can assess their effects on the estimation of true effect sizes, heterogeneity, and moderator effects is an important step to improve the quality of meta-analytic research.

This article provides more insight in the complex and non-linear impact of publication bias on the assessment of heterogeneity in meta-analysis. Publication bias can not only result in incorrect conclusions regarding true effect size, but heterogeneity as well. Furthermore, we have developed a web-application, Q-sense, that allows researchers to investigate the impact of publication bias on their estimates of heterogeneity and the robustness of their meta-analytic estimates to publication bias. As publication bias may strongly affect the assessment of heterogeneity, we acknowledge the importance of developing meta-analytic methods that correct for publication bias, not only when estimating the effect size, but also when estimating heterogeneity.

References

Adams, R. A., & Essex, C. (2013). *Calculus: A complete course* (8 ed.). Toronto: Pearson.

Aguinis, H., Gottfredson, R. K., & Wright, T. A. (2011). Best-practice recommendations for estimating interaction effects using meta-analysis. *Journal of Organizational Behavior, 32*(8), 1033-1043. doi:10.1002/job.719

American Psychological Association. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *American Psychologist, 63,* 839-851.

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

Borenstein, M., Hedges, L.V., Higgins, J.P.T., & Rothstein, H.R. (2009). *Introduction to meta-analysis.* West Sussex: John Wiley & Sons, ltd.

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2017, November 22). Correcting for bias in psychology: A comparison of meta-analytic methods. Retrieved from psyarxiv.com/9h3nu

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics, 10*(1), 101-129.

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*(4), 447.

Der, G., Batty, G. D., & Deary, I. J. (2006). Effect of breast feeding on intelligence in children: prospective study, sibling pairs analysis, and meta-analysis. *Bmj*, 333(7575), 945.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*(3), 177-188.

Ellis, P. D. (2010). The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results. Cambridge University Press.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries.

Scientometrics, 90(3), 891-904. doi:10.1007/s11192-011-0494-7

Fraley, R. C., & Vazire, S. (2014). The N-pact factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PloS one, 9*(10), e109019.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505.

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

Gronau, Q. F., Duizer, M., Bakker, M., & Wagenmakers, E. J. (2015). *Bayesian Mixture Modeling of Significant P Values: A Meta-Analytic Method to Estimate the Degree of Contamination from E0.* Manuscript submitted for publication. Retrieved from: http://www.ejwagenmakers.com/submitted/GronauMixtureModel.pdf

Guan, M., & Vandekerckhove, J. (2016). A Bayesian approach to mitigation of publication bias. *Psychonomic bulletin & review, 23*(1), 74-86.

Hardy, R.J., & Thompson, S.G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in medicine, 17,* 841-859. doi: 10.1371/journal.pone.0109019

Hartgerink, C. H., Wicherts, J. M., & van Assen, M. A. (2017). Too good to be false: Nonsignificant results revisited. *Collabra: Psychology, 3*(1)

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*(3), e1002106. doi:10.1371/journal.pbio.1002106

Hedges, L.V. (1992). Modeling Publication Selection Effects in Meta-Analysis. *Statistical Science, 7*(2), 246-255.

Hedges, L.V., & Vevea, J. (2005). Selection method approaches. In H.R. Rothstein, A.J. Sutton, & M. Borenstein (Eds). *Publication bias in meta-analysis: Prevention, assessment and adjustment* (pp. 145-174). West Sussex, United Kingdom: Wiley doi:10.1002/0470870168.ch1

Higgins, J. P. T., & Green, S. (2011). *Cochrane handbook for systematic reviews of interventions* (Version 5.1.0, updated March 2011 ed.). Londen, UK: The Cochrane Collaboration. Available from http://www.cochrane-handbook.org/.

Higgins, J., & Thompson, S. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine, 21,* 1539-1558. doi:10.1002/sim.1186

Higgins, J., Thompson, S., Deeks, J., & Altman, D. (2002). Statistical heterogeneity in systematic reviews of clinical trials: a critical appraisal of guidelines and practice. *Journal of health services research & policy, 7,* 51-61. doi: 10.1258/1355819021927674

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327,* 557-560.

Huedo-Medina, T.B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analyses: Q static or I2 index? *Psychological methods, 11,* 193-206.

Ioannidis, J.P.A. (2008). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of evaluation in clinical practice, 14,* 951-957. doi:10.1111/j.1365-2753.2008.00986.x

Ioannidis, J.P.A., Trikalinos, T.A., & Zintzaras, E. (2006). Extreme between study homogeneity in meta-analyses could offer useful insights. *Journal of clinical epidemiology, 59,* 1023-1032. doi:10.1016/j.jclinepi.2006.02.013

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

Iyengar, S., & Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science,* 109-117.

Jackson, D. (2006a). The implications of publication bias for meta-analysis' other parameter. *Statistics in medicine, 25,* 2911-1921. doi:10.1002/sim.2293

Jackson, D. (2006b). The power of the standard test for the presence of heterogeneity in meta-analysis. *Statistics in Medicine*, *25*, 2688-2699.

Jackson, D. (2007). Assessing the implications of publication bias for two popular estimates of between-study variance in meta-analysis. *Biometrics*, *63*, 187-193.

Kicinski, M. (2013). Publication bias in recent meta-analyses. *PloS one, 8*(11), e81823.

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. *British Journal of Mathematical & Statistical Psychology, 31*, 107-112.

Langan, D., Higgins, J., & Simmonds, M. (2016). Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Research synthesis methods, 8*(2), 181-198. doi:10.1002.jrsm.1198

McShane, B. B., Böckenholt, U., & Hansen, K. T. (2016). Adjusting for publication bias in meta-analysis: An evaluation of selection methods and some cautionary notes. *Perspectives on Psychological Science*, *11*(5), 730-749.

Nuijten, M. B., van Assen, M. A. L. M., Veldkamp, C. L. S., & Wicherts, J. M. (2015). The replication paradox: Combining studies can decrease accuracy of effect size estimate. *Review of General Psychology, 19* (2), 172-182.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716.

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., Rushton, L., Moreno, S. G. (2010). Assessing publication bias in meta-analyses in the presence of between-study heterogeneity. *Journal of the Royal Statistical Society, 173,* 575-591.

Pietschnig, J., Penke, L., Wicherts, J. M., Zeiler, M., & Voracek, M. (2015). Meta-analysis of associations between human brain volume and intelligence differences: How strong are they and what do they mean? *Neuroscience & Biobehavioral Reviews*, *57*, 411-432.

Rabelo, A. L., Keller, V. N., Pilati, R., & Wicherts, J. M. (2015). No effect of weight on judgments of importance in the moral domain and evidence of publication bias from a meta-analysis. PloS one, 10(8), e0134808.

Raudenbush, S. W. (2009). Analyzing effect sizes: Random-effects models. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (pp. 295-315). New York: Russell Sage Foundation.

Rhodes, K.M., Turner, R.M., & Higgins, J.P.T. (2015). Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continues outcome data. *Journal of epidemiology*, 68, 52-60. doi:10.1016/j.jclinepi.2014.08.012

Rücker, G., Schwarzer, G., Carpenter, J.R., & Schumacher, M. (2008). Undue reliance on *I*2 in assessing heterogeneity may mislead. *BMC Medical research methodology, 8,* 79. doi:10.1186/1471-2288-8-79

Schmidt, F.L., Oh, I.S., Hayes, T.L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *Mathematical and Statistical Psychology, 62*(1), 97-128. doi:10.1348/000711007X255327

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*(11), 1359-1366. doi:10.1177/0956797611417632

Simonsohn, U., Nelson, L.D., & Simmons, J.P. (2014a). *P*-curve: A key to the file drawer. *Journal of Experimental Psychology: General, 143,* 534-547. doi:10.1037/a0033242

Simonsohn, U., Nelson, L.D., & Simmons, J.P. (2014b). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on psychological science, 9,* 666-681. doi:10.1177/1745691614553988

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society open science*, *3*(9), 160384.

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. doi:10.1080/00031305.1995.10476125

Sterne, J.A.C., & Egger, M. (2005). Regression methods to detect publication bias and other bias in meta-analysis. In Borenstein, M., Rothstein, H., & Sutton, A. J. (Eds.). *Publication bias in meta-analysis: Prevention, assessment and adjustments*. (pp. 99-110). West Sussex, United Kingdom: Wiley. doi:10.1002/0470870168.ch1

Thompson, S. G., & Pocock, S. J. (1991). Can meta-analyses be trusted?. *The Lancet, 338*(8775), 1127-1130.

Thorlund, K., Imberger, G., Johnston, B.C., Walsh, M., Awad, T., Thabane, L., ... Wetterslev, J. (2012). Evolution of heterogeneity (I^2) estimates and their 95% confidence intervals in large meta-analyses. *PLoS One*, *7*(7), e39471

van Aert, R.C.M., Wicherts, J.M., & van Assen, M.A.L.M. (2016). Conducting meta-analyses based on p-values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on psychological science*, *11*(5), 713-729. doi: 10.1177/1745691616650874

van Assen, M. A. L. M., van Aert, R. C. M., & Wicherts, J. M. (2015). Meta-Analysis Using Effect Size Distributions of Only Statistically Significant Studies. *Psychological Methods, 20*(3), 293-309. doi:10.1037/met0000025.

Veroniki, A.A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., … Salanti, G. (2015). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods, 7,* 55-79. doi:10.1002/jrsm.1164

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. Journal of Educational and Behavioral Statistics, 30(3), 261-293.

Viechtbauer, W. (2007). Approximate confidence intervals for standardized effect sizes in the two-independent and two-dependent samples design. Journal of Educational and Behavioral Statistics, 32(1), 39-60.

Wicherts, J. M., Veldkamp, C. L.S., Augusteijn, H.E.M, Bakker, M., Van Aert, R.C.M., & Van Assen, M.A.L.M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in psychology*, *7*.