Meta-analysis

van Aert, Robbie

Publication date:
2018

Document Version
Publisher's PDF, also known as Version of record

Meta-analysis: Shortcomings and potential

Robbie Cornelis Maria van Aert

The research in all chapters was funded by The Netherlands Organization for Scientific Research (NWO; grant project number 406-13-050) and the research in Chapter 4 was also funded by the Berkeley Initiative for Transparency in the Social Sciences and the Laura and John Arnold Foundation.

Printing was financially supported by Tilburg University.

Cover design: Mandy van de Watering
Printed by: GVO drukkers & vormgevers
ISBN: 978-94-6332-367-3

Meta-analysis: Shortcomings and potential

Proefschrift ter verkrijging van de graad van doctor
aan Tilburg University
op gezag van de rector magnificus,
prof. dr. E. H. L. Aarts,
in het openbaar te verdedigen
ten overstaan van
een door het college voor promoties aangewezen commissie
in de aula van de Universiteit
op vrijdag 6 juli 2018 om 14:00 uur
door Robbie Cornelis Maria van Aert
geboren te Roosendaal en Nispen.

Promotiecommissie

Promotores:	Prof. dr. M. A. L. M. van Assen Prof. dr. J. M. Wicherts
Overige leden:	Prof. dr. W. Van Den Noortgate Prof. dr. E. M. Wagenmakers Prof. dr. P. Cuijpers Dr. E. M. L. Dusseldorp Dr. M. Bakker

Contents

CHAPTER 1

Introduction

More and more scientific papers get published in an increasing number of journals (e.g., Adair & Vohra, 2003; Byyny, 2012). This information explosion makes it hard for researchers and other users of scientific knowledge to keep up with all the papers that are relevant for them. Hence, tools are needed to synthesize and summarize the existing literature on a particular topic (e.g., Cooper, Hedges, & Valentine, 2009a). Meta-analysis is a tool that statistically combines effect sizes from independent primary studies on the same topic (e.g., Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper, Hedges, & Valentine, 2009b), and is now seen as the "gold standard" for synthesizing and summarizing the results from different primary studies (Aguinis, Gottfredson, & Wright, 2011; Head, Holman, Lanfear, Kahn, & Jennions, 2015). The popularity of meta-analysis is also reflected in the rapid increase of the relative number of published meta-analyses. Ioannidis (2016) studied the publication rate of papers that were tagged as a meta-analysis in PubMed and observed that the publication rate increased by 2,635% between 1991 and 2014 whereas the publication rate of all PubMed-indexed items increased by only 153% in the same period.

As for every statistical technique, confidence in the interpretation of the results of a meta-analysis is limited by the quality of the data that are analyzed. The term *garbage in, garbage out* is often used to denote that meta-analyzing biased primary studies will yield biased results of the meta-analysis as well (Borenstein et al., 2009; Eysenck, 1978). Many researchers argue that the published literature in many fields is affected by biases in the (reporting of) primary study results (e.g., C. J. Anderson et al., 2016; Begley & Ioannidis, 2015; Camerer et al., 2016; Ioannidis, 2005; Maxwell, Lau, & Howard, 2015; Open Science Collaboration, 2015; Pashler & Harris, 2012). One strong signal that the published literature is biased is the evidence for publication bias that has been amassed in various research fields (e.g., Driessen, Hollon, Bockting, Cuijpers, & Turner, 2015; Franco, Malhotra, & Simonovits, 2014; Franco, Simonovits, & Malhotra, 2016; Sterling, Rosenbaum, & Weinkam, 1995). Publication bias refers to statistically nonsignificant effect sizes having a lower probability of getting published than significant effect sizes (Rothstein, Sutton, & Borenstein, 2005a). In its most extreme case, this means that only studies with statistically significant effect sizes get published while studies with nonsignificant effect sizes are left unpublished. Publication bias is seen as a major threat to the validity of meta-analyses (Dickersin & Min, 1993; Easterbrook, Berlin, Gopalan, & Matthews, 1991; Rothstein et al., 2005a). Publication bias might strengthen researcher's belief that only statistically significant results are eligible for publication. Consequently, researchers confronted with a nonsignificant result might decide not to publish or to engage in so-called *p*-hacking behaviors or opportunistically use researcher degrees of freedom in the analysis of data to obtain significant results. Such ad hoc and data driven decisions in the analysis of data violate the assumptions of the statistical procedures to test hypotheses (Simmons, Nelson, & Simonsohn, 2011;

Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011; Wicherts et al., 2016); these practices inflate the false positive rate (Type I error rate) if the null hypothesis of a zero effect is true and inflate effect size estimates for genuine non-null effects.

Evidence of publication bias is apparent in the unrealistic high rate of statistically significant results in the literature. For example, approximately 95% of the tested hypotheses was supported in a random sample of published papers in psychology and psychiatry (Fanelli, 2010a, 2012). This is not in line with the average statistical power estimated at .35 (Bakker, van Dijk, & Wicherts, 2012) and .47 (Cohen, 1990) in psychological research, which would give rise to less than half of studies of genuine effects being significant. This implies that statistically nonsignificant effect sizes are often not published. This, in turn, decreases the likelihood that these nonsignificant effect sizes are included in a meta-analysis.

Another signal that the published literature may be distorted are the results of three large-scale projects examining the replicability of studies (i.e., are study results the same if a study is conducted a second time) published in the fields of psychology (Open Science Collaboration, 2015), economics (Camerer et al., 2016), and cancer biology research (Begley & Ellis, 2012; Errington et al., 2014). The Reproducibility Project: Psychology replicated 100 studies that were published in major journals in 2008. The results of this project revealed that 97 studies were statistically significant in the published original study, but that only 35 (36.1%) of these studies were also statistically significant in the replication. Furthermore, the effect size in the published original study was larger in 81 (83.5%) of the 97 studies than in their replication. In economics, 18 published studies on experimental economics were replicated in the Experimental Economics Replication Project. Sixteen of these 18 published original studies were statistically significant, and of these 16 studies 11 (68.8%) were statistically significant in the replication. Moreover, 13 of the 16 (81.3%) studies had a larger effect in the original study than in the replication. The Reproducibility Project: Cancer Biology (Begley & Ellis, 2012; Errington et al., 2014) replicated 53 landmark studies in the field of hematology and oncology, and only 6 (11.3%) of these studies were confirmed in the replication. The results of these projects raise doubts about the reliability of the effects and their size in the published literature.

Two parameters are usually of primary interest in a meta-analysis; the average effect size and the between-study variance in primary studies' true effect size. The average effect size is a weighted average of the primary studies included in a meta-analysis. Advantages of synthesizing primary studies by means of a meta-analysis instead of evaluating the primary studies in isolation is that the average effect size is generally closer to the true average effect size and statistical power for testing the null hypothesis of no effect is larger in a meta-analysis (Borenstein et al., 2009). The other parameter that is of interest for meta-analysts is the between-study variance in primary studies' true effect sizes. This denotes whether there is one fixed

true effect size underlying all the primary studies included in a meta-analysis (homogeneous true effect sizes) or whether each primary study has its own true effect size (heterogeneous true effect sizes) (Borenstein et al., 2009; Borenstein, Hedges, Higgins, & Rothstein, 2010). Estimating the between-study variance is of importance (Higgins & Thompson, 2002; Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006; Mittlböck & Heinzl, 2006; Veroniki et al., 2016), because it may lead to relevant insights about the consistency of the true effect sizes (Higgins, Thompson, & Spiegelhalter, 2009). Research has shown that heterogeneity is often present in meta-analyses (Higgins, 2008; Kontopantelis, Springate, & Reeves, 2013). Two so-called Many Labs (Klein et al., 2014; Klein et al., 2017) projects replicated published studies in multiple laboratories across the world. Even though the studies were exactly replicated in the laboratories and used the same study protocol, there was still heterogeneity in true effect size in 8 out of 16 (50%) and 13 out of 28 (46.4%) of the replicated studies.

Publication bias is well-known to yield overestimated primary studies' effect sizes (e.g., Ioannidis, 2008b; Lane & Dunlap, 1978), and therefore also overestimated average effect size if these studies are combined in a meta-analysis. Publication bias also results in bias in the estimate of the between-study variance in primary studies' true effect sizes, but this bias can be positive or negative depending on characteristics of a meta-analysis (Augusteijn, van Aert, & van Assen, 2017; Jackson, 2006, 2007). Hence, it is of importance to develop methods that enable accurate estimation of the average effect size as well as the between-study variance in primary studies' true effect sizes. Most importantly, meta-analyses that are biased do not only hamper scientific progress, but are also detrimental for practice since results of meta-analyses are often used for policy making (Polanin, Tanner-Smith, & Hennessy, 2016).

Outline of this dissertation

The goals of this dissertation are to develop new meta-analysis methods and examine the statistical properties of existing methods. In Chapter 2, we propose the *p-uniform* method that is a new meta-analysis method for homogeneous effect sizes that takes into account publication bias using only the statistically significant primary studies' effect sizes. *P*-uniform is able to (i) test for publication bias, (ii) estimate effect size and compute a confidence interval around this estimate, and (iii) test the null hypothesis of no effect. We conduct a Monte-Carlo simulation study and compare effect size estimation of *p*-uniform with other meta-analytic methods, namely the trim-and-fill method that also attempts to correct for publication bias (Duval & Tweedie, 2000a, 2000b) and fixed-effect and random-effects meta-analysis that do not correct for publication bias. *P*-uniform's publication bias test is compared with the test of excess significance (Ioannidis & Trikalinos, 2007b) that was developed to see whether a set of outcomes shows an excess of significant outcomes.

Other researchers independently of us developed the *p*-curve method

(Simonsohn, Nelson, & Simmons, 2014a, 2014b) that is similar to *p*-uniform. However, this method is not able to compute a confidence interval for the effect corrected for publication bias and does not offer a publication bias test. The goal of Chapter 3 is to inform applied researchers about the differences between *p*-uniform and *p*-curve and provide recommendations for applying the methods in practice. To substantiate these recommendations, we assess the statistical properties of *p*-uniform and *p*-curve if there is heterogeneity in primary studies' true effect sizes. We also show the consequences of researchers using *p*-hacking in the primary studies on estimates of *p*-uniform and *p*-curve, and illustrate how *p*-uniform and *p*-curve can be applied by analyzing the meta-analysis of Rabelo, Keller, Pilati, and Wicherts (2015) on the relationship between weight on judgement of importance.

In Chapter 4, we present the results of a pre-registered study examining the presence of publication bias and the overestimation in effect size caused by it in meta-analyses published in the fields of psychology and medicine. This is done by first creating a large-scale data set of 83 meta-analyses published in Psychological Bulletin and 499 systematic reviews from the Cochrane Database of Systematic Reviews representing data form psychology and medicine, respectively. Subsequently, we create homogeneous subsets of the meta-analyses in this large-scale data set, because publication bias methods do not have good statistical properties if the primary studies' true effect sizes are heterogeneous (e.g., Ioannidis & Trikalinos, 2007a, 2007b; McShane, Böckenholt, & Hansen, 2016; Terrin, Schmid, Lau, & Olkin, 2003). We subsequently test for publication bias by applying the rank-correlation test (Begg & Mazumdar, 1994), Egger's test (Egger, Smith, Schneider, & Minder, 1997), the test of excess significance (Ioannidis & Trikalinos, 2007b), and *p*-uniform's publication bias test. Overestimation in effect size caused by publication bias and characteristics of meta-analyses that are related to this overestimation are studied by comparing estimates of *p*-uniform with estimates of traditional random-effects meta-analysis.

In Chapter 5, we present *p-uniform** that extends and improves *p*-uniform in three important ways. First, *p*-uniform* no longer overestimates the effect if there is heterogeneity in primary studies' true effect sizes. Second, the method provides a more efficient estimator than *p*-uniform. Third, *p*-uniform* enables estimating and drawing inferences for the between-study variance. We compare the statistical properties of *p*-uniform* to the selection model proposed by Hedges (1992), because selection model approaches are nowadays seen as the state-of-the-art publication bias method (McShane et al., 2016). To compare the statistical properties of both methods, we use an analytical study revolving around meta-analysis consisting of only one statistically significant and one nonsignificant primary study's effect size and run a Monte-Carlo simulation using conditions that are representative for meta-analyses in practice. We also offer recommendations on how to examine publication bias in a meta-analysis based on the results of our analyses.

In Chapters 6 and 7, we anticipate on a question that emerged mainly because

of the Reproducibility Project: Psychology (Open Science Collaboration, 2015), the Experimental Economics Replication Project (Camerer et al., 2016), and Reproducibility Project: Cancer Biology (Begley & Ellis, 2012; Errington et al., 2014): how to statistically combine the effect size of a published original study with that of a replication? It is important to realize that this is not only an issue in the context of these two projects since many applied researchers also replicate a published original study as the starting point for a multi-study paper (Neuliep & Crandall, 1993). A problem with combining these two studies is that the effect size of the published original study is most likely statistically significant, as findings in the literature generally have been subject to publication bias (e.g., Fanelli, 2010a, 2012; Sterling et al., 1995). Hence, the published original study and replication study effect size cannot be combined using traditional meta-analysis methods, because these methods do not take into account the likely overestimation of effect size in the published original study.

We develop *the hybrid method of meta-analysis* that statistically combines a published original study and a replication by taking into account the statistical significance of the original study in Chapter 6. This method enables accurate estimation of the effect size, computing a confidence interval for the effect size, and testing the null hypothesis of no effect. In Chapter 7, we propose a meta-analysis method using Bayesian statistics that statistically combines a published original study and replication. This method called is *snapshot hybrid method* and computes the posterior model probabilities at different snapshots of true effect size while taking statistical significance of the published original study into account. That is, posterior model probabilities are computed at snapshots equal to a zero, small, medium, and large true effect size when statistically combining the published original study and replication. Hence, the snapshot hybrid method provides insights into the magnitude of effect size underlying the published original study and replication. The statistical properties of the hybrid method of meta-analysis and snapshot hybrid method are both assessed using an analytical study. The hybrid method of meta-analysis is applied to the data of the Reproducibility Project: Psychology and the snapshot hybrid method is applied to the data of the Reproducibility Project: Psychology as well as the Experimental Economics Replication Project, examining the underlying true effect size of original studies and replications. We also provide recommendations for applying these methods in practice.

Chapters 2 to 7 are about meta-analysis methods to correct for publication bias or to statistically combine a published original study and replication. The last two chapters are dedicated to estimating the between-study variance (Chapter 8) and computing a confidence interval for this parameter (Chapter 9) in random-effects meta-analyses where data are assumed not to be affected by publication bias. Many different estimators for the between-study variance exist, but the Paule-Mandel (Paule & Mandel, 1982) and the restricted maximum likelihood estimator (Raudenbush,

2009) are nowadays commonly recommended (Langan, Higgins, & Simmonds, 2016; Veroniki et al., 2016). DerSimonian and Kacker (2007) developed two-step moment based estimators of the between-study variance. We extend these two-step estimators to a multi-step estimator, and illustrate how the multi-step estimator can be applied to data of three published meta-analyses (i.e., Bangert-Drowns, Hurley, & Wilkinson, 2004; Ho & Lee, 2012; Sterne, Bradburn, & Egger, 2001).

Estimates of the between-study variance in a random-effects meta-analysis are preferably reported together with a confidence interval (Chung, Rabe-Hesketh, & Choi, 2013; Kontopantelis et al., 2013; Sidik & Jonkman, 2007) since point estimates are rather imprecise especially if the number of primary studies in a meta-analysis is small (Higgins et al., 2009; Ioannidis, Patsopoulos, & Evangelou, 2007; Kepes, McDaniel, Brannick, & Banks, 2013; Langan et al., 2016). Veroniki et al. (2016) reviewed the existing methods for computing a confidence interval for the between-study variance and recommended the *Q*-profile (Viechtbauer, 2007b) or generalized *Q*-statistic method (Jackson, 2013) for use. Both methods are exact (i.e., coverage probability equal to 1-α) if the assumptions underlying the random-effects meta-analysis hold. However, these assumptions are usually violated in practice (Biggerstaff & Tweedie, 1997; Hardy & Thompson, 1998; Hoaglin, 2016a, 2016b) making the confidence intervals approximate rather than exact confidence intervals. In Chapter 9, we examine the coverage probabilities and width of the confidence interval of the *Q*-profile and generalized *Q*-statistic method in two Monte-Carlo simulation studies with conditions that are representative for actual meta-analyses but that violate the assumptions that the effect size measure follows a normal sampling distribution and that the primary studies' sampling variances are known. Moreover, we offer recommendations for computing confidence intervals for the between-study variance in practice.

In the epilogue, I discuss the findings and their implications, and provide recommendations based upon all findings presented in my dissertation. I also discuss limitations of my work and offer suggestions for future research.

CHAPTER 2

Meta-analysis using effect size distributions of only statistically significant studies

Abstract

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

This chapter is published as 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

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

One of the greatest threats to the validity of meta-analytic results is publication bias (Banks, Kepes, & Banks, 2012; Rothstein et al., 2005a). We narrowly define publication bias here as 'the selective publication of studies with a statistically significant outcome', that is, the overrepresentation in the literature of studies with a significant outcome compared to studies with so-called null results. The evidence of publication bias is overwhelming (e.g., van Assen, van Aert, Nuijten, & Wicherts, 2014). For instance, Kühberger, Fritz, and Scherndl (2014) examined 1,000 randomly drawn psychological studies in 2007 and observed three times as many outcomes just reaching significance than outcomes just failing significance. Furthermore, in psychology about 95% of published articles contain statistically significant outcomes, and this percentage has been increasing over the years (Fanelli, 2012). Neither the high percentage nor its increase can be explained by the studies' statistical power since power is generally low (Ellis, 2010) and there is no evidence that it has grown over the years (Fanelli, 2012). Explanations of publication bias include researchers' reluctance to submit studies with non-significant results (Cooper, DeNeve, & Charlton, 1997; Coursol & Wagner, 1986), and lower appraisal of these studies by reviewers (Coursol & Wagner, 1986; Mahoney, 1977) and editors (Coursol & Wagner, 1986).

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

[1] Aguinis, Dalton, Bosco, Pierce, and Dalton (2010) conclude that meta-analysis is one of the most influential methodological tools in management and related fields after observing that meta-analyses were cited three times as much as other empirical articles from 1963 to 2007 in the Academy of Management Journal, one of the most influential management journals.

of our study.

Three harmful consequences of publication bias are that researchers may exploit degrees of freedom (*df*) in the analysis of data (Simmons et al., 2011), uncertainty of the existence of a true effect underlying a published statistically significant effect, and more generally, overestimation of the population effect (e.g., Asendorpf et al., 2013). Researcher *df*, or researchers' behavior directed at obtaining statistically significant results (Simonsohn et al., 2014b), which is also known as *p*-hacking or questionable research practices in the context of null hypothesis significance testing (e.g., O'Boyle, Gonzalez-Mule, & Banks, 2017), results in a higher frequency of studies with false positives (Simmons et al., 2011) and inflates genuine effects (Bakker et al., 2012). Additionally, even in the absence of researcher *df*, systematic investigations demonstrate that publication bias leads to overestimation of effects, which can be dramatic if sample sizes are small (Bakker et al., 2012; Francis, 2012; Gerber, Green, & Nickerson, 2001; Kraemer, Gardner, Brooks, & Yesavage, 1998). Consider extreme publication bias, i.e., only statistically significant effects are published, and a population effect that is of medium or small size. A study's published effect size is then hardly informative of the underlying population effect and merely reflects sample size (Francis, 2012; Gerber, et al., 2001; Kraemer, et al., 1998). Moreover, a replication of a small study will generally obtain a smaller effect than the original study. For example, Gerber, Green, and Nickerson (2001, p.388) show that in two-group studies with a total sample size of 50, the probability is about .95 that the observed effect in the replication study is smaller than in the original study. This property may at least partly explain why many replication studies fail to confirm results of original studies (Begley & Ellis, 2012; Prinz, Schlange, & Asadullah, 2011; Sarewitz, 2012). Obviously, if individual published studies obtain biased effect size estimates, meta-analyses mainly using these individual studies will yield biased estimates as well, and may falsely give the impression of a consistent research finding (Francis, 2012).

Because of the harmful consequences of publication bias it will not come as a surprise that meta-analysis experts note that publication bias analyses should be included in meta-analytic studies (e.g., Aytug et al., 2012; Banks, Kepes, & McDaniel, 2012; Field & Gillett, 2010; Sterne, Gavaghan, & Egger, 2000; Sutton, 2005). However, publication bias is unfortunately often not adequately addressed in meta-analytic studies. For example, reviews showed that publication bias was assessed in less than 10% of meta-analytic studies in industrial organization psychology studies (Sutton, 2005), less than 10% in management sciences (Aguinis et al., 2010), 18% in organizational sciences (Aytug, et al., 2012), 56% in education research (Banks, Kepes, & Banks, 2012), 31% in management and industrial/organizational psychology (Banks, Kepes, & McDaniel, 2012), 70% in journals published by the American Psychological Association and the Association for Psychological Science (Ferguson & Brannick, 2012), and 33% in judgment and decision making research (Renkewitz,

Fuchs, & Fiedler, 2011). To conclude, the failure to address publication bias is omnipresent, although there is considerable variation across disciplines.

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

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

2.1 Methods for assessing publication bias

We briefly discuss the following methods for assessing publication bias, along with their most important properties: failsafe *N* (Rosenthal, 1979), funnel plot (Light & Pillemer, 1984), Begg and Mazumdar (1994) rank correlation test, Egger's test (Egger et al., 1997), the trim-and-fill method (Duval & Tweedie, 2000b), the TES, and selection models (Hedges & Vevea, 2005). The oldest and most popular (e.g., Banks,

Kepes, & McDaniel, 2012; Ferguson & Brannick, 2012) method is failsafe *N* (Rosenthal, 1979), which provides the number of studies needed to render a statistically significant effect of a meta-analysis insignificant. Because of its problematic assumptions and typically overly optimistic results, experts recommend abandoning failsafe *N* (e.g., Becker, 2005).

The funnel plot (Light & Pillemer, 1984) typically displays studies' effect sizes on the *x*-axis and their standard error or their precision (the inverse of a study's standard error) on the *y*-axis (Sterne & Egger, 2001). Figure 2.1 shows the (contour-enhanced) funnel plot of the meta-analysis of McCall and Carriger (1993; cf. Bakker et al., 2012). Funnel plot asymmetry, with a lower frequency of studies in the lower center of the plot corresponding to studies with a small and statistically insignificant effect size and a small sample size, is interpreted as an indication of publication bias. Hence the funnel plot in Figure 2.1 indicates that publication bias may have affected the results. However, funnel plots can also be asymmetric for other reasons (Sterne, Becker, & Egger, 2005). To overcome this interpretation problem, (Peters, Sutton, Jones, Abrams, & Rushton, 2008) developed the contour-enhanced funnel plot, which explicitly links the presence of studies in the funnel plot to their statistical (in)significance. The contour-enhanced funnel plot in Figure 2.1 suggests publication bias, since the asymmetry of the plot is linked to the statistical significance of the studies. Nonetheless, funnel plot methods are subjective, and many errors are made when identifying publication bias using the funnel plot (Terrin et al., 2003). Even experienced meta-analysts only correctly identified 52.5% of the cases in which a funnel plot was or was not affected by publication bias (Terrin, Schmid, & Lau, 2005). Two methods, Begg and Mazumdar's (1994) rank correlation and Egger's regression method (Egger et al., 1997; Sterne & Egger, 2005), formally test funnel plot asymmetry. Both methods test the association between studies' effect size and corresponding standard error, where a significant (typically positive) association signals publication bias. Because these methods have low statistical power (Borenstein et al., 2009; Sterne & Egger, 2005), both tests are usually applied using a significance level of .10. Due to their low power, their application is only recommended for meta-analyses based on at least 10 (Banks, Kepes, & Banks, 2012; Sterne & Egger, 2005) or even 15 effect sizes (Kepes, et al., 2012). Rothstein and Bushman (2012) also argued that the results of both tests are not meaningful if between-study heterogeneity in effect size is substantial. Finally, a clear limitation of both methods is that they can only be applied if there is reasonable variation in studies' sample size, with preferably at least a few samples with medium or large sample sizes (Borenstein, et al., 2009).

The trim-and-fill method developed by Duval and Tweedie (2000a, 200b) is another method for assessing publication bias on the basis of the funnel plot. It entails an iterative procedure that fills in missing studies that are needed to restore funnel plot symmetry, and provides an estimate of both the number of such missing studies

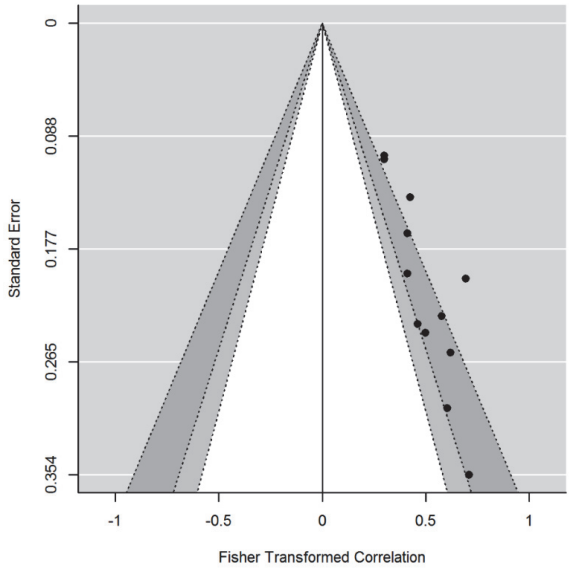

Figure 2.1. Contour-enhanced funnel plot of the meta-analysis of McCall and Carriger (1993). Areas represent studies with two-tailed *p*-values larger than .10 (white), smaller than .05 (light gray), smaller than .01 (dark gray), and smaller than .001 (light gray outside large triangle).

and the effect size. Duval and Tweedie (2000a, 2000b) developed three estimators (R_0, L_0, Q_0) for the number of missing studies. Estimators R_0 and L_0 perform better than Q_0, and L_0 is more robust than R_0 against the occurrence of a few aberrant studies (Duval and Tweedie, 2000a, 200b). L_0 is also used in most applications of the trim-and-fill method. Stated advantages of the trim-and-fill method are that it is relatively simple and provides an estimate of the effect size corrected for publication bias. However, the consensus is that the method should not be regarded as a way of yielding a more "valid" estimate of the overall effect size, but rather as a sensitivity analysis (Duval, 2005; Duval & Tweedie, 2000b; Viechtbauer, 2010). Results on the performance of the trim-and-fill method are mixed; some suggest the method is quite powerful and yields close to unbiased effect size estimates (Duval and Tweedie, 2000b), whereas others suggest it has low power to test the null hypothesis of no effect (Ferguson and Brannick, 2012). Agreement exists, however, that the method should not be used when population effect sizes are heterogeneous, because then it is likely to add non-existing studies (Rothstein & Bushman, 2012; Terrin, et al., 2003).

Ioannidis and Trikalinos (2007b) developed a test for publication bias based on a comparison between the observed (O) and expected (E) number of statistically significant studies in a meta-analysis. The expected number *E* is calculated as the sum of the studies' observed power, based on the effect size as estimated in the meta-

analysis: $E = \sum_{i=1}^{K}(1-\beta_i)$. The test for excess significance (TES) for publication bias is the common χ^2-test, with degrees of freedom equal to 1:

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

If the *p*-value of the test statistic is significant at .10, the test is interpreted as a signal of publication bias for a given meta-analysis. However, a statistically significant test outcome may also be the result of researcher *df* such as data peeking (Francis, 2012, 2013). Any process (publication bias or researcher *df*) leading to an abundance of statistically significant studies may be picked up by the TES. The TES has low power when only a limited number of studies is included in a meta-analysis (Francis, 2012, 2013; Ioannidis & Trikalinos, 2007b), and has particularly low power when population effects are heterogenous (Francis, 2013). Ioannidis and Trikalinos (2007) also recommend not using the test if between-study heterogeneity exists, but to first create homogenous subgroups of effect sizes before applying the test. Finally, the TES neither provides an answer to the question whether the population effect differs from zero, nor does it provide a (corrected) estimate of the effect.

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

routines are available (e.g., Vevea & Woods, 2005), it is unlikely that selection models will be used routinely in meta-analysis (Hedges & Vevea, 2005).

2.2 The *p*-uniform method

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

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

The basic idea of *p*-uniform is that the distribution of *p*-values conditional on the population effect size is uniform. This assumption is equivalent to the assumption underlying standard null hypothesis testing, with the important distinction that we now focus on the (conditional) p^{μ} – value distribution, which is the *p*-value distribution under the alternative hypothesis that the population effect size equals μ. *P*-uniform's effect size estimate will equal the effect size μ yielding a p^{μ} –value

distribution that is fitted best by a uniform distribution. *P*-uniform's test of the hypothesis of no effect is based on the deviation of the p^0 –value distribution from the uniform distribution, where the p^0-value distribution corresponds to the distribution of original *p*-values (i.e., corresponding to *p*-values of the test of no effect, or μ=0). *P*-uniform's test of publication bias is based on the deviation of the $p^{\hat{\mu}}$-value distribution from the uniform distribution, where $\hat{\mu}$ equals the effect size estimate of traditional fixed-effect meta-analysis.

We will explain effect size estimation and the two tests using an artificial example. The example is based on testing the hypothesis of no effect (μ=0) against the alternative of a positive effect ($\mu > 0$) with $\alpha = .05$. However, *p*-uniform can estimate and test any effect size measure. In the example, eighty studies with sample size 25 are generated using a fixed-effect model with μ=.33 and σ=1, where all statistically significant studies and 25% of insignificant studies are published. If each study tests the hypothesis of no effect (μ=0) against the alternative of a positive effect ($\mu > 0$) with $\alpha = .05$, then each study has a power of .5. Figure 2.2a shows the distribution of transformed *p*-values (p^0 –value distribution, or the distribution of *p*-values $\times$ 1/α) of the *K*=40 statistically significant studies of one simulation of the traditional example. Traditional fixed-effect meta-analysis carried out on all fifty published studies using the metafor package (Viechtbauer, 2010) yields a biased effect size estimate of 0.43 (SE = .063, $p < .001$).

Test of μ=0. If μ=0 then the p^0 –value distribution in Figure 2.2a should be close to the uniform distribution. Hence, *p*-uniform tests the hypothesis μ=0 by testing whether the observed p^0 –value distribution deviates from the uniform distribution. Fisher's (1925) method has been used before to test deviations from the uniform distribution. Notably, independently of us, Simonsohn et al. (2014b) have applied exactly the same test of μ=0 as we did. The first step of Fisher's method is to convert each *p*-value into numbers in the interval from 0 to 1 by computing the conditional probability of the *p*-value given its significance ($\alpha = .05$). The probability that a *p*-value is statistically significant is .05 if μ=0, hence all *p*-values are multiplied by 20 in the first step. Applying Fisher's (1925) method, if μ=0 then the test statistic $L^0 = -\sum_{i=1}^{K} \ln(20 p_i)$ is gamma distributed with *K* and 1 degrees of freedom, here denoted by $\Gamma(K,1)$. If the studies' *p*-values are generally small, as in Figure 2.2a, L^0 will be high. *P*-uniform rejects μ=0 whenever the value of L^0 is larger than the 95th percentile of the gamma distribution, denoted by $\Gamma_{.95}(K,1)$. In the example with *K*=40, $\Gamma_{.95}(40,1) = 50.94$. The null hypothesis is rejected since $L^0 = 82.26$ ($p < .001$); the population effect is larger than zero.

Test of publication bias. The test for publication bias by *p*-uniform amounts to a one-tailed test of the null hypothesis $\mu = \hat{\mu}$, i.e., whether the population effect size

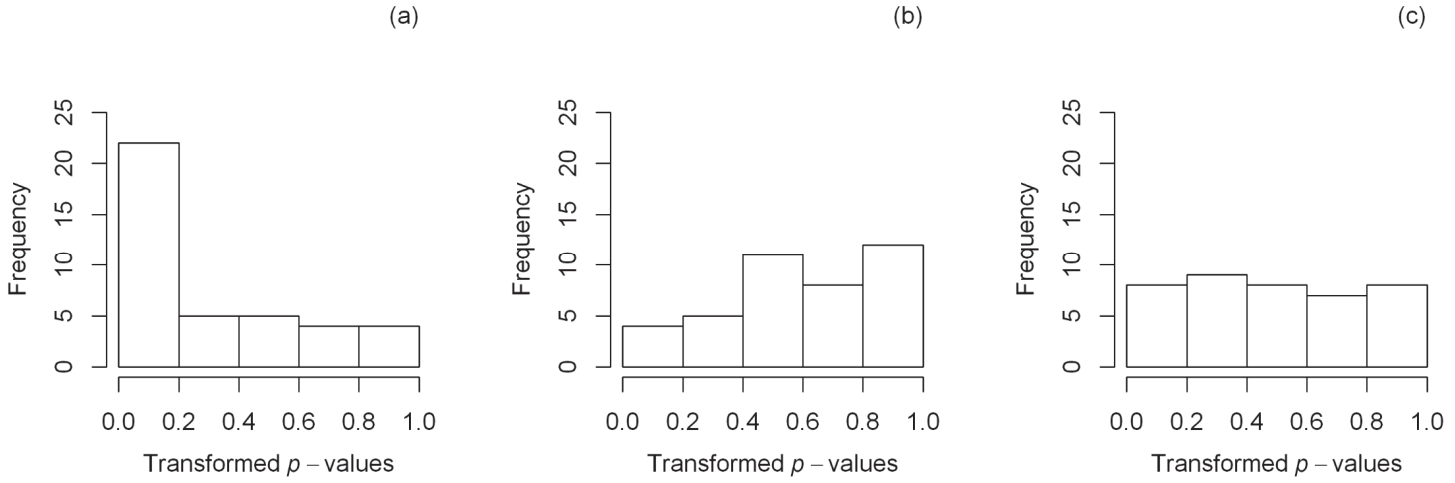

Figure 2.2. P-value distribution for (a) μ = 0, (b) μ = $\widehat{\mu}$, and (c) μ = $\widehat{\mu}^*$ as a function of the transformed significant *p*-values on the *x*-axis and its frequency on the *y*-axis.

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

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

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

Figure 2.2b depicts the distribution of $p_i^{\hat{\mu}}$, i.e., the $p^{\hat{\mu}}$-value distribution. The distribution is not uniform but skewed to the left with many high *p*-values, suggesting publication bias. The hypothesis of no publication bias is rejected if $L^{\hat{\mu}} < \Gamma_{.05}(40,1) = 30.2$, with $L^{\hat{\mu}} = -\sum_{i=1}^{K} \ln(p_i^{\hat{\mu}})$. Applying Fisher's test to the distribution of $p_i^{\hat{\mu}}$ yields $L^{\hat{\mu}} = 28.11$ ($p = .020$), indeed suggesting publication bias; the population effect is smaller than its value estimated by the traditional fixed-effect meta-analysis.

Interval and point estimation of μ. The 100(1– α)% confidence interval $\hat{\mu}_L^* \leq \mu \leq \hat{\mu}_U^*$ is obtained by $L^{\hat{\mu}_L^*} = \Gamma_{1\text{-}0.5\alpha}(K,1)$ and $L^{\hat{\mu}_U^*} = \Gamma_{0.5\alpha}(K,1)$. That is, each border of the interval is a value of μ for which the null hypothesis is only just accepted in a two-tailed test at significance level α. The probability that the null hypothesis is rejected that effect size equals μ is exactly .05, because (only) for μ is the *p*-value distribution exactly uniform. Consequently, this proofs that the interval estimate of *p*-uniform is unbiased: 95% of all confidence intervals contain μ, or the coverage probability of *p*-uniform is exactly .95. The borders of the confidence interval are easily obtained, since $L^{\hat{\mu}^*}$ decreases monotonically in $\hat{\mu}^*$.[2] The confidence interval in the example is 0.21 ≤

[2] *P*-uniform's point estimates and the bounds of its confidence interval are obtained by the R CRAN function uniroot. The input of uniroot is a function and an interval. It searches for a value in the interval for which the function equals zero. The functions were $L^{\hat{\mu}_L^*}$ –Γ1-0.5α(K,1), $L^{\hat{\mu}^*}$ – K, $L^{\hat{\mu}_U^*}$ –Γ0.5α(K,1).

$\mu \leq 0.43$. *P*-uniform's point estimate $\hat{\mu}^*$ equals the effect size yielding a $p^{\hat{\mu}^*}$ -value distribution that is fitted best by a uniform distribution. The point estimate is defined as the value of $\hat{\mu}^*$ for which L^{μ} equals *K*, which is the expected value of $\Gamma(K,1)$. In the example, $\hat{\mu}^*$ =.32. Figure 2.2c depicts the distribution of $p_i^{0.32}$. Note that 0.32 closely corresponds to μ = .33 used to generate the studies in this hypothetical example.

Alternative estimators in *p*-uniform: 1–*p*. The basic idea of *p*-uniform is that the *p*-value distribution conditional on the population effect size is uniform. However, the distribution of some transformation of *p*-values are then also uniform. For instance, if *p* is uniformly distributed, then so is 1–*p*. Consequently, we can also (i) test μ=0, (ii) test publication bias, and (iii) estimate $\hat{\mu}^*$, using 1– $p_i^{\mu^*}$ rather than $p_i^{\mu^*}$. The two estimators are differently sensitive to outliers, i.e., studies with extreme effect size estimates, where the estimator based on $p_i^{\mu^*}$ is very much affected by outliers, whereas the other is not. A very large effect size will yield a small $p_i^{\mu^*}$, hence a large – ln($p_i^{\mu^*}$), resulting in a large positive effect of that effect size on estimate $\hat{\mu}^*$. However, one very large effect size hardly affects $\hat{\mu}^*$ whenever the estimator based on 1– $p_i^{\mu^*}$ is used, because then –ln(1– $p_i^{\mu^*}$) approaches 0. To conclude, we expect the estimator based on 1– $p_i^{\mu^*}$ to be more robust to outliers and a violation of the homogeneity assumption than the estimator based on $p_i^{\mu^*}$. Properties of both estimators are examined in this study.

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

One assumption of *p*-uniform is that no questionable research practices were used in the studies, or, as (Simonsohn et al., 2014b) put it, "*p*-hacking" did not occur. *P*-hacking will typically result in *p*-values just below .05 (Simonsohn et al., 2014b). Because *p*-values close to .05 provide evidence for a low or even negative population effect size in *p*-uniform, *p*-hacking will in general result in an *under*estimation of the population effect size whenever *p*-uniform is applied. We consider this conservatism to be a positive quality of *p*-uniform; it will give estimates on the safe side, rather than traditional meta-analysis methods that *over*estimate population effect size because of *p*-hacking.

Another important assumption of *p*-uniform is that the population effect size

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

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

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

accurate estimates.

2.3 Method

All methods for assessing publication bias work for any effect size measure (cf. Borenstein et al., 2009). For illustrative purposes, we compare the methods in the most simple research situation. Effect sizes of studies were generated with a fixed population mean μ and standard deviation $\sigma = 1$ in all conditions, and a right-tailed test of the null hypothesis H_0: $\mu = 0$ was conducted in each individual study with $\alpha = 0.05$. The performance of *p*-uniform and other techniques for assessing publication bias were examined by means of Monte-Carlo simulations. In these simulations, equal sample sizes of 25 were imposed for each study in the meta-analysis. A sample size of 25 resembles the median cell size of 24 in both between- and within-subjects designs in experimental psychology observed by Wetzels et al. (2011).

Due to using equal sample sizes, not all available techniques for assessing publication bias can be applied. Neither the rank-correlation test, nor Egger's test can deal with equal sample sizes (e.g., Ioannidis & Trikalinos, 2007a), and were therefore excluded from the simulation study. The fixed-effect model was applied because studies' effect sizes were generated from the same population with one fixed mean. The trim-and-fill method was imposed to impute only studies in the left-hand side of the funnel plot because studies were tested for being significantly larger than zero. Two-tailed tests ($\alpha = 0.05$) were conducted for testing the effect size estimates obtained by the fixed-effect model, the trim-and-fill method, and *p*-uniform. The TES and *p*-uniform's publication bias test were also conducted two-tailed with an alpha level of 0.05; a 0.05 significance level rather than the more common 0.10 was selected to be consistent with the tests of effect size and its 95%-confidence interval.

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

Three parameters were varied in the main simulation study: the number of studies (*K*), the population effect size μ, and the proportion of statistically non-

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

We also ran an additional simulation study to examine the robustness of p-uniform to violations of the homogeneity assumption. Four cells of the design of the main simulation study were selected ($K/\mu = (0; 0.33) \times p_p = (0; 0.25)$), and heterogeneity was manipulated using three levels ($\tau^2 = (0.013333; 0.04; 0.12)$ in each

[3] Assume that the probability that an effect truly exists (P(H1)) is 0.5. Ioannidis (2005) used this value as starting point in his paper and argued that this value may be lower in fields with less confirmatory research. Also assume that the statistical power accompanied with the applied statistical test is 0.5 (using α = 0.05). Statistical power is often lower than the convention of 0.8. Bakker et al. (2012) even suggested that the typical power in psychological research is 0.35. These findings suggest that assuming a statistical power of 0.5 may even be liberal. The proportion of statistically significant studies in the literature (P('H1'| lit)) can then be found after entering values for p in the following equation:

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

where P('H0') is the proportion of statistically non-significant findings in the literature. For instance, the proportion of statistically significant findings in the literature if $p = 0.025$ is:

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

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

of these four cells. Parameter τ^2 represents the variance of true study means. The levels of τ^2 correspond to low (I^2=.25), moderate (I^2=.50), and high (I^2=.75) heterogeneity (Borenstein et al., 2009). The main dependent variables in the simulation were the point and interval estimates of traditional meta-analysis, the trim-and-fill method, and *p*-uniform.

To summarize, the main simulation study consisted of $K/\mu \times p_p = 4 \times 6 = 24$ conditions, whereas the additional simulation study had $K/\mu \times p_p \times \tau^2 = 2 \times 2 \times 3 = 12$ conditions. Simulations and *p*-uniform were programmed in R (R Core Team, 2017). The metafor package (Viechtbauer, 2010) was used for conducting the trim-and-fill method and fixed-effect (main simulation) and random-effects (small simulation) meta-analyses. See the supplementary information for the R code of our simulations (https://osf.io/drav5/).

2.4 Results

2.4.1 Estimation of effects when effects are homogenous

Convergence rates for the effect size estimates with *p*-uniform were above 98.9% and 96.3% across conditions for the $p_i^{\mu^*}$ and 1– $p_i^{\mu^*}$ estimator, respectively.[4] Table 2.1 shows average effect size estimates, standard errors or standard deviations of the effect size estimates, confidence intervals, and coverage probabilities of the fixed-effect model, the trim-and-fill method, and *p*-uniform. The performance of *p*-uniform was only evaluated as a function of the population effect size μ because the method does not take statistically non-significant studies into account. Coverage probabilities of *p*-uniform were 95% in all conditions for both estimators (see last row of Table 2.1), so exactly equal to the nominal coverage rates, confirming that *p*-uniform performs very well when its assumptions are satisfied. Figure 2.3 presents the average effect size estimates with the proportion of statistically non-significant studies included in the meta-analysis (p_p) on the *x*-axis, and on the *y*-axis the population effect size μ (horizontal dotted lines) and the average effect size estimates ($\widehat{\mu}$ and $\widehat{\mu}^*$). *P*-uniform's average effect size estimates are indicated by an asterisk (estimator $p_i^{\mu^*}$) and a cross (estimator 1– $p_i^{\mu^*}$) on the *y*-axis. *P*-uniform's average effect size estimate ($\widehat{\mu}^*$) had a slight negative bias for both estimators, which was significantly different from zero in some conditions. That is, bias of estimator $p_i^{\mu^*}$ for μ=0, μ=0.16, μ=0.33, μ=0.5 was −0.059 (−*z* = 5.9, *p* < .001), −0.048 (−*z* = 4.8, *p* < .001), −0.035 (−*z* = 3.5, *p* < .001), −0.018 (−*z* = 1.8, *p* = .065), respectively, and of estimator 1– $p_i^{\mu^*}$ it was −0.011 (−*z* = 1.1, *p* = .27), −0.020 (−*z* = 2.0, *p* = .046), −0.020 (−*z* = 2.0,

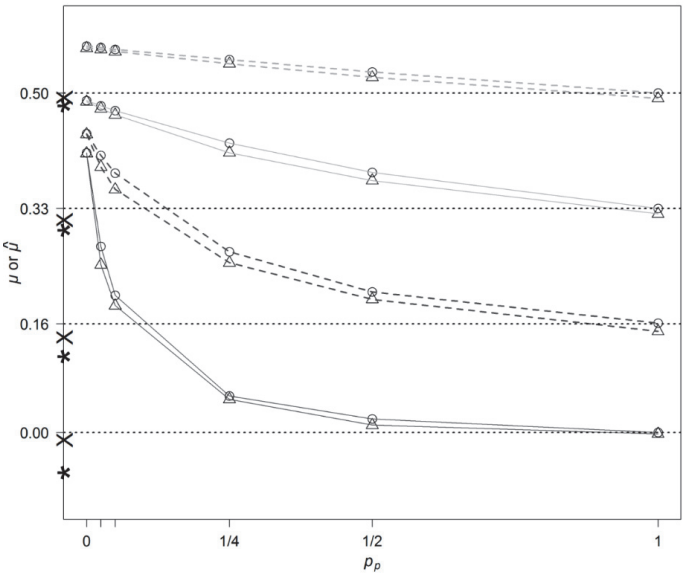

Figure 2.3. Average effect size estimates of the fixed-effect model, the trim-and-fill method, and *p*-uniform as a function of the proportion *p* of non-significant studies included in the meta-analysis and the population effect size μ. Average effect size estimates are indicated by open bullets (traditional fixed-effect model), triangles (trim-and-fill), asterisks (*p*-uniform estimator *p*), and crosses (*p*-uniform estimator 1–*p*). Dotted black lines illustrate the population effect size μ. Solid black lines refer to $\mu = 0$, dashed black lines refer to $\mu = 0.16$, solid gray lines refer to $\mu = 0.33$, and dashed gray lines refer to $\mu = 0.5$.

p = .046), −0.007 (−*z* = 0.7, *p* = .48), respectively.[4] Apparently, for the conditions in the simulations, the estimator $1-p_i^{\mu^*}$ slightly outperformed the estimator $p_i^{\mu^*}$.

Average effect size estimates of the fixed-effect model and the trim-and-fill method are presented as a function of p_p and population effect size μ using lines in Figure 2.3. Unsurprisingly, the fixed-effect model and the trim-and-fill method yielded accurate average effect size estimates in cases of no publication bias ($p_p = 1$). In particular, average effect size estimates obtained by the fixed-effect model (open bullets) fell exactly on the dotted lines reflecting the population effect size μ. Without publication bias ($p_p = 1$), the average effect size estimates of the trim-and-fill method (triangles in Figure 2.3) slightly underestimated the population effect size μ ($\widehat{\mu} =$ 0.49). This underestimation of the average effect size was caused by the imputation of

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

studies while no studies were missing. Table 2.2 shows the average number of studies imputed by the trim-and-fill method in each condition. The first row of the last column indicates that, on average, nine studies were imputed when there was no effect ($\mu = 0$) and no publication bias ($p_p = 1$), resulting in an underestimated effect. The other rows in Table 2.2 also illustrate the poor performance of the trim-and-fill method. If the proportion of statistically non-significant studies included in the meta-analysis (p_p) decreases, more studies are omitted from the meta-analysis and the trim-and-fill method should impute more studies. However, this is not the case because the trim-and-fill method hardly ever imputed studies if there was extreme publication bias (cf. $p_p = 0$, third column in Table 2.2).

In conditions with publication bias ($p_p < 1$), the fixed-effect model and the trim-and-fill method severely overestimated effect sizes. This is shown in Figure 2.3: As publication bias increased, the lines representing the fixed-effect model and the trim-and-fill method deviated more strongly from the population effect size μ. These average effect size estimates deviated more from the population effect size μ when there was at the same time no effect ($\mu = 0$) and extreme publication bias ($p_p = 0$), with $\widehat{\mu} = 0.41$ for both the fixed-effect model and the trim-and-fill method (see the first two rows of the first column in Table 2.1). If there was actually an effect in the population ($\mu > 0$), the overestimation in average effect sizes of both the fixed-effect model and the trim-and-fill method decreased in μ. The lines belonging to the fixed-effect model and the trim-and-fill method for $\mu = 0.5$ (dashed gray lines in Figure 2.3) diverged less from its population effect size μ than the lines belonging to both methods for $\mu = 0$ (solid black lines).

Coverage probabilities of both the fixed-effect model and the trim-and-fill method were far below the nominal 95% rate for $\mu < 0.5$ and whenever publication bias was present ($p_p < 1$). For conditions without an effect ($\mu = 0$) and extreme publication bias ($p_p = 0$), coverage probabilities were even close to 0 (see the first two rows of the first column in Table 2.1). Coverage probabilities became closer to the nominal rate as the effect increased and the amount of publication bias decreased. However, the coverage probability was still unsatisfactory in condition $\mu = 0.33$ and $p_p = .5$ for the fixed-effect model (0.88) and the trim-and-fill method (0.86). Coverage probabilities of both methods approached 95% when $\mu = 0.5$ and in conditions without publication bias ($p_p = 1$).

To conclude, coverage probabilities of *p*-uniform were 95% in all conditions, which did not apply to the fixed-effect model and the trim-and-fill method. Average effect size estimates of *p*-uniform were accurate, albeit slightly underestimated. Average effect size estimates of the fixed-effect model and the trim-and-fill method substantially deviated from the population effect size μ except for a medium size population effect ($\mu = 0.5$) and no publication bias ($p_p = 1$). At the same time, the standard deviations of *p*-uniform's effect size estimates were substantially

Table 2.1. Average effect size estimates and corresponding standard errors/standard deviations, confidence intervals, and coverage probabilities for the fixed-effect model, the trim-and-fill method, and *p*-uniform based on Monte-Carlo simulations of homogenous effects (10,000 replications)

			μ (K)			
			0 (160)	0.16 (40)	0.33 (16)	0.5 (10)
p_p	0	Fixed-effect model	0.412 (0.028) [0.267;0.557] CP: < .0001	0.440 (0.035) [0.295; 0.585] CP: .009	0.489 (0.045) [0.347; 0.631] CP: .359	0.569 (0.054) [0.429; 0.708] CP: .901
		Trim-and-fill	0.411 (0.028) [0.267; 0.556] CP: < .0001	0.439 (0.035) [0.295; 0.583] CP: .009	0.487 (0.044) [0.346; 0.629] CP: .362	0.566 (0.054) [0.428; 0.705] CP: .906
	1/40	Fixed-effect model	0.274 (0.071) [0.157; 0.392] CP: .026	0.408 (0.050) [0.271; 0.546] CP: .037	0.481 (0.047) [0.341; 0.621] CP: .402	0.567 (0.055) [0.427; 0.706] CP: .903
		Trim-and-fill	0.247 (0.066) [0.135; 0.358] CP: .035	0.391 (0.067) [0.259; 0.524] CP: .080	0.477 (0.051) [0.339; 0.616] CP: .415	0.564 (0.055) [0.426; 0.702] CP: .908
	1/20	Fixed-effect model	0.202 (0.066) [0.101; 0.304] CP: .074	0.382 (0.056) [0.251; 0.512] CP: .074	0.474 (0.049) [0.336; 0.612] CP: .443	0.564 (0.056) [0.426; 0.703] CP: .907
		Trim-and-fill	0.187 (0.062) [0.088; 0.286] CP: .092	0.358 (0.071) [0.233; 0.482] CP: .142	0.468 (0.057) [0.331; 0.604] CP: .464	0.561 (0.056) [0.424; 0.699] CP: .910
	1/4	Fixed-effect model	0.054 (0.035) [-0.004; 0.112] CP: .542	0.266 (0.056) [0.166; 0.365] CP: .434	0.426 (0.056) [0.300; 0.551] CP: .696	0.549 (0.059) [0.414; 0.684] CP: .928
		Trim-and-fill	0.049 (0.037) [-0.009; 0.107] CP: .577	0.250 (0.054) [0.153; 0.347] CP: .533	0.412 (0.066) [0.289; 0.534] CP: .726	0.543 (0.062) [0.410; 0.677] CP: .926

Table 2.1. Continued

		μ (K)			
		0 (160)	0.16 (40)	0.33 (16)	0.5 (10)
1/2	Fixed-effect model	0.020 (0.024) [-0.023; 0.063] CP: .833	0.207 (0.044) [0.127; 0.288] CP: .772	0.383 (0.056) [0.270; 0.497] CP: .859	0.531 (0.061) [0.400; 0.662] CP: .946
	Trim-and-fill	0.011 (0.032) [-0.031; 0.053] CP: .791	0.196 (0.046) [0.117; 0.275] CP: .821	0.371 (0.061) [0.259; 0.482] CP: .875	0.523 (0.065) [0.394; 0.652] CP: .937
1	Fixed-effect model	0.000 (0.016) [-0.031;0.031] CP: .949	0.161 (0.032) [0.099; 0.223] CP: .948	0.330 (0.050) [0.232; 0.428] CP: .952	0.500 (0.063) [0.376; 0.624] CP: .951
	Trim-and-fill	-0.020 (0.030) [-0.050; 0.011] CP: .634	0.149 (0.039) [0.088; 0.209] CP: .869	0.322 (0.053) [0.225; 0.418] CP: .926	0.492 (0.066) [0.370; 0.615] CP: .932
p-uniform	Estimator *p*	-0.059 (0.224) [-0.427; 0.313] CP: .952	0.112 (0.187) [-0.224; 0.418] CP: .950	0.298 (0.142) [0.031; 0.539] CP: .951	0.481 (0.103) [0.282; 0.677] CP: .952
	Estimator 1–*p*	-0.011 (0.271) [-0.468; 0.367] CP: .948	0.140 (0.240) [-0.307; 0.459] CP: .949	0.313 (0.188) [-0.094; 0.564] CP: .952	0.493 (0.137) [-0.183; 0.686] CP: .947

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

Table 2.2. Average number of imputed studies by the trim-and-fill method based on Monte-Carlo simulations (10,000 replications)

		p_p					
		0	1/40	1/20	1/4	1/2	1
μ (K)	0 (160)	0.06	1.41	0.82	0.59	2.01	9.00
		(0.25)	(1.45)	(1.18)	(1.84)	(4.84)	(12.06)
	0.16 (40)	0.05	0.82	1.17	0.90	0.81	1.49
		(0.25)	(1.36)	(1.48)	(1.33)	(1.52)	(2.74)
	0.33 (16)	0.07	0.17	0.27	0.61	0.61	0.49
		(0.28)	(0.59)	(0.74)	(1.07)	(1.09)	(1.08)
	0.5 (10)	0.10	0.11	0.12	0.21	0.27	0.30
		(0.36)	(0.39)	(0.50)	(0.59)	(0.68)	(0.72)

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

larger than those of the fixed-effect model and the trim-and-fill method. As a consequence, average effect size estimates of *p*-uniform were accurate but more uncertain. In contrast, the results of the fixed-effect model and the trim-and-fill method provided false certainty. These estimates were precise but highly inaccurate if the population effect size μ was smaller than medium ($\mu < 0.5$) and publication bias was present ($p_p < 1$).

2.4.2 Test of an effect when effects are homogenous

In Table 2.3, Type I error rates and statistical power of the fixed-effect model, the trim-and-fill method, and estimator $p_i^{\mu^*}$ of *p*-uniform are presented for testing whether the population effect size equals 0. *P*-uniform's Type I error rates were exactly equal to the nominal rate in all conditions (see third row of the last column in Table 2.3). Statistical power of *p*-uniform increased in μ from 0.26 for $\mu = 0.16$ to 0.98 for $\mu = 0.5$. Consequently, *p*-uniform already has very high power to detect a medium effect size ($\mu = d = 0.5$) when only eight studies with $n = 25$ are statistically significant.

If there was no effect ($\mu = 0$) and no publication bias ($p_p = 1$), the Type I error rate of the trim-and-fill method was lower than the nominal rate ($\alpha = 0.035$). This was caused by the imputation of studies while no publication bias was present (see Table 2.2). If there was publication bias ($p_p < 1$) the Type I error rates were grossly overestimated by the fixed-effect model and the trim-and-fill method. The Type I error rates increased as publication bias became more severe. If there was no effect ($\mu = 0$) and extreme publication bias ($p_p = 0$), Type I error rates of the fixed-effect model and the trim-and-fill method equaled 1 (see first two rows of the first column in Table 2.3) meaning that both methods always yielded a Type I error in this condition. This Type I

error rate was severely inflated due to overestimated average effect size estimates by both methods as explained in the previous section (see also Table 2.1 and Figure 2.3).

The fixed-effect model and the trim-and-fill method were powerful in detecting an effect when it truly existed ($\mu > 0$) and no publication bias was present ($p_p = 1$). The levels of statistical power rapidly approached one for these conditions (see last column in Table 2.3). If there was an effect ($\mu > 0$) and publication bias was present ($p_p < 1$), the statistical power of the fixed-effect model and the trim-and-fill method was close to 1 or equaled 1 in every condition. However, these results reflect false certainty because effect size estimates of both methods were overestimated due to the presence of publication bias (see previous section).

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

2.4.3 Publication bias test when population effects are homogenous

Table 2.4 shows Type I error rates and statistical power of two publication bias tests: estimator $p_i^{\mu^*}$ of *p*-uniform and the TES. Type I error rates of *p*-uniform were close to 5% in the conditions $\mu < 0.5$ without publication bias ($p_p = 1$) (see last column in Table 2.4). With $\mu = 0.5$ and without publication bias ($p_p = 1$), Type I error rates obtained by *p*-uniform were lower than the nominal rate ($\alpha = 0.012$). *P*-uniform had reasonable statistical power when a considerable number of studies had been excluded from the meta-analysis. For example, statistical power of the method was 0.75 for $\mu = 0.16$ and extreme publication bias ($p_p = 0$) (see fourth row of first column in Table 2.4).

The last column in Table 2.4 illustrates that in conditions without publication bias ($p_p = 1$) the TES was more conservative than *p*-uniform. Type I error rates of the TES ranged from 0.022 for no effect ($\mu = 0$) to 0.003 for $\mu = 0.5$. With one exception, the TES was less powerful than *p*-uniform in detecting publication bias. This exception was that the TES had more power if no effect existed ($\mu = 0$) and at least some statistically non-significant studies were published ($p_p > 0$). *P*-uniform had more statistical power to detect publication bias if there was no effect ($\mu = 0$) and extreme publication bias ($p_p = 0$), and if an effect indeed existed ($\mu > 0$).

The statistical power of the TES and *p*-uniform was low for the two largest population effect sizes ($\mu = 0.33$ and $\mu = 0.5$). For example, for $\mu = 0.5$ the statistical power was not higher than 0.03 for *p*-uniform and 0.001 for the TES. The statistical power of *p*-uniform was low for two reasons. First, few studies were statistically

significant (eight on average) resulting in a wide confidence interval for the average effect size estimate. Second, few studies were not statistically significant (on average two for $\mu = 0.5$ or eight for $\mu = 0.33$), such that the average effect size estimate based on all studies was close to the average effect size estimate based on only the statistically significant studies. In conditions where only few studies were omitted from the meta-analysis, which occurred when the population effect size or a study's power is high, publication bias was hard to detect.

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

Table 2.3. Results of Monte-Carlo simulations (10,000 replications) on Type I error rates and statistical power for testing whether the effect size is significantly different from zero

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

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

Table 2.4. Results of Monte-Carlo simulations (10,000 replications) on Type I error rates and statistical power for publication bias tests

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

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

2.4.4 Estimation of effects when population effects are heterogeneous

Here we study the performance of the methods under violations of a homogeneous population effect. Convergence rates for the effect size estimates with *p*-uniform were above 98.3% and 99.2% across conditions for the $p_i^{\mu^*}$ and 1– $p_i^{\mu^*}$ estimator, respectively[5]. Table 2.5 shows average effect size estimates, standard errors or standard deviations of the effect size estimates, and coverage probabilities of the random-effects model (with the most frequently used DerSimonian Laird procedure), the trim-and-fill method, and *p*-uniform. We compare the results of the three methods to each other, but also compare them to the results of these methods when effects are homogenous (Table 2.1). First, note how introducing heterogeneity increases the number of significant studies from 8 when effect size is homogenous or μ=.33, to 32.8 when heterogeneity is high and μ=0 (second row of Table 2.5). Consequently, *p*-uniform uses relatively more than 5% (up to about 20%) of the studies if no effect exists and effects are heterogeneous.

From the results of Table 2.5 and comparing its results to those in Table 2.1, it follows that random-effects meta-analysis and the trim-and-fill method perform worse as heterogeneity increases; both bias increases and the coverage probability

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

decreases in heterogeneity. Moreover, the estimate of heterogeneity (τ^2) is biased in random-effects meta-analysis as well; e.g., τ^2 is severely underestimated if only statistically significant studies are published, whereas τ^2 is grossly overestimated if 25% of the statistically insignificant studies are published (not shown in Table 2.5). The trim-and-fill method on average imputed less than .1 studies if only statistically significant studies are published (also when about 130 or more studies were omitted), and up to 6.3 studies when 25% of the statistically insignificant studies are published and no effect exists (when on average about 95 studies were omitted) (not shown in Table 2.5). To conclude, the performance of random-effects meta-analysis and the trim-and-fill method is bad in case of publication bias and worsens when heterogeneity increases.

Whereas the performance of *p*-uniform is excellent when effects are homogenous (Table 2.1), performance worsens when heterogeneity increases; both bias increased and the coverage probability decreased in heterogeneity (Table 2.5). As expected, estimator 1– $p_i^{\mu^*}$ is more robust to heterogeneity than estimator $p_i^{\mu^*}$. However, in our opinion the performance of 1– $p_i^{\mu^*}$ is only acceptable when heterogeneity is low, with coverage probabilities of .895 and .926 and bias of .086 and .047 for μ=0 and μ=.33, respectively. Both *p*-uniform estimators outperformed traditional random-effects meta-analysis and the trim-and-fill method under conditions of heterogeneity when statistically insignificant studies are not published (p_p = 0), but not when p_p = 0.25. This suggests that if effects are heterogeneous, *p*-uniform only outperforms the other methods when publication bias is extreme (with p_p close to 0). To conclude, *p*-uniform is generally not robust to heterogeneous effects, only provides acceptable estimates if heterogeneity is low, and outperforms other methods only if publication bias is extreme under conditions of heterogeneity.

2.5 Application to meta-analysis of McCall and Carriger (1993)

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

The apparent negative association between effect size and standard error in the contour-enhanced funnel plot (Figure 2.1) suggests publication bias. This is confirmed by both Begg and Mazumdar's rank-correlation test (τ = 0.636, p = .003)

and Egger's test (z = 2.24, p = .025). The TES also provides evidence for the presence of publication bias (χ^2 = 6.22, p = .013). The funnel plot after application of the trim-and-fill technique using statistic L_0 is presented in Figure 2.4. Six studies were imputed to the left. Trim-and-fill's estimate of the Fisher-transformed correlation was .35 (p < .001), corresponding to an estimated correlation of .34 (CI 95%: .26, .41). Based on the R_0 statistic, nine studies were imputed reducing the Fisher-transformed correlation to 0.31 (p < .001). The untransformed correlation coefficient based on the R_0 statistic became .30 (CI 95%: .23, .37). Hence, the trim-and-fill method reduced the estimated correlation somewhat for both statistics (from .39 to .34 for L_0 and .30 for R_0), but still suggested a significant and medium correlation.

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

2.6 Discussion

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

[6] All test statistics of *p*-uniform are compared to a gamma distribution with df1=1 and df2=11.

Table 2.5. Average effect size estimates and corresponding standard errors/standard deviations, and coverage probabilities for the random-effects model, the trim-and-fill method, and *p*-uniform based on Monte-Carlo simulations of heterogeneous population effects (10,000 replications)

		μ = 0, *K* = 160			μ = 0.33, *K* = 16		
	Heterogeneity (τ^2)	Low (0.0133)	Mod. (0.04)	High (0.12)	Low (0.0133)	Mod. (0.04)	High (0.12)
$p_p = 0$	# significant studies	12.33 (3.35)	19.62 (4.11)	32.80 (5.16)	8.04 (2.01)	8.02 (2.00)	8.02 (2.00)
	Random-effects	0.433 (.059)	0.469 (.046)	0.554 (.036)	0.514 (.073)	0.554 (.075)	0.648 (.075)
	model	CP < .0001	CP = < .0001	CP = < .0001	CP = .225	CP = .107	CP = .035
	Trim-and-fill	0.433 (.059)	0.469 (.046)	0.554 (.036)	0.512 (.073)	0.553 (.075)	0.645 (.089)
		CP < .0001	CP < .0001	CP < .0001	CP = .229	CP = .112	CP = .039
$p_p = 0.25$	Random-effects	0.081 (.039)	0.126 (.044)	0.211 (.054)	0.447 (.072)	0.473 (.081)	0.532 (.097)
	model	CP = .441	CP = .185	CP = .026	CP = .604	CP = .535	CP = .505
	Trim-and-fill	0.071 (.039)	0.099 (.045)	0.145 (.055)	0.428 (.071)	0.449 (.081)	0.497 (.111)
		CP = .539	CP = .404	CP = .268	CP = .655	CP = .594	CP = .570
p-uniform		0.091 (.160)	0.262 (.102)	0.503 (.075)	0.367 (.137)	0.464 (.137)	0.641 (.153)
	Estimator *p*	CP = .827	CP = .223	CP < .0001	CP = .887	CP = .644	CP = .206
		0.086 (.219)	0.228 (.141)	0.406 (.080)	0.357 (.187)	0.428 (.173)	0.535 (.163)
	Estimator 1–*p*	CP = .895	CP = .593	CP = .045	CP = .926	CP = .840	CP = .610

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

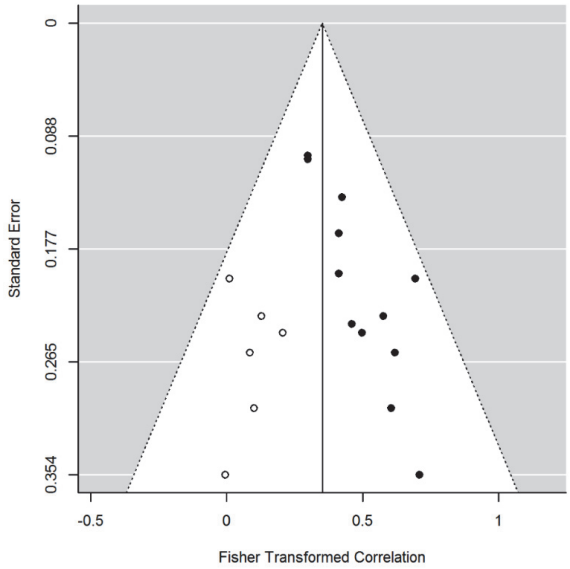

Figure 2.4. Funnel plot of the studies of McCall and Carriger's (1993) meta-analysis after the trim-and-fill method imputed six studies (open circles) based on the L_0 statistic. The vertical line corresponds to trim-and-fill's effect size of 0.352.

examining the performance of *p*-uniform were selected, with small numbers of studies included in the meta-analysis and small sample sizes for each individual study.

Results of the main simulation study on homogenous population effect sizes showed good statistical properties of *p*-uniform in comparison to the trim-and-fill method, TES, and standard fixed-effects meta-analysis. Coverage probabilities of *p*-uniform were always 95%, whereas *p*-uniform's slightly underestimated the population effect. Our results and those of others (Moreno, Sutton, Ades, et al., 2009; Peters et al., 2007; Terrin et al., 2003) clearly show that the fixed-effect model and the trim-and-fill method cannot be trusted when there is publication bias. The average effect size estimates and coverage probabilities of existing methods were only acceptable in the absence of publication bias ($p_p = 1$) or sufficient power in the primary studies (.80 for $\mu = 0.5$). For testing whether the population effect is zero, the Type I error rate of *p*-uniform was exactly equal to the nominal rate, and *p*-uniform's statistical power was high to detect a population effect of medium size. The fixed-effect model and the trim-and-fill method yielded too many Type I errors if publication bias was present. Both *p*-uniform and the TES for the presence of publication bias were too conservative. However, *p*-uniform's publication bias test outperformed the TES in most conditions of homogenous population effects. An additional simulation study on heterogeneous population effects revealed that both *p*-uniform and other fixed-effects techniques performed poorly under increasing

heterogeneity. Our transformed estimator 1– $p_i^{\mu^*}$ was more robust to heterogeneity than estimator $p_i^{\mu^*}$, but its performance was only acceptable if heterogeneity was low. However, the transformed estimator did outperform other fixed-effect techniques when publication bias was extreme.

P-uniform did not converge to an effect size estimate in a small percentage of the simulations (< 2%) when no effect existed. The reason of the non-convergence is the small number of studies in combination with the distribution of *p*-values under the null hypothesis of no effect; *p*-uniform sometimes cannot estimate μ if all *p*-values are higher than .025 and close to .05. Since this is unlikely as *K* increases, the non-convergence problem quickly disappears if *K* increases. For instance, *p*-uniform's convergence rates were all above 99.9% if the number of studies was twice as large as in the conditions with homogeneous population effects, with 16 rather than 8 expected statistically significant studies.

The effect size estimates of both estimators *p*-uniform based on Fisher's method (Fisher, 1925) were slightly negatively biased. The negative bias is a consequence of the estimate $\widehat{\mu}$ being a nonlinear function of *p*. We first examined the bias for estimating $\widehat{\mu}$ on the basis of one single statistically significant study. The expected value of $\widehat{\mu}$ turned out negative because *p*-values close to .05 yielded very negative estimates of μ. The negative bias decreases in the study's sample, with factor $\sqrt{N}$, and in population effect size μ. Additional simulations, with on average twice as many statistically significant studies in a meta-analysis (16 instead of 8), suggested that the bias also decreases in the number of statistically significant studies whenever effect size is larger than zero, although the bias did not disappear entirely. Future studies should consider examining systematically the performance of other statistical tests for uniformity than those based on Fisher's method (such as *p*-uniforms $p_i^{\mu^*}$ and 1- $p_i^{\mu^*}$ estimator) to assess whether this decreases bias in effect size estimates and provide lower standard errors than estimates obtained with the Fisher's method. Statistical tests that can be used for this purpose are the Kolmogorov-Smirnov test (Massey, 1951), and the Anderson-Darling test (T. W. Anderson & Darling, 1954).

The newly proposed *p*-uniform method has numerous advantages over existing techniques in examining and correcting for publication bias. First of all, it is the first method that can provide an effect size estimate, test whether the population effect is zero, and test for publication bias at the same time. An important second advantage of *p*-uniform is that, even though power may be low for testing publication bias in applications with a small number of studies, the average effect size is accurately estimated by *p*-uniform when its assumptions are satisfied. When there is publication bias and effects are homogenous, *p*-uniform has good statistical properties compared to fixed-effect meta-analyses, the TES, and the trim-and-fill method. Our study did not compare *p*-uniform's performance to that of Egger's and the rank

correlation test. However, since other studies (e.g., Moreno et al., 2009) showed that the latter two methods had low power for the conditions with eight studies examined in our simulation study, it is likely *p*-uniform also outperforms them. Third, no sophisticated assumptions or choices have to be made when applying *p*-uniform. No additional (unpublished) data have to be collected and interpretation of the results is straightforward. Hence, in principle, meta-analysts should be able to easily apply *p*-uniform in their research. We are currently working on developing a website that will have R programs enabling researchers to apply *p*-uniform to their research. Finally, *p*-uniform will provide conservative effect size estimates in case of researcher *df*, rather than further overestimating effect size.

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

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

perform just as well when studies vary in sample size; in conditions with studies varying in sample size the performance of *p*-uniform can then also be compared to Egger's test and the rank correlation test. Finally, following others (Banks, Kepes, & Banks, 2012; Banks, Kepes, & McDaniel, 2012; McDaniel, Rothstein, & Whetzel, 2006), we recommend conducting publication bias analyses in both past and future meta-analytic studies. Moreover, following Banks, Kepes, and Banks (2012), we encourage journals to publish re-evaluations of previous meta-analytic reviews regardless of their results to avoid 'publication bias in publication bias results'.

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

CHAPTER 3

Conducting meta-analyses based on *p*-values: Reservations and recommendations for applying *p*-uniform and *p*-curve

Abstract

Because evidence of publication bias in psychology is overwhelming, it is important to develop techniques that correct meta-analytic estimates for publication bias. Van Assen et al. (2015) and Simonsohn et al. (2014a) developed *p*-uniform and *p*-curve, respectively. The methodology on which these methods are based has great promise for providing accurate meta-analytic estimates in the presence of publication bias. However, we show that in some situations *p*-curve behaves erratically while *p*-uniform may yield implausible negative effect size estimates. Moreover, we show that (and explain why) *p*-curve and *p*-uniform overestimate effect size under moderate to large heterogeneity, and may yield unpredictable bias when researchers employ *p*-hacking. We offer hands-on recommendations on applying and interpreting results of meta-analysis in general and *p*-uniform and *p*-curve in particular. Both methods as well as traditional methods are applied to a meta-analysis on the effect of weight on judgments of importance. We offer guidance for applying *p*-uniform or *p*-curve using R and a user-friendly web application for applying *p*-uniform (https://rvanaert.shinyapps.io/p-uniform).

This chapter is published as van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses 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

Meta-analysis is the standard technique to synthesize effect sizes of several studies on the same phenomenon. A well-known problem of meta-analysis is that effect size can be overestimated because of publication bias (e.g., Ioannidis, 2008b; Lane & Dunlap, 1978). Publication bias is here defined as the tendency of studies with statistically significant results to be published at a higher rate than studies with results that are not statistically significant. Because evidence of publication bias is overwhelming across many scientific disciplines (Fanelli, 2012), it is important to develop techniques that correct the meta-analytic estimate for publication bias (Moreno, Sutton, Ades, et al., 2009). Recently, van Assen, van Aert, and Wicherts (2015) and Simonsohn et al. (2014a) have independently developed methods aiming to provide an accurate meta-analytic estimate in the presence of publication bias. Their methods, *p*-uniform and *p*-curve, respectively, both make use of the distribution of statistically significant results yet differ in implementation. The goals of this chapter are to introduce and explain both methods and their differences, to provide straightforward recommendations for applying meta-analysis, and to formulate guidelines for applying and interpreting results of *p*-uniform and *p*-curve.

3.1 A primer on *p*-uniform and *p*-curve

Simonsohn et al. (2014b) described how statistically significant *p*-values of studies on an effect could be used to test this effect against the null hypothesis that the effect equals zero. This idea was not new; Fisher (1925) already developed a method for testing the null hypothesis of no effect by means of combining *p*-values. However, the novelty of *p*-curve lies with its use of only the statistically significant *p*-values, which are arguably not affected by publication bias. The method was called *p*-curve because it analyzed the curve or distribution of *p*-values. The logic of *p*-curve suggests that there is no effect in the studies in the meta-analysis if the *p*-values are uniformly distributed (i.e., *p*-curve is flat), whereas there is an effect whenever the *p*-value distribution or *p*-curve is right-skewed (Hung, O'Neill, Bauer, & Köhne, 1997).

A disadvantage of *p*-curve at that time was that effect size could not be estimated. Van Assen et al. (2015) developed another method analyzing statistically significant *p*-values, called *p*-uniform, which is able to estimate the effect size in a set of studies. Van Assen et al. (2015) called their method *p*-uniform, because the effect size estimate is equal to the value for which the *p*-value distribution conditional on that value is uniform (as we explain below). Besides estimating the effect size, *p*-uniform can also estimate a confidence interval around the effect size estimate, includes a test of publication bias, and, similar to *p*-curve (Simonsohn et al., 2014a), tests the null hypothesis of no effect. Simonsohn et al. (2014a) later extended *p*-curve to also estimate effect size. However, *p*-curve neither provides a confidence interval nor a test for publication bias. In the present study, we will focus on effect size estimation by both *p*-curve and *p*-uniform.

Van Assen et al. (2015) and Simonsohn et al. (2014a) convincingly illustrated

the strengths of *p*-uniform and *p*-curve and the logic upon which it is based for carrying out meta-analysis. They showed that the methods provide accurate effect size estimates in the presence of publication bias, even when the number of statistically significant studies is small. Similarly, both methods were found to perform well when studies have the same sample sizes, when studies differ in sample size, and in the scenario where there is (small) heterogeneity of effect size (i.e., when the underlying (population) effect sizes actually differ between studies in the meta-analysis). Moreover, results of Simonsohn et al. (2014a) suggested that *p*-hacking, or the original researcher's use of strategies to achieve statistical significance (Simmons et al., 2011), leads to an underestimation of effect size in analyses based on *p*-curve, whereas it leads to overestimation of effect size in traditional meta-analysis (Bakker et al., 2012).

3.2 Three reservations

Although we are convinced of the potential and validity of the logic of *p*-uniform and *p*-curve, we add three important reservations to the application of the methods and the general methodology in its current state. More specifically, we first demonstrate that *p*-uniform and *p*-curve may yield implausible negative (*p*-uniform) or inaccurate (*p*-curve) estimates in meta-analyses with *p*-values close to the significance level (considered equal to .05 in the present chapter). Second, we explain why and show that *p*-hacking does not always cause *p*-curve's and *p*-uniform's effect sizes to be underestimated as was stated in Simonsohn et al. (2014a). Finally, we show that, in contrast to the results in Simonsohn et al. (2014a), *p*-uniform and *p*-curve cannot deal with a substantial amount of heterogeneity (i.e., there is no single true effect size underlying the studies in the meta-analysis, but rather a distribution of true effect sizes). Based on our explanation of the methods and the reservations, we formulate recommendations for applying meta-analysis in general and interpreting results of *p*-uniform and *p*-curve in particular. These hands-on recommendations are summarized in Table 3.1. Scientists who consider using these methods have to be aware of conditions in which the methods should not be interpreted, or interpreted with caution.

In the remainder of the paper, we illustrate major issues involved in applying *p*-curve and *p*-uniform by considering a recent meta-analysis of studies on the effect of weight on judgment of importance (Rabelo et al., 2015). We will briefly describe other meta-analysis methods using statistically significant effect sizes, introduce the basic idea underlying *p*-uniform and *p*-curve, and illustrate the logic of and computations in of *p*-uniform and *p*-curve in Appendix A. The analyses that form the basis of our three reservations and recommendations are presented in the next sections. Readers who do not want to delve into the (technical) details of *p*-uniform and *p*-curve can skip these sections and move over to the Discussion and Conclusion section, where we explain the recommendations in Table 3.1. R code of all our analyses is available at https://osf.io/5nk4y/.

Table 3.1. Recommendations for meta-analysis and applying *p*-uniform and *p*-curve

Recommendations for meta-analysis and applying *p*-uniform and *p*-curve:
1) Check for evidence of *p*-hacking in the primary studies
→ In case of strong evidence or strong indications of *p*-hacking, be reluctant with interpreting estimates of traditional meta-analytic techniques and *p*-uniform and *p*-curve, because their effect size estimates may be biased in any direction depending on the type of *p*-hacking.
2) Apply fixed-effect and random-effects meta-analysis, as well as *p*-uniform or *p*-curve, and report their results conforming to the Meta-Analysis Reporting Standards (MARS; American Psychological Association, 2010)
3) Check for direct or indirect evidence of publication bias
→ In case of evidence of publication bias, interpret results of *p*-uniform or *p*-curve rather than of fixed-effect and random-effects meta-analysis; in the absence of such evidence, interpret results of fixed-effect and random-effects meta-analysis
4) Set the effect size estimate of *p*-uniform or *p*-curve equal to zero if the average *p*-value of the statistically significant studies is larger than .025
5a.) If effect size is homogenous or heterogeneity small to moderate ($I^2 < 0.5$), interpret the estimate of *p*-uniform and *p*-curve as estimates of the average *population* effect size; otherwise they overestimate average *population* effect size and should be interpreted as estimates of the average true effect size of only the set of statistically significant studies 5b) In case of substantial heterogeneity and if desired, create homogeneous subgroups of primary studies based on theoretical or methodological considerations in order to estimate with *p*-uniform and *p*-curve the average *population* effect size underlying the studies in each subgroup

3.3 Example

Rabelo et al. (2015) conducted a meta-analysis on the effect of weight on judgments of importance. The theory underlying the studies included in the meta-analysis is that the physical experience of weight (e.g., holding a heavy object) influences how much importance people assign to things, issues, and people (IJzerman, Padiotis, & Koole, 2013; Jostmann, Lakens, & Schubert, 2009). For instance, in their second study, Jostmann et al. (2009) found that participants who held a heavy clipboard attributed more importance to fairness in decision-making as opposed to participants holding a light clipboard. Table 3.B1 in the appendices provides the full references, sample sizes (n_i^1 and n_i^2), *t*-values, and *p*-values from the 25 studies of this kind published in the embodiment literature.

According to the first recommendation, we should consider the presence of *p*-hacking in the primary studies included in the meta-analysis. We believe that the

studies on the link between weight and importance are mostly studies in which the specifics of the analysis are often neither preregistered nor clearly restricted by theory. Hence, according to Recommendation 1, we would use caution in interpreting the current results and await new (preferably pre-registered) studies in this field.

Four different meta-analytic estimates of the (mean) effect size underlying the weight-importance studies are presented in Table 3.2. In line with Recommendation 2, we first fitted traditional fixed-effect and random-effects meta-analysis. Both analyses yielded the same effect size estimate of 0.571 (95% confidence interval: [0.468;0.673]), which is highly statistically significant (z = 10.90, p < .001) and suggests a medium to large effect of the experience of weight on how much importance people assign to things (see Table 3.2). *P*-uniform's publication bias test suggested that there is evidence for publication bias (z = 5.058, p < .001), so we should interpret the results of *p*-uniform or *p*-curve rather than the standard meta-analytic estimates (Recommendation 3). Because the average *p*-value of the 23 statistically significant studies equals .0281, we set the effect size estimate of *p*-uniform and *p*-curve equal to 0, in line with Recommendation 4. When not setting the estimate to 0, applying *p*-curve and *p*-uniform yields a nonsignificant *negative* effect size (see Table 3.2), and *p*-uniform's 95% confidence interval (-0.676; 0.160) suggests that the effect size is small at best.

Table 3.2. Results of *p*-uniform, *p*-curve, fixed-effect meta-analysis (FE MA), and random-effects meta-analysis (RE MA) when applied to the meta-analysis reported in Rabelo et al. (2015) of the effect of weight on the judgment of importance in the moral domain.

	p-uniform	*p*-curve	FE MA	RE MA
Effect size estimate	-0.179	-0.172	0.571	0.571
95% CI	(-0.676;0.160)	-	(0.468;0.673)	(0.468;0.673)
Test of H_0: $\delta = 0$	z=0.959; p=.831	$\chi^2(46)$=55.833 p=.848	z=10.904; p<.001	z=10.904; p<.001
Pub. bias test	z=5.058; p<.001	-		-

The null hypothesis of no heterogeneity among the included studies was not rejected ($Q(24) = 4.55$, $p = 1$, $I^2=0$), which suggests that *p*-uniform and *p*-curve may accurately estimate the average population effect size (Recommendation 5a). Note that due to the absence of heterogeneity, effect size estimates of fixed-effect and random-effects meta-analysis were identical. Although the lack of heterogeneity

suggests that the effects are homogeneous, in this particular instance, homogeneity is excessive (with a *p*-value of the *Q*-test very close to 1). Such excessive homogeneity is unlikely to occur under normal sampling conditions (Ioannidis, Trikalinos, & Zintzaras, 2006) and could be caused by publication bias (Augusteijn et al., 2017), possibly in combination with *p*-hacking. Our preliminary conclusion about the effect of physical experience of weight on importance would be that there is as yet no evidence in the literature for such an effect.

3.4. Other methods using *p*-values for estimation

Several other methods were developed that use *p*-values in order to obtain an effect size estimate corrected for publication bias. Hedges (1984) developed a method for correcting meta-analytic effect sizes for publication bias that is similar to *p*-uniform and *p*-curve. He derived the maximum likelihood estimator of effect size under a model with only statistically significant results and studied the bias in the effect size estimate. Although Hedges (1984) discussed the application to meta-analyses, he only examined the bias in effect size of one statistically significant study. Hedges' method and its performance is not further examined in this chapter because it is currently not applied in practice.

Other methods for obtaining effect size estimates corrected for publication bias are selection models (Hedges & Vevea, 2005). Selection models use an effect size model and a weight function for correcting the effect size estimates for publication bias. The effect size model describes the distribution of effect sizes in case all studies get published. The weight function yields probabilities of observing a particular study given its effect size or *p*-value. Studies' effect sizes are then weighted by these probabilities in order to get an effect size corrected for publication bias (for an overview on selection models see Hedges & Vevea, 2005). Drawbacks of selection models are that they require a large number of studies (i.e., more than 100) in order to avoid non-convergence (e.g., Field & Gillett, 2010; Hedges & Vevea, 2005), often yield implausible weight functions (Hedges & Vevea, 2005), are hard to implement, and require sophisticated assumptions and difficult choices (Borenstein et al., 2009). A recently proposed alternative for selection models based on Bayesian statistics showed promising results and does not suffer from convergence problems when the number of studies in the meta-analysis is small (Guan & Vandekerckhove, 2015). However, a disadvantage of the latter method is that it makes stronger assumptions on weight functions than *p*-uniform and *p*-curve. *P*-uniform and *p*-curve assume that the probability of publishing a finding is independent of its *p*-value given its statistical significance, whereas the models in the method described in Guan and Vandekerckhove (2015) assume specific weights of findings depending on their *p*-value, significant or not. Because both significant and nonsignificant *p*-values are included, this Bayesian method makes assumptions about the extent of publication

bias, and its estimates are affected by the extent of publication bias. For these reasons, we also no longer discuss selection models and their properties.

3.5 Basic idea underlying *p*-uniform and *p*-curve

P-uniform and *p*-curve use the distribution of only the statistically significant *p*-values for estimating effect size, for at least two reasons. First, collecting unpublished studies without the existence of study (or trial) registers is often hard, and these unpublished studies may provide biased information on effect size just like published studies do (Ferguson & Brannick, 2012). Second, evidence for publication bias is overwhelming. For instance, researchers have estimated that at least 90% of the published literature within psychology contains statistically significant results (e.g., Bakker et al., 2012; Fanelli, 2012; Sterling et al., 1995), yielding overestimated effect sizes (e.g., Ioannidis, 2008; Lane & Dunlap, 1978). Because most published findings are statistically significant, only a relatively small number of published but statistically nonsignificant studies (on average up to 10%) need to be omitted from meta-analyses by *p*-curve and *p*-uniform.

Both *p*-uniform and *p*-curve are founded on the statistical principle that the distribution of *p*-values conditional on the true effect size is uniform.[13] This same statistical principle underlies standard null hypothesis significance testing, where the *p*-values are uniformly distributed when the true effect size equals zero. In contrast to null hypothesis significance testing, *p*-values from *p*-uniform and *p*-curve are computed conditional not only on an effect size of zero (which would yield a simple transformation of the traditional *p*-values), but also conditional on other effect sizes (in which case the conditional *p*-value is not a simple transformation of the traditional *p*-value anymore). The effect size estimate of *p*-uniform and *p*-curve represents the effect size for which the conditional *p*-values are uniformly distributed.[14] So what both procedures do is to find an underlying effect, compute for each study the (conditional) *p*-value given this effect, and subsequently check whether these conditional *p*-values show a flat (i.e., uniform) distribution, like they should if indeed the studies reflect that underlying effect. The assumptions of *p*-uniform and *p*-curve are that all statistically significant studies have the same probability of getting published and being included in the meta-analysis, and are statistically independent (i.e., they should not be based on the same sample) (van Assen et al., 2015). We describe the logic

[13] This principle is one of the most fundamental principles of probability and statistics. For instance, this principle is applied when sampling from distributions using so-called "inverse transform sampling" or the "inverse CDF method" (Gentle, 2004). In this method, one starts sampling a random number from a uniform distribution from 0 to 1. Next, the random number, which is considered a cumulative percentage of the distribution, is used to calculate the *x*-value of the distribution that one wished to sample in the first place.
[14] The distribution of (conditional) *p*-values based on the true effect size is only uniform when the assumptions of the underlying statistical model (e.g. independence, effect distribution are valid [Bland, 2013]).

underlying *p*-uniform and *p*-curve as well as how the conditional *p*-value and *p*-uniform's and *p*-curve's effect size estimate are computed in Appendix A.

3.6 *P*-curve and *p*-uniform overestimate effect size if heterogeneity is moderate to large

Simonsohn et al. (2014a) stated that *p*-curve provides accurate effect size estimates in the presence of heterogeneity, i.e., in cases where true effects underlying the studies' observed effects differ. In a blog post Simonsohn (2015, February 9) qualified this statement as follows; "if we apply *p*-curve to a set of studies it tells us what effect we expect to get if we run those studies again". In other words, applying *p*-curve (and *p*-uniform) to a set of studies yields an accurate estimate of the average true effect size of *this exact set of studies*. We note, however, that it may be impossible to run exactly the same studies again since there will always be differences in, for instance, the participants included in the studies and the context in which the studies were conducted.

Because of the importance of its implications for the interpretation of *p*-curve's estimate, we provide a simple example with heterogeneous effect sizes. Assume that the true effect size is equal to either 0 or 1 and that both underlying effects are equally likely, implying an average true effect size $\mu = .5$. Also assume that both true effect sizes are investigated with the same number of studies with a huge sample size, implying 5% and 100% of studies with true effects equal to 0 and 1 are statistically significant, respectively. Because studies' sample sizes are huge, the observed effect sizes of statistically significant studies are equal to (a number very close to) 0 and 1. As a result, *p*-curve's estimate equals $(0.05\times0 + 1\times1)/1.05 = .952$, which is indeed equal to the average underlying true effect size of all the statistically significant studies. However, it is *much* larger than the true population average of .5. Moreover, traditional random-effects meta-analysis will provide a more accurate estimate of true average effect size (i.e., less positively biased) than *p*-curve, even under extreme publication bias.

It is often unrealistic to assume homogeneous true effect sizes underlying primary studies in psychological meta-analyses (e.g., Borenstein et al., 2009). Moreover, researchers often want to estimate the true effect size in the population instead of the average true effect size in the studies included in the meta-analysis. That is, meta-analysts wish to obtain an estimate of .5, rather than .952 in our example. The reason why *p*-curve overestimates effect size under heterogeneity is that studies with an underlying true effect of 0 have a lower probability to be statistically significant, such that these studies are underrepresented in the meta-analysis. In our example, studies with large true effect size are 20 times more likely to be included in the meta-analysis than those with a zero effect size. Finally, we note that in this simple example, we may deal with the heterogeneity rather easily if true effect size (0 or 1) is perfectly linked to an observed dichotomous study characteristic;

applying *p*-curve or *p*-uniform to studies of both groups (a so called subgroup analysis [e.g., Borenstein et al., 2009]) yields the correct estimates of 0 and 1. We therefore recommend applying these methods to subgroups of studies based on the different levels of a moderator in order to create more homogeneous sets of studies (Recommendation 5b, Table 3.1). However, in other realistic situations, the causes of heterogeneity are not simply observed, and subgroup analysis will not completely solve the heterogeneity problem.

To illustrate the effect of heterogeneity of effect sizes on the (over)estimation of effect size by *p*-curve and *p*-uniform, we also ran a simulation study where we varied heterogeneity from moderate to large under the usual scenario where heterogeneity is modeled continuously using a normal distribution of true effects, which is commonly assumed in meta-analysis (Raudenbush, 2009). As in Simonsohn et al. (2014a), 5,000 studies with statistically significant results were generated on which the meta-analysis was conducted. All studies had two conditions with 50 cases each, with population variance equal to 1 in both conditions. Average population effect size was .397, and standard deviations of true effect size (denoted by τ) were 0, 0.2, 0.4, 0.6, and 1, roughly corresponding to I^2 (i.e., ratio of heterogeneity to total variance [Higgins & Thompson, 2002]) values of 0, .5 (moderate heterogeneity), .8 (large heterogeneity), .9, and .96 in the population of studies. Table 3.3 provides the estimates of *p*-curve, *p*-uniform, fixed-effect meta-analysis, and random-effects meta-analysis (with restricted maximum likelihood estimator for estimating the amount of heterogeneity) of all studies with a statistically significant positive effect. For *p*-uniform we used the Irwin-Hall estimator and the so-called "1–*p*" estimator, a variant based on Fisher's method, because this estimator is least affected by extreme effect sizes, and therefore provides better estimates in case of heterogeneity (van Assen et al., 2015).

The first column confirms that *p*-curve and *p*-uniform provide accurate estimates under homogeneity (effect size estimates are close to the true effect size .397), whereas fixed-effect and random-effects meta-analysis (both .553) overestimate effect size. The other columns, however, show that both *p*-curve and *p*-uniform *overestimate* the mean population effect size of .397 for moderate to large heterogeneity, and that this bias increases with larger heterogeneity. Note that the bias of fixed-effect and random-effects meta-analysis also increases with larger heterogeneity, and exceeds the bias of *p*-curve and *p*-uniform in these cases. Although *p*-uniform's "1-*p*" estimator provides the best estimates, its bias is still so large that we do not recommend applying the methodology in its current state to estimate the average population effect size in situations where moderate or large heterogeneity is present or suspected (Recommendation 5a, Table 3.1).

For illustrative purposes, we show how *p*-curve and *p*-uniform may still diagnose heterogeneity by applying *p*-uniform to one simulated meta-analysis of 20 studies with the aforementioned specifications; mean population effect size equal to

.397, and large heterogeneity (τ = 1; I^2 = .96). *P*-uniform's "1-p" estimator yielded an effect size estimate of $\hat{\delta}$ = .795. However, a comparison of the expected conditional *p*-values to the observed conditional *p*-values for $\hat{\delta}$ = .795 in the probability or P-P plot in Figure 3.1 clearly indicated systematic misfit. Specifically, observed conditional *p*-values should be uniformly distributed, as the expected conditional *p*-values. That is, all dots should fall on or close to the diagonal. But, assuming a fixed effect size of .795, the observed conditional *p*-values were either (much) too small (dots below the diagonal to the left) or (much) too large (dots above the diagonal to the right), signifying a large effect size variance. In other words, deviations from the diagonal in the P-P plot may be used to diagnose heterogeneity of effect size.

Table 3.3. Estimates of effect size using *p*-curve, *p*-uniform with Irwin-Hall estimator (IH), *p*-uniform with "1-*p*" estimator, fixed-effect meta-analysis (FE MA), and random-effects meta-analysis (RE MA; using restricted maximum likelihood for estimating the amount of heterogeneity) under different levels of heterogeneity (true effect .397), based on 5,000 studies with statistically significant positive effects.

	τ = 0, I^2 = 0	τ = .2, I^2 = .5	τ = .4, I^2 = .8	τ = .6, I^2 = .9	τ = 1, I^2 = .96
p-curve	.393	.530	.703	.856	1.094
p-uniform (IH)	.383	.535	.724	.874	1.110
p-uniform ("1-p")	.387	.522	.679	.776	.903
FE MA	.553	.616	.738	.875	1.104
RE MA	.553	.616	.743	.897	1.185

To conclude, if moderate to large heterogeneity is present, then *p*-curve and *p*-uniform will estimate the average true effect underlying all *significant* studies in the meta-analysis. When the main goal of the meta-analysis is to estimate the average true effect of the *whole* population of studies in the presence of heterogeneity ($I^2 \geq .5$), we do not recommend using *p*-curve or *p*-uniform, because they then generally overestimate average true effect size (Recommendation 5a, Table 3.1). As opposed to mainstream meta-analytic thinking, Simonsohn et al. (2014a) argued that "the" average true effect size under heterogeneity often does not exist, and even that it is meaningless since studies cannot be run randomly. However, we believe the average true effect size may be meaningfully interpreted in the presence of heterogeneity in some situations, and consider heterogeneity to be both realistic for psychological

studies (e.g., in 50% of the replicated psychological studies in the "Many Labs Replication Project," heterogeneity was present [Klein et al., 2014]) and important to take into consideration when estimating average effect size.

3.7 Sensitivity to *p*-values close to .05

Statistically significant *p*-values that are uniformly distributed in the interval (0; .05) are in line with a zero true effect size. Interestingly, a distribution of *p*-values with many *p*-values close to .05 (and say, an average *p*-value above .025) are not in line with a zero true effect size, but may indicate a *negative* true effect size. We will now show that if the majority of studies in the meta-analysis have a *p*-value just below the significance criterion of .05, then *p*-uniform yields implausible highly negative effect size estimates and a very wide confidence interval. Similarly, under these conditions *p*-curve will behave erratically.

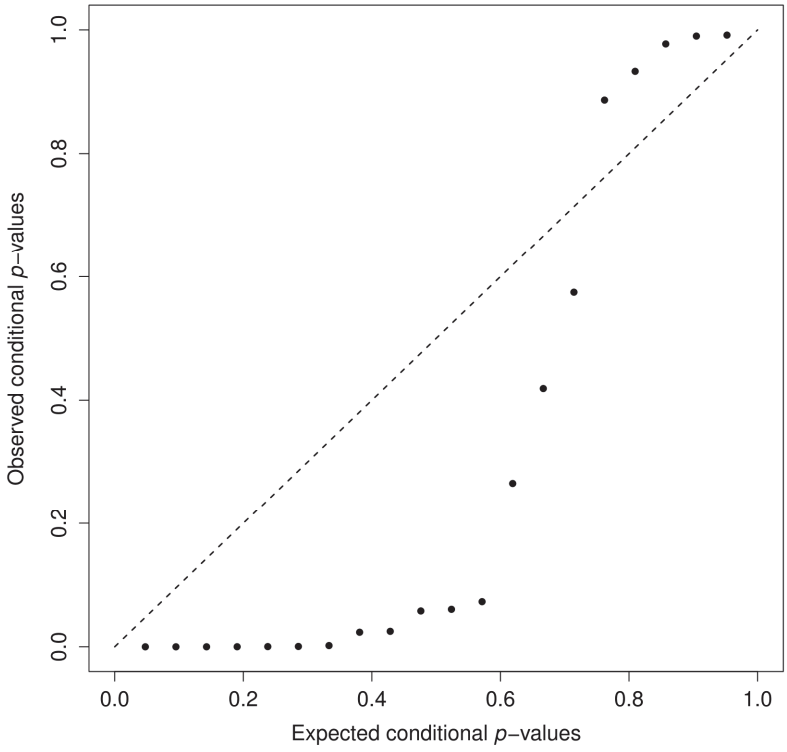

Figure 3.1. Probability or P-P plot for a meta-analysis of 20 studies with large heterogeneity.

To illustrate the consequences of many *p*-values just below .05 on the estimates of *p*-uniform and *p*-curve, consider doing a meta-analysis on the following three observed effect sizes with two conditions having equal sample sizes; Effect 1 with $d = .963$, $t(18) = 2.154$, $p = .045$ (two-tailed), Effect 2 with $d = .582$, $t(48) = 2.058$, $p = .045$, and Effect 3 with $d = .4$, $t(98) = 2.002$, $p = .048$. Several explanations exist for observing multiple *p*-values that barely pass the significance criterion as in this example. First, *p*-hacking such as optional stopping or data peeking (Hartgerink, van Aert, Nuijten, Wicherts, & van Assen, 2015; Lakens, 2014) or the deletion of outliers to achieve statistical significance may yield a preponderance of *p*-values just below .05 (Bakker & Wicherts, 2014b). Another explanation is (bad) luck when the meta-analysis consists of a small number of studies and multiple studies coincidentally have *p*-values close to .05. The fixed-effect meta-analytic estimate for these three observed effect sizes is .506 ($p < .001$), with a 95% confidence interval excluding zero (.199, .812).[15]

Applying *p*-curve to this set of studies yields an effect size estimate of $d = -1.898$. Figure 3.2 displays the behavior of the Kolmogorov-Smirnov test statistic in *p*-curve with dots as a function of effect size. It shows that the Kolmogorov-Smirnov statistic in *p*-curve does not behave as it should (decrease to one minimum, and then increase, and being continuous for all effect sizes). This erratic behavior is caused by *p*-curve's implementation using the *t*-distribution from the software R (R Core Team, 2017), because R yields inaccurate probabilities for very high *t*-values in combination with an extreme non-centrality parameter (Witkovský, 2013). This inaccuracy may cause conditional *p*-values to be negative or undefined (division by zero), which yield the discontinuities in Figure 3.2. Therefore, *p*-curve's estimate cannot be trusted for this example.

P-uniform differs in implementation from *p*-curve because it uses the normal distribution instead of the *t*-distribution for computing conditional *p*-values. The studies' effect sizes are transformed into standardized effect sizes (Hedges' *g*) before the effect size is estimated. Consequently, extreme tail probabilities can be computed, and therefore *p*-uniform behaves as it should, as can be seen from the dashed line in Figure 3.2. At the same time, *P*-uniform's estimate, also based on the Kolmogorov-Smirnov statistic to ease comparison with *p*-curve, is -5.296, which is clearly peculiar. Because a confidence-interval cannot be computed with the Kolmogorov-Smirnov statistic, we also calculated the Irwin-Hall estimates with *p*-uniform; $\hat{\delta} = -5.484$, 95% confidence interval (-15.219, -1.009). Although the behavior of *p*-uniform's estimator is correct, its effect size estimate (< -5) is unrealistically low; the probability of

[15] If *p*-hacking and publication bias were absent in these three studies, the fixed-effect meta-analytic estimator is unbiased and most efficient. In case publication bias was present but none of the three studies involved *p*-hacking, only the estimators of *p*-uniform and *p*-curve would have been accurate. Estimators of both fixed-effect meta-analysis and *p*-uniform and *p*-curve are inaccurate if *p*-hacking was used in the primary studies.

obtaining three *positive* statistically significant studies when δ = -5.484 is essentially zero. Furthermore, *p*-uniform's confidence interval is very wide. We explain in the supplementary materials (https://osf.io/pfmqt/) why these implausible negative estimates can be obtained and what can be concluded from these estimates

In order to deal with *p*-uniform's implausibly negative estimates and *p*-curve's erratic behavior, we recommend setting the effect size estimate of *p*-uniform and *p*-curve to zero in meta-analyses where the mean of the significant *p*-values of the primary studies is larger than .025 (Recommendation 4, Table 3.1). The cutoff of .025 is natural for two reasons. First, if the average *p*-value equals .025, *p*-uniform actually estimates $\hat{\delta}$= 0. Second, higher average *p*-values than .025 would yield negative effect size estimates, and testing is then redundant because the *p*-value of the test will be above .5 and hence cannot be statistically significant. The true effect size can, of course, be below zero, but a left-tailed hypothesis test is then required to examine whether the effect is smaller than zero.

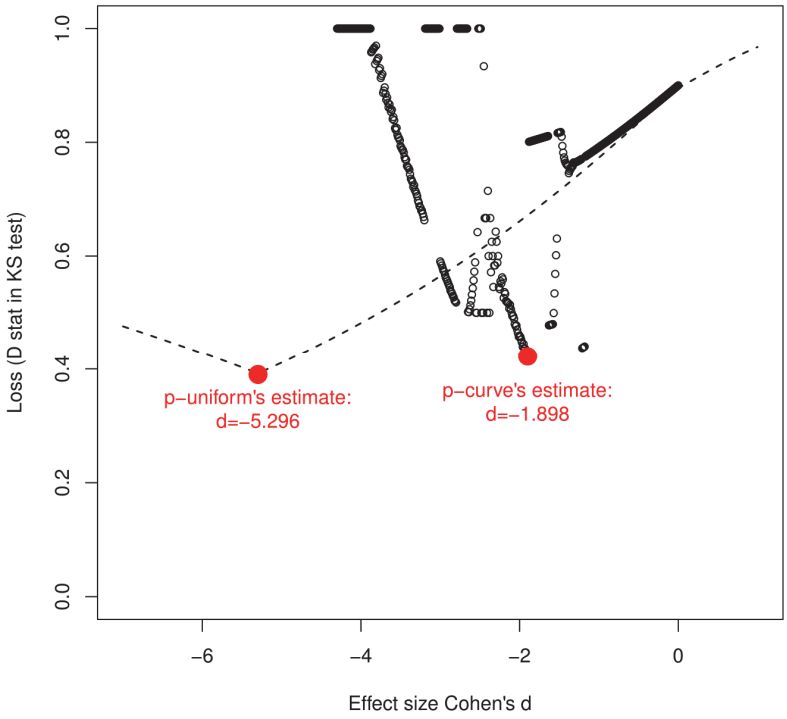

Figure 3.2. Values for Kolmogorov-Smirnov's test statistics in *p*-curve's and *p*-uniform's implementation for the example with three observed effect sizes and *p*-values close to .05.

3.8 *P*-hacking may cause bias in effect size estimates for *p*-uniform and *p*-curve

Simonsohn et al. (2014a) examined the effect of *p*-hacking on *p*-curve's effect size estimation, considering three different *p*-hacking strategies; data-peeking, selectively reporting by using three dependent variables, and selectively excluding outliers. In data-peeking (or optional stopping), observations are added whenever a test is not yet statistically significant. Their *p*-hacking strategy with multiple dependent variables refers to a practice where dependent variables are considered one by one, until one is found for which the test was statistically significant, which is then published. Selectively excluding outliers refers to deleting outliers whenever a test is not yet statistically significant. From their simulations of specific examples of these three practices, they concluded that *p*-curve *under*estimates effect sizes. However, *p*-hacking comprises a very large number of behaviors, and Simonsohn et al. (2014a) examined only three of these behaviors. We will now show that other types of *p*-hacking will lead to *over*estimation of effect size by *p*-curve and *p*-uniform.

As Simonsohn et al. (2014a) explain, *p*-hacking affects *p*-curve's estimate through the conditional *p*-value distribution. For instance, data peeking and selectively excluding outliers lead to a distribution with relatively more conditional *p*-values corresponding to just statistically significant results, which pulls *p*-curve's (and *p*-uniform's) estimate downward, as we have explained in the foregoing section. On the other hand, *p*-hacking behaviors yielding relatively more small *p*-values will result in an overestimation of effect size. Ulrich and Miller (2015) and Bruns and Ioannidis (2016) illustrate that multiple *p*-hacking behaviors may result in relatively more small *p*-values, which will lead to overestimation of effect size by *p*-curve (and *p*-uniform).

We examined the effect of two types of *p*-hacking on effect size estimation by *p*-curve and *p*-uniform. The first behavior again involves selectively reporting among three dependent variables, but differs from the procedure in Simonsohn et al. (2014a) in one crucial aspect; rather than reporting the *first* significant *p*-value, the *smallest* of three significant *p*-values is reported. The second behavior involves a "multiple conditions" scenario, where multiple experimental conditions are run and compared to the same control condition, and only the comparison yielding the largest difference (and smallest *p*-value) is reported. We note that a large portion of surveyed psychologists have admitted to using at least once selective reporting among different dependent variables (63.4%) and not reporting all experimental conditions (27.7%) in their work (John, Loewenstein, & Prelec, 2012).

Figure 3.3 presents the estimates of *p*-uniform, as well as the true effect size and the effect size of fixed-effect meta-analysis (see the supplementary materials [https://osf.io/pfmqt/] for the details of our simulations). We do not show *p*-curve's results because these are almost indistinguishable from *p*-uniform's results. Condition "First significant DV" and "Data peeking" are a replication of the simulations in Simonsohn et al. (2014a), showing that *p*-uniform and *p*-curve indeed underestimate

effect size under these conditions. The estimate is slightly below the true effect size for "First significant DV", and about .2 lower on the scale of Cohen's *d* for "Data peeking" for all true effect sizes from 0 (no effect) to .8 (considered a large effect). Conversely, and as anticipated, both "DV with lowest *p*-value" and "Multiple conditions" overestimate effect size, and this overestimation increases for larger true effect sizes. What should also be mentioned is that *p*-uniform and *p*-curve did not always outperform traditional fixed-effect meta-analysis in the *p*-hacking scenarios we simulated. For instance, fixed-effect meta-analysis outperformed *p*-uniform and *p*-curve (i.e., presented less biased estimates) in the case of "Data peeking" (e.g., (Francis, 2013; van Aert, Maassen, Wicherts, & van Assen, 2016). We therefore conclude that (i) *p*-hacking may bias *p*-uniform's and *p*-curve's estimate in any direction depending on the type of *p*-hacking, (ii) *p*-uniform's and *p*-curve's estimate are not *necessarily* better than those of fixed-effect meta-analysis when *p*-hacking occurs. Thus, *p*-uniform and *p*-curve can deal with publication bias, but (just like traditional fixed-effect and random-effects meta-analysis) neither corrects for *p*-hacking nor reacts predictably to it.

Because the validity of results of both traditional meta-analysis methods and *p*-curve and *p*-uniform may be lowered by *p*-hacking, we recommend scrutinizing both data and studies included in the meta-analysis before applying meta-analytic methods. Underpowered primary studies (i.e., statistical power substantially below 0.8) and a preponderance of *p*-values just below .05 are signals for *p*-hacking. Other signals are unsystematic deletion of outliers and reporting results of other than commonly used measurement instruments. If there are signs of *p*-hacking, we recommend applied researchers to be reluctant in interpreting the results of any meta-analysis (Recommendation 1, see Table 3.1).

3.9 Discussion and conclusion

Recently, new methods were developed aiming to provide an accurate meta-analytic estimate in the presence of publication bias (Simonsohn et al., 2014a; van Assen et al., 2015). These methods, *p*-uniform and *p*-curve, are based on the same basic idea but differ in implementation. The methods' idea is selecting only the statistically significant results and estimating the effect size using the principle of statistical theory that the distribution of (conditional) *p*-values based on the true effect size is uniform. Van Assen et al. (2015) and Simonsohn et al. (2014a) convincingly demonstrated the power of *p*-uniform and *p*-curve and the principles upon which the methods are based to carry out meta-analysis. In this chapter, we explained the rationale and basics of both methods, added three reservations (concerning heterogeneity, incredible estimates, and *p*-hacking) to the application of both methods, and offered hands-on recommendations for researchers.

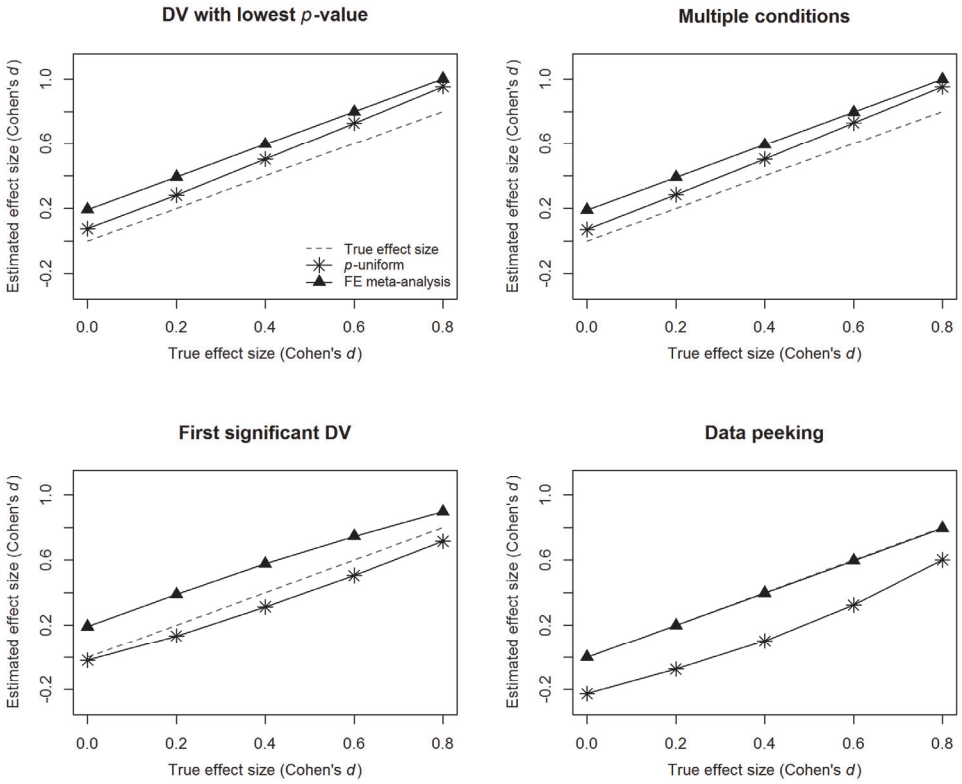

Figure 3.3. Effect size estimates of *p*-uniform and fixed-effect meta-analysis in case of four types of *p*-hacking.

We explained that *p*-curve behaves erratically and yields inaccurate estimates in situations where multiple studies in a meta-analysis have *p*-values close to .05. Due to a different implementation, *p*-uniform does not suffer from this erratic behavior, but provides implausible negative estimates. These problems are solved by setting *p*-uniform's and *p*-curve's estimate to zero whenever the mean of statistically significant studies' *p*-values exceeds .025, i.e., whenever *p*-uniform's estimate is lower than zero. We also showed that *p*-hacking may bias *p*-uniform's and *p*-curve's estimate in any direction depending on the particular type of *p*-hacking, and these methods' estimates are not necessarily better than those of fixed-effect meta-analysis when *p*-hacking has taken place. Finally, we explained that *p*-curve and *p*-uniform estimate the average true effect underlying all *significant* studies in the meta-analysis, but overestimate the average true effect of the *whole* population of studies whenever moderate to large heterogeneity is present.

On the basis of these and contemporary insights we formulated the recommendations summarized in Table 3.1. These recommendations hold for any

meta-analysis and extend the Meta-Analysis Reporting Standards (MARS) as proposed by the APA (American Psychological Association, 2010).

First, we recommend researchers to be reluctant to interpret the results of any meta-analytic technique if there are indicators for *p*-hacking in the primary studies (Recommendation 1), because *p*-hacking may bias the effect size estimates of meta-analysis in any direction. Indicators of potential *p*-hacking include the unsystematic deletion of outliers in many primary studies, the usage and reporting of multiple and different measures for the same dependent variable across primary studies, the common use of small underpowered studies, inconsistencies between sample size descriptions and degrees of freedom (Bakker & Wicherts, 2014a), and grossly misreported *p*-values (Nuijten, Hartgerink, van Assen, Epskamp, & Wicherts, 2015). *P*-hacking can be characteristic of a particular research field (e.g., different measures of dependent variables in a research field) as well as of a single study or a set of studies. Researchers can conduct a sensitivity analysis by comparing the results of traditional meta-analysis methods and *p*-uniform and *p*-curve with the results of these methods applied to only the studies where no *p*-hacking is suspected, for instance, because they involved the use of pre-registered data collection and analysis plans. Meta-analysts will probably observe indicators for *p*-hacking (if these are present) during the literature search and data extraction and do not have to go through all the primary studies again to gather information about the potential presence of *p*-hacking.

Second, we recommend applying fixed-effect and random-effects meta-analysis and *p*-uniform or *p*-curve (Recommendation 2). The selection of a fixed-effect or random-effects meta-analysis should be based on whether a researcher wants to draw inferences on only the studies included in the meta-analysis (fixed-effect) or wants to generalize the meta-analytic results to the whole population of studies (random-effects) (see for a more elaborate discussion Borenstein et al., 2009; F. L. Schmidt, Oh, & Hayes, 2009). Moreover, the estimate of fixed-effect meta-analysis, when compared to the estimate of random-effects meta-analysis, may signal publication bias; publication bias generally results in higher estimates of random-effects than fixed-effect meta-analysis because the studies with smaller sample sizes and usually overestimated effect sizes get less weight in fixed-effect meta-analysis (Greenhouse & Iyengar, 2009).

Next, we recommend checking for direct and indirect evidence of publication bias (Recommendation 3). Direct evidence can be obtained using the publication bias test in *p*-uniform. Previous research suggests *p*-uniform's test for publication bias has higher statistical power than traditional tests (van Assen et al., 2015), which are known to have low statistical power (e.g., Borenstein et al., 2009; Sterne & Egger, 2005). Moreover, use of the quite popular trim-and-fill method is discouraged because it often provides inaccurate results (Moreno, Sutton, Abrams, et al., 2009; Simonsohn et al., 2014a; Stanley & Doucouliagos, 2014; van Assen et al., 2015). However, for a

small number of studies in the meta-analysis or a small amount of publication bias, *p*-uniform's publication bias test will lack sufficient statistical power. In these cases, indirect evidence of publication bias may be used. An example of indirect evidence is if 80% or more of the primary studies' effect sizes are statistically significant when at the same time these studies' sample sizes imply a power of .5 or less to detect a medium effect size (e.g., see Francis, 2013). In case of (direct or indirect) evidence of publication bias, we recommend that conclusions be based on *p*-uniform's or *p*-curve's results, rather than on fixed-effect and random-effects meta-analysis, because these traditional methods overestimate effect size in the presence of publication bias (e.g., Bakker et al., 2012; Ioannidis, 2008b; Lane & Dunlap, 1978; van Assen et al., 2015). Although *p*-uniform and *p*-curve also provide accurate effect size estimates even in the absence of publication bias (Simonsohn et al., 2014a; van Assen et al., 2015), we recommend interpreting fixed-effect and random-effects meta-analysis in this case because these traditional methods yield more efficient and precise estimates.

We recommend setting *p*-uniform's and *p*-curve's estimate to 0 if the average *p*-value of statistically significant studies is larger than .025 (Recommendation 4); an average larger than .025 signals no evidence of an effect and/or the use of *p*-hacking in the set of included studies (in which case, meta-analytic methods' effect size estimation may be biased in any direction depending on the type of *p*-hacking; see Recommendation 1). Interpreting *p*-uniform's and *p*-curve's estimate as the average population effect size estimate is discouraged when effect size heterogeneity is large (Recommendation 5a). In this case, *p*-uniform's and *p*-curve's estimate reflects the average true effect underlying all *significant* studies in the meta-analysis. The average population effect size is overestimated (although the addition of *p*-hacking could complicate this pattern further) when there is moderate or large heterogeneity ($I^2 \geq .5$) and the average true effect of the *whole* population of studies is estimated. In order to deal with heterogeneous effect sizes and still be able to accurately estimate the average true effect of the *whole* population of studies, *p*-uniform or *p*-curve can be applied to homogeneous subgroups of primary studies which were created based on theoretical (e.g., same population of participants being studied) or methodological considerations (using the same methodology, i.e. study design and measures) (Recommendation 5b). The implication of recommendations 3 and 5 is that, currently, no method provides accurate estimates of average population effect size in the presence of both publication bias and heterogeneity.

In the example meta-analysis described earlier, we applied *p*-uniform and *p*-curve to a set of primary studies on the effect of weight on judgment of importance (Rabelo et al., 2015). Researchers can also easily apply *p*-uniform or *p*-curve to their own data. User-friendly R code for applying *p*-uniform can be readily installed.[16] Moreover, we developed a user-friendly web application for researchers who are not

[16] Functions for applying *p*-uniform can be loaded in R by means of running the following code: devtools::install_github("RobbievanAert/puniform"); library(puniform)

familiar with R (https://rvanaert.shinyapps.io/p-uniform). R code for estimating effect size with *p*-curve can be found in the supplementary materials of Simonsohn et al. (2014a). *P*-uniform has the advantage over *p*-curve that it also includes a publication bias test and yields a confidence interval around the effect size estimate.

To conclude, even though both *p*-uniform and *p*-curve are promising meta-analytic methods, the methodology underlying them is still under development, and properties of these methods still need to be examined under more stringent conditions (e.g., different forms of *p*-hacking). Moreover, both methods need to be extended to allow estimation of other effect sizes such as odds ratios, which have their own idiosyncrasies. Once the current methodology is further refined, particularly by enabling accurate estimation in case of heterogeneity, we believe it has the potential to become the standard meta-analytic tool correcting for publication bias. At present, however, researchers should follow the recommendations provided in Table 3.1 to avoid drawing erroneous conclusions from these still developing methods.

3.10 Appendix A: Illustration of logic of and computations in *p*-uniform and *p*-curve

P-curve and *p*-uniform employ the conditional *p*-values, that is, conditional on the effect size being statistically significant. More precisely, the conditional *p*-value of an observed effect size refers to the probability of observing this effect size or larger, conditional on the observed effect size being statistically significant *and* given a particular population (or "true") effect size. Statistical significance has to be taken into account because *p*-uniform and *p*-curve only focus on the interval with *p*-values between 0 and .05 rather than the interval from 0 to 1. Figure 3.A1 depicts how this conditional *p*-value of Effect 3 is computed for three different candidates of the underlying effect size, namely δ=0, δ=0.5 (i.e., the true effect size), and δ=0.748 (i.e., estimate of fixed-effect meta-analysis). Figure 3.A1a reflects the conditional *p*-value for δ=0, which is calculated by dividing the probability of observing an effect size larger than the observed Effect 3 (dark grey area in Figure 3.A1a to the right of d_{obs}) by the probability of observing an effect size larger than the critical value (light and dark grey area in Figure 3.A1a to the right of d_{cv}). For δ=0, the null hypothesis being tested, this boils down to dividing the *p*-value (.0257) by α=.05, yielding a conditional *p*-value denoted by *q)* for Effect 3 of q_3=.0257/.05=.507.[17] Thus, for δ=0 the conditional *p*-value is simply 20 times the traditional *p*-value.

Computation of the conditional *p*-values under effects that differ from zero uses calculations closely resembling the computation of statistical power of a test. Consider the conditional *p*-value of Effect 3 at δ=0.5 (Figure 3.A1b). The critical value (d_{cv}) and the observed effect size (d_{obs}) on the Cohen's *d* scale remain the same, but the distribution of true effect size δ is now shifted to the right. The numerator in computing the conditional *p*-value expresses the probability that the observed effect size d_{obs} is 0.641 or larger given δ=0.5 (dark grey area in Figure 3.A1b to the right of d_{obs}), which equals 0.314, whereas the denominator expresses the probability that the observed effect size is statistically significant given δ=0.5 (light and dark grey area in Figure 3.A1b to the right of d_{cv}), which equals 0.419 (i.e., the traditional power of the study given its degrees of freedom and δ=0.5). This yields a conditional *p*-value for Effect 3 at δ=0.5 of q_3=0.314/0.419=0.75. The conditional *p*-value of Effect 3 at

[17] Due to transformation of d_{obs} and d_{cv} to *z*-values (see later on in this section), conditional *p*-values in *p*-uniform are divided by a value that is slightly larger than .05. Furthermore, dividing by .05 is only feasible if all observed effect sizes are statistically significant in the same direction. Imagine a situation where the observed effect size of Effect 1 is changed into *d* = -0.872, *t*(48) = -3.133. The two-tailed *p*-value of Effect 1 remains .00294, but the observed effect size is statistically significant in the opposite direction than Effect 2 and 3. *P*-uniform and *p*-curve use one-tailed *p*-values and consequently, effects with opposite sign will be omitted when applying *p*-uniform or *p*-curve. If statistically significant effect sizes in a meta-analysis are observed in both tails of the distribution, it is advised to apply *p*-uniform and *p*-curve to both the statistically significant observed positive effect sizes and the statistically significant observed negative effect sizes. An example of such an analysis is described in Simonsohn et al. (2014a).

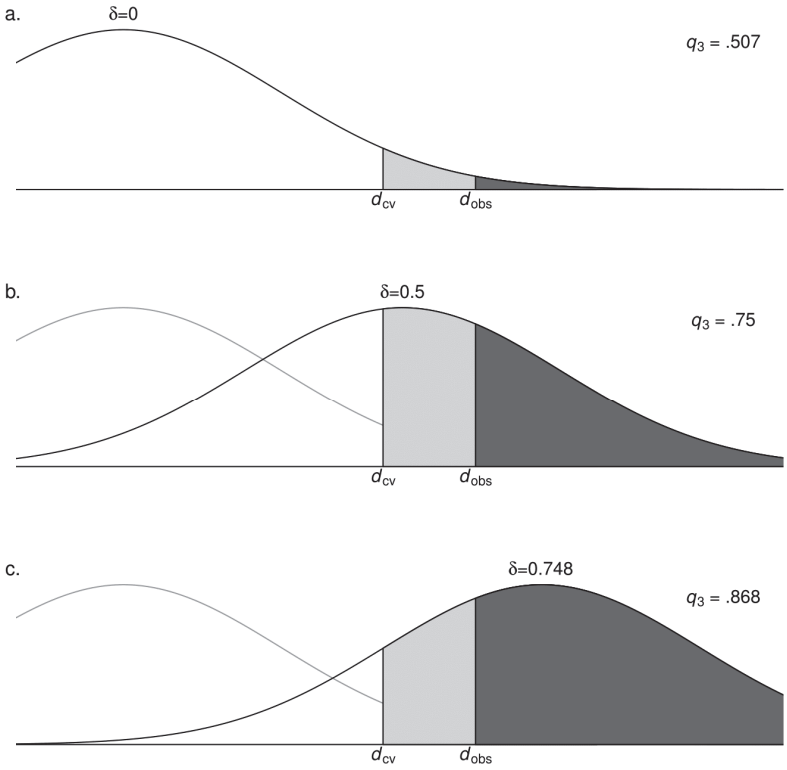

Figure 3.A1. Illustration of computation of conditional *p*-values for Effect 3 (q_3) for three effect sizes: a. δ=0; b. δ=0.5 (true effect size); c. δ=0.748 (estimate of fixed-effect meta-analysis). Critical value on Cohen's *d* scale is denoted by d_{cv} and observed effect size is denoted by d_{obs}.

δ=0.748, as displayed in Figure 3.A1c, can be computed in a similar way: q_3=0.644/0.742=0.868.

The conditional *p*-values of all three observed effect sizes in our example under the three different true effect sizes are presented in Figure 3.A2. The solid black lines in the left panel of Figure 3.A2 shows the conditional *p*-values for δ=0:

q_1=.00294/.05=.0558 q_2=.0110/.05=.213 q_3=.0257/.05=.507.

The dashed grey lines in the left panel illustrate uniformly distributed conditional *p*-values. In case of three studies these uniformly distributed conditional *p*-values should equal ¼, ½, and ¾. Note that the observed conditional *p*-values, summing to .0558+.213+.507=.776, are lower than their corresponding expected uniformly distributed conditional *p*-values, which sum to ¼+½+¾=1.5. Hence, we see that the

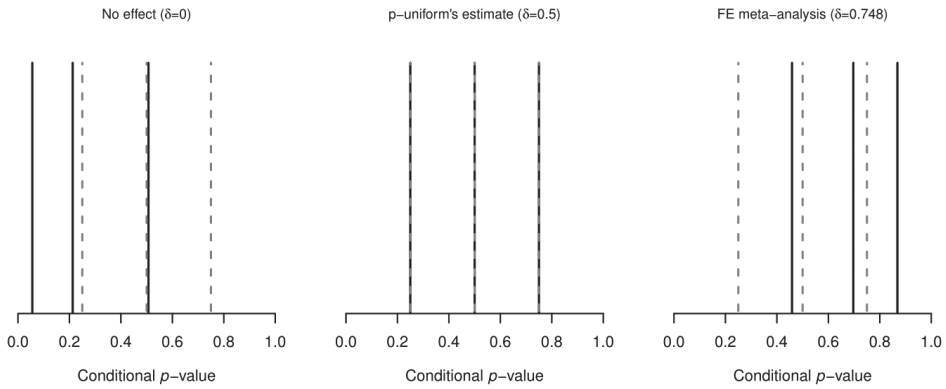

Figure 3.A2. Observed conditional *p*-values (solid black lines) and conditional *p*-values under uniformity (dashed gray lines) for the example with three observed effect sizes. The three panels refer to the conditional *p*-values for *p*-uniform's hypothesis test of no effect (δ=0), *p*-uniform's effect size estimate (δ=0.5), and effect size obtained by fixed-effect meta-analysis (δ=0.748).

conditional *p*-values under the null hypothesis (δ=0) as given in the left-hand side of Figure 3.A2 do *not* fit a uniform distribution.

To obtain the effect size estimate of *p*-uniform, effect size (δ) has to be shifted until the sum of conditional *p*-values equals 1.5, which is the expected value of the sum under uniformity, i.e. given the true effect size. Figure 3.A3 shows the effect of shifting δ on the conditional *p*-values from -.5 to 1.5 for the three observed effect sizes in our example. Each conditional *p*-value increases when the true effect size gets larger. For instance, the conditional *p*-value of Effect 1 increases from .0558 to .25 when the true effect size is increased from 0 to .5, and further increased to .459 if true effect size is increased to .748. As a consequence of these increases, the sum of conditional *p*-values also increases as true effect size increases.

The middle panel in Figure 3.A2 presents the conditional *p*-values in case the effect size is shifted to δ=0.5. These conditional *p*-values are also shown in Figure 3.A3 as the intersections of the three curves with the vertical line representing δ=0.5, and equal:

$$q_1=.25 \qquad q_2=.50 \qquad q_3=.75.$$

These conditional *p*-values exactly match (and studies were selected to exactly match) the expected conditional *p*-values under uniformity. Consequently, the sum of the conditional *p*-values also equals the sum of the conditional *p*-values under uniformity (1.5). This indicates that the effect size estimate of *p*-uniform will be equal to the true effect size of 0.5.

The right panel in Figure 3.A2 and the intersections of the studies' curves with

line δ=0.748 in Figure 3.A3 show the conditional *p*-values conditional on the effect size δ=0.748, which was the estimate of traditional fixed-effect meta-analysis:

$$q_1=.459\ q_2=.697\ q_3=.868.$$

All are higher than their corresponding expected conditional *p*-values under uniformity, and their sum (2.031) is larger than the expected sum under uniformity (1.5). These results indicate that traditional fixed-effects meta-analysis overestimated the effect size. If this occurs, it is not farfetched to suppose that publication bias exists, i.e. some nonsignificant results are missing from the set of studies included in the meta-analysis.

Table 3.A1 shows the results of applying *p*-uniform and *p*-curve to the example. The estimated effect size by *p*-uniform is exactly equal to the true effect size of δ = 0.5. Other output of *p*-uniform is the 95% confidence interval (-0.300; 0.960), and that both the null hypothesis of no effect (p = .0737) and the hypothesis of no publication bias (p = .147) cannot be rejected.[18] The output of *p*-curve incorporate neither a confidence interval nor a publication bias test. *P*-curve's estimate of .511 is slightly larger than the true effect size[19], and *p*-curve's result of the test of no effect is p = .086. Why are the results of both methods different, if they are based on the same logic? This is because the methods differ in implementation, which we explain in the supplementary materials (https://osf.io/pfmqt/).

[18] *P*-uniform's confidence interval is obtained by means of test inversion (e.g., Casella & Berger, 2002), so the lower (upper) bound of the confidence interval equals that effect size for which the sum of conditional *p*-values is equal to the 2.5th (97.5th) percentile of the Irwin-Hall distribution. The statistical test of the null hypothesis of no effect of *p*-uniform examines whether the conditional *p*-values follow a uniform distribution if δ=0 (van Assen et al., 2015).

For the publication bias test of *p*-uniform, all studies (significant and nonsignificant) in a meta-analysis are used for computing the effect size estimate based on fixed-effect meta-analysis. Only the statistically significant studies are then used to examine whether the conditional *p*-values follow a uniform distribution conditional on this fixed-effect meta-analytic effect size estimate. If the statistically significant *p*-values are not uniformly distributed conditional on this effect size estimate, the null hypothesis of no publication bias is rejected (van Assen et al., 2015).

[19] For illustrative purposes we designed an example where *p*-uniform's effect size estimate equals the true effect size; just as easily an example can be constructed where *p*-curve's estimate equals the true effect size.

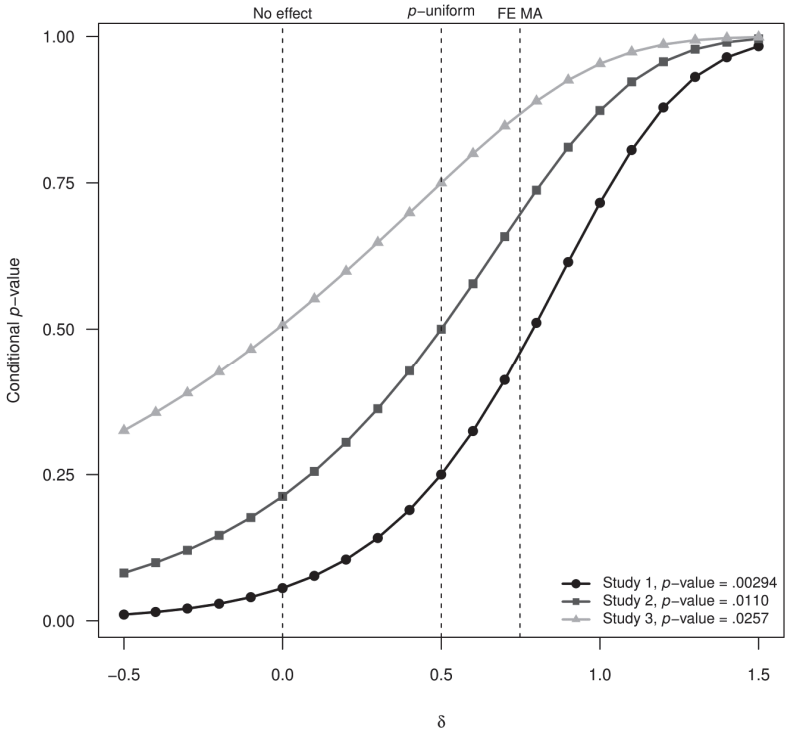

Figure 3.A3. Conditional *p*-values as a function of true effect size (*x*-axis) for each of the three observed effect sizes in the example. Effect sizes zero, true effect size, and estimated by fixed-effects meta-analysis are indicated by vertical lines.

Table 3.A1. Results of *p*-uniform and *p*-curve when applied to the artificial example based on three observed effect sizes, with δ=0.5.

	p-uniform	*p*-curve
Effect size estimate	0.500	0.530
95% CI	(-0.308;0.964)	-
Test of H_0: δ = 0	*z*=-1.44; *p*=.0753	$\chi^2(6)$=1.97; *p*=.0772
Publication bias test	*z*=1.06; *p*=.144	-

3.11 Appendix B

Table 3.B1. Studies and corresponding sample sizes (group 1: n_i^1 and group 2: n_i^2), *t*-values and two-tailed *p*-values as included in the meta-analysis described in Rabelo et al. (2015).

Study no.	Article and experiment	n_i^1	n_i^2	*t*-value	*p*-value
1	Ackerman et al. (2010), Exp. 1	26	28	2.016	0.0489
2	Ackerman et al. (2010), Exp. 2	21	22	1.867	0.0690
3	Chandler et al. (2012), Exp. 2	30	30	2.554	0.0133
4	Chandler et al. (2012), Exp. 1	50	50	2.113	0.0372
5	Chandler et al. (2012), Exp. 3	50	50	2.390	0.0188
6	Hafner (2013), Exp. 1	30	30	2.042	0.0457
7	Jostmann et al. (2009), Exp. 1	20	20	2.245	0.0307
8	Jostmann et al. (2009), Exp. 2	22	28	2.081	0.0428
9	Jostmann et al. (2009), Exp. 3	25	24	2.191	0.0335
10	Jostmann et al. (2009), Exp. 4	20	20	2.294	0.0274
11	Kaspar & Krull (2013)	45	45	3.049	0.0030
12	Kouchaki et al. (2014), Exp. 1a	15	15	2.020	0.0531
13	Kouchaki et al. (2014), Exp. 1c	27	27	2.184	0.0335
14	Kouchaki et al. (2014), Exp. 2	26	25	2.307	0.0254
15	Kouchaki et al. (2014), Exp. 3	35	36	2.308	0.0240
16	Kaspar (2013), Exp. 1	20	20	3.268	0.0023
17	Kaspar (2013), Exp. 2	25.5	25.5	2.306	0.0254
18	Kaspar (2013), Exp. 3	31	31	2.278	0.0263
19	Kaspar (2013), Exp. 4	48.5	48.5	2.053	0.0429
20	Kaspar (2013), Exp. 5	30	30	2.452	0.0172
21	Kouchaki et al. (2014), Exp. 4	31	31	2.139	0.0365
22	Maglio and Trope (2012), Exp. 2	18	18	2.284	0.0287
23	Zhang and Li (2012), Exp. 1	35	35	2.382	0.0200
24	Zhang and Li (2012), Exp. 2	39	39	1.994	0.0498
25	Zhang and Li (2012), Exp. 4	40	40	2.530	0.0134

CHAPTER 4

Publication bias in meta-analyses from psychology and medicine: A meta-meta-analysis

Abstract

Publication bias is a substantial problem for the credibility of research in general and of meta-analyses in particular, as it yields overestimated effects and may suggest the existence of non-existing effects. Although there is consensus that publication bias is widespread, how strongly it affects different scientific literatures is currently less well-known. We examined evidence of publication bias in a large-scale data set of 83 meta-analyses published in Psychological Bulletin (representing meta-analyses from psychology) and 499 systematic reviews from the Cochrane Database of Systematic Reviews (representing meta-analyses from medicine). Publication bias was assessed on homogeneous subsets of the meta-analyses, because publication bias methods do not have good statistical properties if the true effect size is heterogeneous. The rank-correlation test, Egger's test, the test of excess significance, and *p*-uniform's publication bias test yielded evidence for publication bias in approximately 10% of homogeneous subsets. Furthermore, we found hardly any evidence of overestimation of effect size because of publication bias, using the *p*-uniform method or when comparing the meta-analyses' estimates with the estimates based on their largest studies. We therefore conclude that evidence for publication bias in the included meta-analyses is weak at best.

This chapter is submitted as van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2018). Publication bias in meta-analyses from psychology and medicine: A meta-meta-analysis

Meta-analysis is the standard technique for synthesizing different studies on the same topic, and is defined as "the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings" (Glass, 1976, p. 3). One of the greatest threats to the validity of meta-analytic results is publication bias, meaning that the publication of studies depends on the direction and statistical significance of the results (Rothstein et al., 2005a). Publication bias generally leads to effect sizes being overestimated and the dissemination of false-positive results (e.g., Lane & Dunlap, 1978; Nuijten, van Assen, Veldkamp, & Wicherts, 2015). Hence, publication bias results in false impressions about the magnitude and existence of an effect (van Assen et al., 2015) and is considered one of the key problems in contemporary science (Bouter, Tijdink, Axelsen, Martinson, & ter Riet, 2016).

Evidence of publication bias exists in various research fields. The social sciences literature consists of approximately 90% statistically significant results (Fanelli, 2012; Sterling et al., 1995), which is not in line with the on average low statistical power of about 50% or less in, for instance, psychology (Bakker et al., 2012; Cohen, 1990). Franco et al. (2014) examined publication bias in studies that received a grant within the social sciences and found that 64.6% of the studies where most or all null hypotheses failed to be rejected was not written up compared to 4.4% of the studies where most or all the null hypotheses were rejected (cf. Cooper et al., 1997; Coursol & Wagner, 1986). In a highly similar project within the psychological literature, Franco et al. (2016) showed that 70% of the included outcomes in a study were not reported, and that this selective reporting depended on statistical significance of the outcomes.

Compared to the social sciences, more attention has been paid to publication bias in medicine (Hopewell, Clarke, & Mallet, 2005). Medicine has a longer history in registering clinical trials before conducting the research (e.g., Dickersin, Chan, Chalmers, Sacks, & Smith, 1987; Jones et al., 2013). As of 2007, the US Food and Drug Administration Act (FDA) even requires US researchers to make the results of different types of clinical trials publicly available independent of whether the results have been published or not. With registers like *clinicaltrials.gov*, it is easier for meta-analysts to search for unpublished research, and to include it in their meta-analysis. Furthermore, it is straightforward to study publication bias by comparing the reported results in registers with the reported results in publications. Studies comparing the reported results in registers and publications show that statistically significant outcomes are more likely to be reported, and clinical trials with statistically significant results have a higher probability of getting published (Dwan et al., 2008; Kirkham et al., 2010).

A number of methods exist to test for publication bias in a meta-analysis and to estimate a meta-analytic effect size corrected for publication bias. However, publication bias is often not routinely assessed in meta-analyses (Aguinis et al., 2010; Aytug et al., 2012; Banks, Kepes, & Banks, 2012) or is often analyzed with suboptimal

methods that lack statistical power to detect it (Ioannidis & Trikalinos, 2007a). It has been suggested to reexamine publication bias in published meta-analyses (Banks, Kepes, & Banks, 2012; Banks, Kepes, & McDaniel, 2012; Ioannidis, 2008a) by applying recently developed methods to better understand the severity and prevalence of publication bias in different fields. These novel methods have better statistical properties than existing publication bias tests and methods developed earlier to correct effect sizes for publication bias. Moreover, several authors have recommended to not rely on a single method for examining publication bias in a meta-analysis, but rather to use and report a set of different publication bias methods (Coburn & Vevea, 2015; Kepes, Banks, McDaniel, & Whetzel, 2012). This so-called triangulation takes into account that none of the publication bias methods outperforms all the other methods under each and every condition; one method can signal publication bias in a meta-analysis whereas another one does not. Using a set of methods to assess the prevalence and severity of publication bias may yield a more balanced conclusion.

We set out to answer three research questions in this chapter. The first research question concerned the prevalence of publication bias: "What is the prevalence of publication bias in meta-analyses published on psychological and medical topics?" (1a), and "Is publication bias more prevalent in psychology than in medicine?" (1b). Medicine was selected to be compared to psychology, because more attention has been paid to publication bias in general (Hopewell et al., 2005) and study registration in particular (e.g., Dickersin et al., 1987; Jones et al., 2013) within medicine. We also evaluated the amount of agreement between different publication bias methods. In the second research question, we examined whether effect size estimates of traditional meta-analysis and corrected for publication bias by the *p*-uniform method can be predicted by characteristics of a meta-analysis: "What characteristics of a meta-analysis are predictors of the estimates of traditional meta-analysis and *p*-uniform?". Our third research question also consisted of two parts and is about overestimation of effect size caused by publication bias: "How much is effect size overestimated by publication bias?" (3a), and "What characteristics of a meta-analysis predict overestimation?" (3b).

The hypotheses as well as our planned analyses were preregistered (see https://osf.io/8y5ep/) meaning that hypotheses and the analysis plan were specified before the data were analyzed. Some additional analyses were conducted that were not included in the pre-analysis plan. We will explicate which analyses were exploratory when describing these analyses and their results. The chapter continues by providing an overview of publication bias methods. Next, we describe the criteria for a meta-analysis to be included in our study. Then we describe how the data of meta-analyses were extracted and analyzed, and list our hypotheses. Subsequently, we provide the results of our analyses and conclude with a discussion.

4.1 Publication bias methods

Methods for examining publication bias can be divided into two groups: methods that assess or test the presence of publication bias, and methods that estimate effect sizes corrected for publication bias. Methods that correct effect sizes for publication bias usually also provide a confidence interval and test the null hypothesis of no effect corrected for publication bias. Table 4.1 summarizes the methods together with their characteristics and recommendations on when to use each method. The last column of the table lists whether the method is included in our analyses.

4.1.1 Assessing or testing publication bias

The most often used method for assessing publication bias is fail-safe *N* (Banks, Kepes, & McDaniel, 2012; Ferguson & Brannick, 2012). This method estimates how many effect sizes with a zero effect size have to be added to a meta-analysis for changing a statistically significant summary effect size in a meta-analysis to a nonsignificant result (Rosenthal, 1979). Applying the method is discouraged, because it makes the unrealistic assumption that all nonsignificant effect sizes are equal to zero, does not take study sample size into account, and focuses on statistical significance and not on the magnitude of an effect that is of substantial importance (Becker, 2005; Borenstein et al., 2009).

Another popular method is the funnel plot (Light & Pillemer, 1984). In a funnel plot, the effect size estimates of the included studies in a meta-analysis are presented on the *x*-axis and some measure of the effect sizes' precision is displayed on the *y*-axis. The left panel in Figure 4.1 shows a funnel plot for a meta-analysis in the systematic review by Jürgens and Graudal (2004) studying the effect of sodium intake on different health outcomes. Solid circles in the funnel plot indicate studies' Hedges' *g* effect sizes (*y*-axis) and their standard errors or precision (*x*-axis) included in this meta-analysis. A funnel plot illustrates whether small-study effects are present. That is, whether there is a relationship between effect size and its precision. The funnel plot should be symmetric and resemble an inverted funnel in the absence of small-study effects, whereas a gap in the funnel indicates that small-study effects exist. Publication bias is one of the causes of small-study effects (Egger et al., 1997), but funnel plot asymmetry is often interpreted as evidence for publication bias in a meta-analysis. Small-study effects can also be caused by, for instance, researchers basing their sample size on statistical power analyses in combination with heterogeneity in true effect size (see supplemental materials of Open Science Collaboration [2015] and Hedges and Vevea [2005]). In this case, larger true effect sizes are associated with studies using smaller sample sizes, resulting in funnel plot asymmetry.

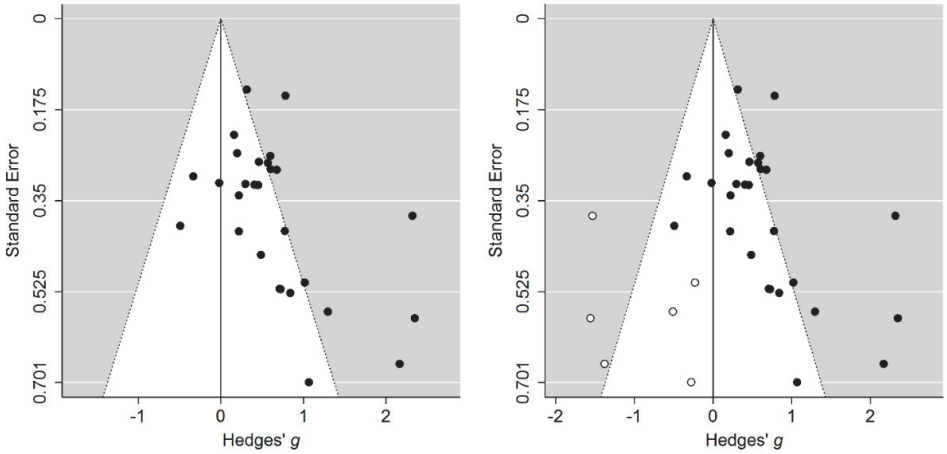

Figure 4.1. Funnel plot showing the relationship between the observed effect size (Hedges' *g*; solid circles) and its standard error in a meta-analysis by Jürgens and Graudal (2004) on the effect of sodium intake on Noradrenaline (left panel). The funnel plot in the right panel also includes the Hedges' *g* effect sizes that are imputed by the trim and fill method (open circles).

Evaluating whether small-study effects exist by eyeballing a funnel plot is rather subjective (Terrin et al., 2005). Hence, Egger's regression test (Egger et al., 1997) and the rank-correlation test (Begg & Mazumdar, 1994) were developed to test whether small-study effects are present in a meta-analysis. Egger's regression test fits a regression line through the observed effect sizes in the funnel plot, and evidence for small-study effects is obtained if the slope of this regression line is significantly different from zero. The rank-correlation test computes the rank correlation (Kendall's τ) between the study's effect sizes and their precision to test for small-study effects. Drawback of these funnel plot asymmetry tests is that statistical power to detect publication bias is low especially if there are few effect sizes in a meta-analysis (Begg & Mazumdar, 1994; Sterne et al., 2000). Hence, these methods are recommended to be only applied to meta-analyses with ten or more effect sizes (Sterne et al., 2011).

The test of excess significance (TES) compares the number of statistically significant effect sizes in a meta-analysis with the expected number of statistically significant effect sizes (Ioannidis & Trikalinos, 2007b). More statistically significant results than expected indicate that some effect sizes are (possibly because of publication bias) missing from the meta-analysis. Ioannidis and Trikalinos (2007b) recommend to not apply the method if heterogeneity in true effect size is present. Moreover, the TES is known to be conservative (Francis, 2013; van Assen et al., 2015).

Another more recently developed method for examining publication bias is

Table 4.1. Summary of publication bias methods to assess publication bias and estimate effect sizes corrected for publication bias. The penultimate column lists principal references of the different methods and the final column indicates whether a method is included in the analyses of this chapter.

Method	Description	Characteristics/Recommendations	Included in analyses
Assessing publication bias			
Fail-safe *N*	Estimates number of effect sizes in the file-drawer	Method is discouraged to be used, because it, for instance, assumes that all nonsignificant effect sizes are equal to zero and focuses on statistical instead of practical significance (Becker, 2005; Borenstein et al., 2009).	No
Funnel plot	Graphical representation of small-study effects where funnel plot asymmetry is an indicator of small-study effects	Publication bias is not the only cause of funnel plot asymmetry (Egger et al., 1997). Eyeballing a funnel plot for asymmetry is subjective (Terrin et al., 2005), so recommendation is to use a statistical test (i.e., Egger's test [Egger et al., 1997] or rank-correlation test [Begg & Mazumdar, 1994]).	No
Egger's and rank-correlation test	Statistical tests for testing funnel plot symmetry	Publication bias is not the only cause of funnel plot asymmetry (Egger et al., 1997). Methods are recommended to be applied when there are 10 or more effect sizes (Sterne et al., 2011) otherwise the methods have low statistical power (Begg & Mazumdar, 1994; Sterne et al., 2000).	Yes
Test of Excess Significance	Computes whether observed and expected number of statistically significant results are in agreement	Do not apply the method in case of heterogeneity in true effect size (Ioannidis & Trikalinos, 2007b). Method is known to be conservative (Francis, 2013).	Yes

Table 4.1 Continued

Method	Description	Characteristics/Recommendations	Included in analyses
p-uniform's publication bias test	Examines whether statistically significant *p*-values are uniformly distributed at the estimate of the fixed-effect model	Method does not use information of nonsignificant effect sizes and, assumes homogeneous true effect size (van Aert, Wicherts, & van Assen, 2016; van Assen et al., 2015).	Yes
Correcting effect size for publication bias			
Trim and fill method	Method corrects for funnel plot asymmetry by trimming most extreme effect sizes and filling these effect sizes to obtain funnel plot symmetry	Method is discouraged to be used because it falsely imputes effect sizes when none are missing and other methods have shown to outperform trim and fill (Moreno, Sutton, Ades, et al., 2009; Simonsohn et al., 2014a; van Assen et al., 2015). Moreover, funnel plot asymmetry is not only caused by publication bias (Egger et al., 1997), and the method does also not perform well if heterogeneity in true effect size is present (Terrin et al., 2003; van Assen et al., 2015).	No
PET-PEESE	Extension of Egger's test where the corrected estimate is the intercept of a regression line fitted through the effect sizes in a funnel plot	Method yields unreliable results if it is based on less than 10 effect sizes (Stanley, Doucouliagos, & Ioannidis, 2017).	No

Table 4.1 Continued

Method	Description	Characteristics/Recommendations	Included in analyses
p-uniform/*p*-curve	Estimate is the effect size for which the distribution of conditional *p*-values is uniformly distributed	Method does not use information of nonsignificant effect sizes and assumes homogeneous true effect size (Simonsohn et al., 2014a; van Aert, Wicherts, et al., 2016; van Assen et al., 2015).	Yes
Selection model approach	Method makes assumptions on the distribution of effect sizes (effect size model) and mechanism of observing effect sizes (selection model). Estimation is performed by combining these two models.	User has to make sophisticated assumptions and choices (Borenstein et al., 2009). Large number of effect sizes (more than 100) are needed to avoid convergence problems (Field & Gillett, 2010; Terrin et al., 2003), but recent research showed that convergence problems of the approach by Iyengar and Greenhouse (Carter, Schönbrodt, Gervais, & Hilgard, 2017; McShane et al., 2016) were only severe if there was no or extreme publication bias in combination with no or a small amount of heterogeneity in true effect size.	No
10% most precise effect sizes	Only the 10% most precise effect sizes are used for estimation with a random-effects model	90% of the available effect sizes is discarded and bias in estimates increases as a function of heterogeneity in true effect size (Stanley, Jarrell, & Doucouliagos, 2010).	Yes

the *p*-uniform method (van Aert, Wicherts, et al., 2016; van Assen et al., 2015). This method is based on the statistical principle that the distribution of *p*-values at the true effect size is uniform. Since in the presence of publication bias not all statistically nonsignificant effect sizes get published, *p*-uniform discards nonsignificant effect sizes and computes *p*-values conditional on being statistically significant. These conditional *p*-values should be uniformly distributed at the (fixed-effect) meta-analytic effect size estimate based on the significant and nonsignificant effect sizes, and deviations from the uniform distribution signals publication bias. *P*-uniform's publication bias test was compared to the TES in a simulation study (van Assen et al., 2015), and statistical power of *p*-uniform was in general larger than the TES except for conditions with a true effect size of zero in combination with statistically nonsignificant studies included in a meta-analysis. This simulation study also showed that Type I error rate of *p*-uniform's publication bias test was too low if the true effect size was of medium size. Limitations of *p*-uniform's publication bias test are that it assumes that the true effect size is homogeneous (which is not very common, see for instance Higgins, 2008; Ioannidis et al., 2006; Stanley & Doucouliagos, 2014), and that the method may inefficiently use the available information by discarding statistically nonsignificant effect sizes in a meta-analysis.

4.1.2 Correcting effect sizes for publication bias

Publication bias tests provide evidence about the presence of publication bias in a meta-analysis. However, statistical power of publication bias tests is often low in practice (Moreno, Sutton, Ades, et al., 2009), because the number of effect sizes in a meta-analysis is often small. For instance, the median number of effect sizes in meta-analyses published in the Cochrane Database of Systematic Reviews was equal to 3 (Rhodes, Turner, & Higgins, 2015; Turner, Jackson, Wei, Thompson, & Higgins, 2015). Furthermore, the magnitude of the effect after correcting for publication bias is more of interest from an applied and theoretical perspective.

The most popular method to correct for publication bias in a meta-analysis is trim and fill (Duval & Tweedie, 2000a, 2000b). This method corrects for funnel plot asymmetry by trimming the most extreme effect sizes from one side of the funnel plot and filling these effect sizes in the other side of the funnel plot to obtain funnel plot symmetry. The corrected effect size estimate is obtained by computing the meta-analytic estimate based on the observed and imputed effect sizes. Trim and fill can also be used to create a confidence interval and test the null hypothesis of no effect after adjusting for funnel plot asymmetry. The procedure of trim and fill is illustrated in the right panel of Figure 4.1. The most extreme effect sizes from the right-hand side of the funnel plot are trimmed and imputed in the left-hand side of the funnel plot (open circles in the right panel of Figure 4.1). A drawback of trim and fill that it shares with other methods based on the funnel plot, is that it corrects for small-study effects that are not necessarily caused by publication bias. Furthermore, the method cannot

accurately correct for publication bias when the true effect size is heterogeneous (e.g., Terrin et al., 2003; van Assen et al., 2015). Simulation studies have shown that results of trim and fill also cannot be trusted because it incorrectly adds studies when none are missing (Peters et al., 2007; Rothstein & Bushman, 2012; Terrin et al., 2003). Hence, the use of trim and fill is discouraged (Moreno, Sutton, Ades, et al., 2009; Simonsohn et al., 2014a; van Assen et al., 2015).

The PET-PEESE method (Stanley & Doucouliagos, 2014) is an extension of Egger's regression test to estimate an effect size in a meta-analysis corrected for small-study effects. PET-PEESE is based on a regression analysis where the observed effect sizes are regressed on their standard errors by means of a weighted least squares regression with the inverse of the effect sizes' sampling variances as weights. If the intercept is not significantly different from zero, the estimate of the intercept is interpreted as the effect size estimate corrected for publication bias. The estimate of the intercept reflects the effect size estimate in a study with a standard error of zero (i.e., a study with an infinite sample size). However, the intercept is biased if the intercept is significantly different from zero (Stanley & Doucouliagos, 2014). Hence, in case the intercept is different from zero, the intercept of another weighted least squares regression analysis (with the inverse sampling variances as weights) is interpreted as the effect size estimate. Simulation studies have shown that PET-PEESE substantially reduced the overestimation caused by small-study effects (Stanley & Doucouliagos, 2014), yet also that it is unlikely to provide reliable results when based on less than 10 effect sizes (Stanley et al., 2017).

The *p*-uniform method can also be used for estimating effect size (and a confidence interval) and testing the null hypothesis of no effect corrected for publication bias. *P*-uniform's effect size estimate is equal to the effect size for which the *p*-values conditional on being statistically significant are uniformly distributed. A similar method that uses the distribution of conditional *p*-values for estimating effect size in the presence of publication bias is *p*-curve (Simonsohn et al., 2014a). This method is similar to the *p*-uniform method, but differs in implementation (for a description of the difference between the two methods see van Aert, Wicherts, et al., 2016). A limitation of *p*-uniform and *p*-curve is that effect sizes are overestimated in the presence of heterogeneity in true effect size (van Aert, Wicherts, et al., 2016). Especially if the heterogeneity in true effect size is more than moderate ($I^2 > 50\%$; more than half of the total variance in effect size is caused by heterogeneity) both methods overestimate the effect size, and their results should be interpreted as a sensitivity analysis. Another limitation of both methods is that they are not efficient if many nonsignificant effect sizes exist. Such results are discarded by the methods, yielding imprecise estimates and wide confidence intervals of *p*-uniform (*p*-curve does not estimate a confidence interval). Yet, *p*-uniform and *p*-curve both outperformed trim and fill in simulation studies (Simonsohn et al., 2014a; van Assen et al., 2015).

A selection model approach (Hedges & Vevea, 2005) can also be used for

estimating effect size corrected for publication bias. A selection model makes assumptions on the distribution of effect sizes (i.e., effect size model) and the mechanism that determines which studies are selected (for publication) and hence observed (i.e., selection model). The effect size estimate (and confidence interval) corrected for publication bias is obtained by combining the effect size and selection model. Many different selection model approaches exist (e.g., Copas, 1999; Dear & Begg, 1992; Hedges, 1984, 1992; Iyengar & Greenhouse, 1988a; Vevea & Hedges, 1995). Some approaches estimate the selection model (Iyengar & Greenhouse, 1988a; Vevea & Hedges, 1995) whereas others assume a known selection model (Vevea & Woods, 2005). A recently proposed selection model approach (Guan & Vandekerckhove, 2015) estimates effect size corrected for publication bias by using Bayesian model averaging over multiple selection models. Selection model approaches are hardly used in practice, because they require sophisticated assumptions and choices (Borenstein et al., 2009) and a large number of effect sizes (more than 100) to avoid convergence problems (Field & Gillett, 2010; Terrin et al., 2003). However, two recent simulation studies (Carter et al., 2017; McShane et al., 2016) included the three-parameter selection model approach by Iyengar and Greenhouse (Iyengar & Greenhouse, 1988a, 1988b) and showed that convergence problems of this approach were limited to conditions including only 10 studies, or to conditions with extreme publication bias.

Stanley et al. (2010) proposed to correct for publication bias in the effect size estimate by computing the unweighted mean of the 10% most precise observed effect sizes, or the single most precise study when there are fewer than ten effect sizes. The rationale underlying only using the 10% most precise observed effect sizes is that these primary studies' effect sizes are less affected by publication bias than the 90% less precise discarded effect sizes. We propose to do not combine the 10% most precise observed effect sizes with an unweighted mean, but with a random-effects model to take differences in primary studies' sampling variances and heterogeneity in true effect size into account. A disadvantage of this method is that it is not efficient, leading to imprecise estimates and wider confidence intervals than estimation based on all effect sizes since up to 90% of the data is discarded. Moreover, bias in the method's estimates increases as a function of the heterogeneity in true effect size (Stanley et al., 2010).

4.2 Methods

4.2.1 Data

A large-scale data set was created with meta-analyses published between 2004 and 2014 in Psychological Bulletin and in the Cochrane Library to study the extent and prevalence of publication bias in psychology and medicine. Psychological Bulletin was selected to represent meta-analyses in psychology, because this journal

publishes many meta-analyses on a variety of topics from psychology. Meta-analyses published in the Cochrane Database of Systematic Reviews (CDSR) of the Cochrane Library were used to represent medicine. This database is a collection of peer-reviewed systematic reviews conducted in the field of medicine.

A first requirement for the inclusion of a meta-analysis was that either fixed-effect or random-effects meta-analysis had to be used in the meta-analysis (i.e., no other meta-analytic methods as, for instance, meta-analytic structural equation modelling or multilevel meta-analysis). Another requirement was that sufficient information in the meta-analysis had to be available to compute the primary study's standardized effect size and its sampling variance. The same effect size measure (e.g., correlation and standardized mean difference) as in the original meta-analysis was used to compute the primary study's effect size and its sampling variance. Formulas as described in Borenstein (2009), Fleiss and Berlin (2009), and Viechtbauer (2007a) were used for computing the standardized effect sizes and their sampling variances. For each included primary study, we extracted information on effect size and sampling variance, as well as information on all categorical moderator variables. Based on these moderators, we created homogeneous subsets of effect sizes. That is, a homogeneous subset consisted of the effect sizes that had the same scores on all the extracted moderators. Consequently, each meta-analysis could contain more than one subset of effect sizes if multiple homogeneous subsets were extracted based on the included moderators.

We only included subsets with less than moderate heterogeneity (I^2<50%, i.e., less than half of the total variance in effect sizes is caused by residual heterogeneity in true effect size) (Higgins, Thompson, Deeks, & Altman, 2003), because none of the publication bias methods has desirable statistical properties under extreme heterogeneity in true effect size (Ioannidis & Trikalinos, 2007a, 2007b; van Aert, Wicherts, et al., 2016; van Assen et al., 2015). Different effect size measures were sometimes used within a meta-analysis. This may cause heterogeneity in a meta-analysis, so the type of effect size measure was also used for creating homogeneous subsets. Publication bias tests have low statistical power (e.g., Begg & Mazumdar, 1994; Macaskill, Walter, & Irwig, 2001; van Assen et al., 2015) if the number of effect sizes in a meta-analysis is small. Hence, another criterion for including a subset in the analyses was that a subset should contain at least five effect sizes.

We searched within the journal Psychological Bulletin for meta-analyses published between 2004 and 2014 by using the search terms "meta-analy*" and *not* "comment", "note", "correction", and "reply" in the article's title. This search resulted in 137 meta-analyses that were published between 2004 and 2014 and that were eligible for inclusion. A flowchart is presented in Figure 4.2 describing the data extraction for the meta-analyses published in Psychological Bulletin. Eighty-three meta-analyses met the inclusion criteria and could be included since the data were available in the paper or were obtained by emailing the corresponding author. Data of

these meta-analyses were extracted by hand and resulted in 9,568 subsets. Data from a random sample of 10% of the included meta-analyses was extracted a second time by a different researcher to verify the procedure of extracting data. Four additional subsets were excluded after verifying the data, because these subsets were heterogeneous instead of homogeneous. After excluding subsets with less than five effect sizes and heterogeneous subsets, a total number of 366 subsets from 83 meta-analyses were available for the analyses.

Data of all systematic reviews in the CDSR are stored online in a standardized format, and data of these reviews can therefore be extracted by an automated procedure. We used the Cochrane scraper developed by Springate and Kontopantelis (2014) to automatically extract data from systematic reviews. The total number of meta-analyses in the CDSR is larger than in Psychological Bulletin, so we drew a simple random sample without replacement of systematic reviews from the CDSR to represent meta-analyses published in medicine. Each systematic review in the database has an identification number. We sampled identification numbers, extracted subsets from the sampled systematic review, and included a subset in our study if (i) I^2<50%, (ii) the number of effect sizes in a subset was at least five, and (iii) the subset was independent of previous included subsets (i.e., no overlap between effect sizes in different subsets). We continued sampling systematic reviews and extracting subsets till the same number of eligible subsets for inclusion were obtained as extracted from Psychological Bulletin (366). Data and/or descriptions of the data of the meta-analyses are available at https://osf.io/9jqht/. The next section describes how the research questions were answered, and how the variables were measured.

4.3 Analysis

Prevalence of publication bias. The prevalence of publication bias in subsets from psychology and medicine was examined to answer research question 1 by using the methods listed in the last column of Table 4.1. Egger's test and the rank-correlation test were used in the analyses to test for funnel plot asymmetry instead of eyeballing a funnel plot. *P*-uniform's publication bias test can be applied to observed effect sizes in a subset that are either significantly smaller or larger than zero. Hence, *p*-uniform was applied to negative or positive statistically significant effect sizes in a subset depending on where the majority of statistically significant effect sizes was observed (using a two-tailed hypothesis test with α=.05). The estimator based on the Irwin-Hall distribution was used for *p*-uniform, because this estimator seems to have the best statistical properties and provides a confidence interval (van Aert, Wicherts, et al., 2016). Publication bias tests have low statistical power, so we followed a recommendation by Egger et al. (1997) to conduct two-tailed hypothesis tests with α=.1 for all methods. Unintentionally, one-tailed *p*-values of *p*-uniform's publication bias test were computed in the preregistered R code for subsets of CDSR instead of the intended two-tailed *p*-values. Since two-tailed *p*-values were computed for all the

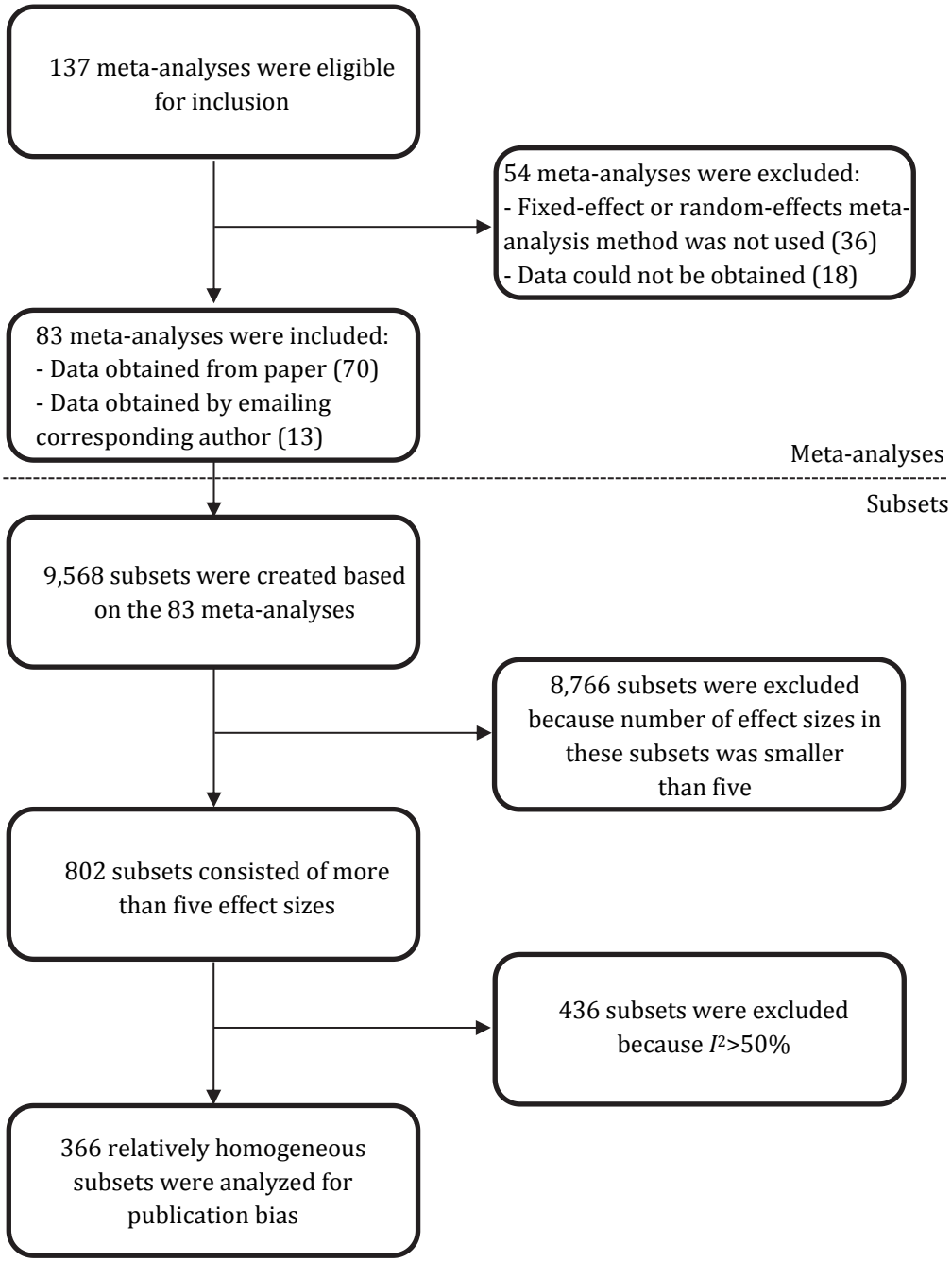

Figure 4.2. Flowchart illustrating the extraction procedure of data from meta-analyses published in Psychological Bulletin between 2004 and 2014.

other publication bias tests were computed, we corrected the pre-registered R code such that two-tailed *p*-values were also computed for *p*-uniform's publication bias test.

We answered research question 1a about the prevalence of publication bias in meta-analyses published in Psychological Bulletin and CDSR by counting how often each method rejected the null hypothesis of no publication bias. Agreement among the publication bias tests was examined by computing Loevinger's *H* values (Loevinger, 1948) for each combination of two methods. Loevinger's *H* is a statistic to quantify the association between two dichotomous variables (i.e., statistically significant or not). The maximum value of Loevinger *H* is 1 indicating a perfect association where the minimum value depends on characteristics of the data. For subsets with no statistically significant effect sizes, *p*-uniform could not be applied, so we computed the association between the results of *p*-uniform and other methods only for subsets with statistically significant effect sizes.

We studied whether publication bias was more prevalent in subsets from Psychological Bulletin and CDSR (research question 1b) by conducting for each publication bias test a logistic regression with as dependent variable whether a publication bias test was statistically significant or not and as predictor a dummy variable indicating whether a subset was obtained from Psychological Bulletin or CDSR (reference category). The number of effect sizes in a subset (or statistically significant effect sizes for *p*-uniform) was included as control variable, because statistical power of publication bias tests depends on the number of effect sizes in a subset and the number of effect sizes in subsets from meta-analyses published in Psychological Bulletin and CDSR were expected to differ. We hypothesized that publication bias would be more severe in subsets from Psychological Bulletin than CDSR after controlling for the number of effect sizes in a subset (or number of statistically significant effect sizes for *p*-uniform). This relationship was expected because medical researchers have been longer aware of the consequences of publication bias whereas broad awareness of publication bias recently originated in psychology. One-tailed hypothesis tests with α=.05 were used for answering research question 1b.

Predicting effect size estimation. Characteristics of subsets were used to predict the estimates of random-effects meta-analysis and estimates of *p*-uniform in research question 2. All effect sizes and their sampling variances were transformed to Cohen's *d* to enable interpretation of the results by using the formulas in section 12.5 of (Borenstein, 2009). If Cohen's *d* and their sampling variances could not be computed based on the available information, Hedges' *g* was used as an approximation of Cohen's *d* (6.4% of all subsets).

Random-effects meta-analysis was used to estimate the effect size rather than fixed-effect meta-analysis. Random-effects meta-analysis assumes that there is no single fixed true effect underlying each effect size (Raudenbush, 2009), and was

preferred over fixed-effect meta-analysis because a small amount of heterogeneity in true effect size could be present in the subsets. The Paule-Mandel estimator (Paule & Mandel, 1982) was used in random-effects meta-analysis for estimating the amount of between-study variance in true effect size since this estimator has the best statistical properties in most situations (Langan et al., 2016; Veroniki et al., 2016). Effect sizes corrected for publication bias were estimated with *p*-uniform and based on the 10% most precise observed effect sizes (see last column of Table 4.1). Estimation based on the 10% most precise observed effect sizes was included as an exploratory analysis to examine whether estimates of *p*-uniform were in line with another method to correct effect sizes for publication bias. If the number of observed effect sizes in a subset was smaller than ten, the most precise estimate was interpreted as estimate of the 10% most precise observed effect sizes. When applying *p*-uniform, we used the estimator based on the Irwin-Hall distribution and two-tailed hypothesis tests in the primary studies (with α=.05). The underlying true effect size in a subset can be either positive or negative. Hence, the dependent variables of these analyses were the absolute values of the estimates of random-effects meta-analysis and *p*-uniform.

Selection model approaches and PET-PEESE methods were not incorporated in the analyses, because the number of effect sizes included in meta-analyses in medicine is often too small for these methods. Selection model approaches suffer from convergence problems when applied to data with these characteristics (e.g., Carter et al., 2017; Field & Gillett, 2010), and PET-PEESE is not recommended to be used since it yields unreliable results if there are less than 10 observed effect sizes (Stanley et al., 2017). *P*-uniform was preferred over trim and fill and *p*-curve, because applying trim and fill is discouraged (Moreno, Sutton, Ades, et al., 2009; Simonsohn et al., 2014a; van Assen et al., 2015) and because *p*-curve is not able to estimate a confidence interval around its effect size estimate.

Two weighted least squares (WLS) regressions were performed with as dependent variables the absolute values of the effect size estimates of either random-effects meta-analysis or *p*-uniform. Since we meta-analyzed the effect sizes estimated with meta-analysis methods, we refer to these analyses as meta-meta-regressions. The inverse of the variance of a random-effects model was selected as weight in both meta-meta-regressions, because it is a function of both the sample size of the primary studies and the number of effect sizes in a subset. *P*-uniform can only be applied to subsets with statistically significant effect sizes, so the meta-meta-regression with the effect size estimates of *p*-uniform as dependent variable was only based on these subsets.

We included five predictors in the meta-meta regressions. The predictors and the hypothesized relationships are listed in the first two columns of Table 4.2. The meta-meta-analytic effect size estimate was expected to be larger in subsets from Psychological Bulletin, because publication bias was expected to be more severe in psychology than medicine. No relationship was hypothesized between the I^2-statistic

and the meta-analytic effect size estimate, because heterogeneity can be either over- or underestimated depending on the extent of publication bias (Augusteijn et al., 2017; Jackson, 2006). Primary studies' precision in a subset was operationalized by computing the harmonic mean of the primary studies' standard error. A negative relationship was expected between primary studies' precision and the meta-analytic estimate, because less precise effect size estimates (i.e., larger standard errors) were expected to be accompanied by more bias and hence larger meta-analytic effect size estimates. The proportion of statistically significant effect sizes in a subset was expected to have a positive relationship on the meta-analytic effect size estimate, because effect sizes with the same sample size that are statistically significant are larger than statistically nonsignificant effect sizes. The predictor indicating the number of effect sizes in a subset was included to control for differences in the number of studies in a meta-analysis, but no effect was expected.

Table 4.2. Hypotheses between predictors and effect size estimate based on random-effects model, *p*-uniform, and overestimation in effect size when comparing estimate of the random-effects model with *p*-uniform (*Y*).

	Hypotheses		
Predictor	Random-effects model	*p*-uniform	Overestimation (*Y*)
Discipline	Larger estimates in subsets from Psychological Bulletin	No specific expectation	Overestimation more severe in Psychological Bulletin
I^2-statistic	No relationship	Positive relationship	Negative relationship
Primary studies' precision	Negative relationship	No relationship	Negative relationship
Proportion of significant effect sizes	Positive relationship	No specific expectation	No specific expectation

The hypotheses concerning the effects in the meta-meta regression on *p*-uniform's estimates are presented in the third column of Table 4.2. No hypothesis was specified for the effect of discipline since *p*-uniform is supposed to correct for possible differences between both disciplines in effect sizes due to publication bias. We expected a positive relationship with the I^2-statistic, because *p*-uniform overestimates

the true effect size in the presence of heterogeneity in true effect size (van Aert, Wicherts, et al., 2016; van Assen et al., 2015). No specific relationship was predicted with primary studies' precision as *p*-uniform is supposed to correct for publication bias. A specific relationship was also not hypothesized for the effect of the proportion of statistically significant effect sizes in a subset. Many statistically significant effect sizes in a subset suggest that the studied effect size is large, sample size of the primary studies were large, or there was severe publication bias in combination with many conducted (but not published) primary studies. These partly opposing effects might cancel each other out or yield a positive or a negative relationship. The number of effect sizes in a subset was again included as control variable.

The effect size estimate of *p*-uniform can become extremely positive or negative if there are multiple *p*-values just below the α-level (van Aert, Wicherts, et al., 2016; van Assen et al., 2015). These outliers may affect the results of the meta-meta-regression with *p*-uniform's estimate as dependent variable. Hence, we used quantile regression (Koenker, 2005) as a sensitivity analysis, because this procedure is less influenced by outliers in the dependent variable. In these quantile regressions, the predictors were regressed on the median of the estimates of *p*-uniform.

Overestimation of effect size. Estimates of random-effects meta-analysis and *p*-uniform obtained for answering research question 2 were used to examine the overestimation caused by publication bias. It is possible that estimates of the meta-analysis and *p*-uniform have opposite signs (i.e., negative estimate of *p*-uniform and positive meta-analytic estimate or the other way around). An effect size estimate of *p*-uniform in the opposite direction than the meta-analytic estimate is often unrealistic, because this suggests that, for instance, a negative true effect size results in multiple positive observed effect sizes. Effect size estimates in opposing directions by meta-analysis and *p*-uniform may be caused by many *p*-values just below α-level (van Aert, Wicherts, et al., 2016). Hence, *p*-uniform's estimate was set equal to zero in these situations. Setting *p*-uniform's estimate to zero when its sign is opposite to that of random-effects meta-analysis is in line with the recommendation in van Aert, Wicherts, et al. (2016).

A new variable *Y* was created to reflect the overestimation of random-effects meta-analysis when compared with *p*-uniform. If the meta-analytic estimate was larger than zero, *Y*=MA-corrected where "MA" is the meta-analytic estimate and "corrected" is the estimate of *p*-uniform. If the meta-analytic estimate was smaller than zero, *Y*=-MA+corrected. Variable *Y* was zero if the estimates of *p*-uniform and meta-analysis were the same, positive if *p*-uniform's effect size estimate was closer to zero than the meta-analytic estimate (if they originally had the same sign), and negative if *p*-uniform's estimate was farther away from zero than the meta-analytic estimate (if they originally had the same sign). The *Y* variable was computed for each subset with statistically significant effect sizes, and we computed the mean and median of *Y* for subsets from Psychological Bulletin and CDSR in order to get an

overall estimate of the amount of overestimation in effect size (research question 3a).

To answer research question 3b, we carried out meta-meta regressions on *Y* with the inverse of the variance of the random-effects meta-analytic estimate as weights. We used the predictors that we also included in research question 2. The hypothesized relationships are summarized in the fourth column of Table 4.2. A larger value on *Y* was expected for subsets from Psychological Bulletin than CDSR, because overestimation was expected to be more severe in psychology than in medicine. We hypothesized a negative relation between the I^2-statistic and *Y*, because *p*-uniform overestimates the effect size in the presence of heterogeneity in true effect size (van Aert, Wicherts, et al., 2016; van Assen et al., 2015). Primary studies' precision was hypothesized to be negatively related to *Y*, because overestimation of the meta-analytic estimate was expected to decrease as a function of primary studies' precision. We had no specific expectations on the relationships between the number of effect sizes in a subset and the proportion of statistically significant effect sizes in a subset. Although a positive effect of this proportion on the meta-analytic effect size estimate was expected, the effect of the proportion on *p*-uniform's estimate was unclear. We included the number of effect sizes in a subset in the meta-meta-regression as a control variable.

Estimates of *p*-uniform that were in the opposite direction than traditional meta-analysis were set equal to zero before computing the *Y* variable. This may have affected the results of the meta-meta-regression since the dependent variable *Y* did not follow a normal distribution. Hence, quantile regression (Koenker, 2005) was used as sensitivity analysis with the median of *Y* as dependent variable instead of the mean of *Y* in the meta-meta regression.

4.4 Results

4.4.1 Descriptive statistics

The total number of included subsets was 732 (366 representing Psychological Bulletin and 366 representing CDSR). Table 4.3 shows descriptive results (number of effect sizes, primary study sample sizes, and positive and negative meta-analytic effect size estimates) of applying random-effects meta-analysis, *p*-uniform, and random-effects meta-analysis based on the 10% most precise observed effect sizes.

The number of effect sizes in subsets was similar in Psychological Bulletin and CDSR. The majority of subsets contained less than 10 effect sizes (third quartile 9 for Psychological Bulletin and 8 for CDSR) meaning that the characteristics of the subsets were very tough for publication bias methods. Statistical power of publication bias is low in these conditions (Begg & Mazumdar, 1994; Sterne et al., 2000) and effect size estimates corrected for publication bias are imprecise (van Aert, Wicherts, et al., 2016; van Assen et al., 2015). The number of statistically significant effect sizes in the

Table 4.3. Median number of effect sizes and median of average sample size per subset and effect size estimates when the subsets were analyzed with random-effects meta-analysis, *p*-uniform, and random-effects meta-analysis based on the 10% most precise effect sizes.

	RE meta-analysis	*p*-uniform	10% most precise
Psychological Bulletin			
Median (IQR) number of effect sizes	6 (5;9)	1 (0;4)	1 (1;1)
Median (IQR) sample size	97.8 (52.4;173.2)	109 (56.5;206.2)	207.3 (100;466)
Positive RE meta-analysis estimates:			
Mean, median, [min.;max.], (SD) of estimates	0.332, 0.279, [0;1.456] (0.264)	-0.168, 0.372, [-21.584;1.295] (2.367)	0.283, 0.22, [-0.629;1.34] (0.289)
Negative RE meta-analysis estimates:			
Mean, median, [min.;max.], (SD) of estimates	-0.216, -0.123, [-1.057;-0.002] (0.231)	-0.041, -0.214, [-5.166;13.845] (1.84)	-0.228, -0.204, [-0.972;0.181] (0.247)
CDSR			
Median (IQR) number of effect sizes	6 (5;8)	1 (0;2)	1 (1;1)
Median (IQR) sample size	126.6 (68.3;223.3)	123.3 (71.9;283.5)	207 (101.2;443)
Positive RE meta-analysis estimates:			
Mean, median, [min.;max.], (SD) of estimates	0.304, 0.215, [0.001;1.833] (0.311)	-1.049, 0.323, [-60.85;1.771] (6.978)	0.284, 0.201, [-0.709;1.757] (0.366)
Negative RE meta-analysis estimates:			
Mean, median, [min.;max.], (SD) of estimates	-0.267, -0.19, [-1.343;0] (0.253)	1.51, -0.239, [-1.581;163.53] (15.064)	-0.214, -0.182, [-1.205;0.644] (0.286)

Note. RE meta-analysis is random-effects meta-analysis, IQR is the interquartile range, min. is the minimum value, max. is the maximum value, SD is the standard deviation, and CDSR is Cochrane Database of Systematic Reviews.

subsets based on a two-tailed hypothesis test with α=.05 was also small (listed in column with results of *p*-uniform). The median number of statistically significant effect sizes in the subsets was 1 for both Psychological Bulletin and CDSR. Moreover, 267 (73%) of the subsets from Psychological Bulletin and 214 (58.5%) of the subsets from CDSR contained at least one statistically significant effect size. Hence, *p*-uniform could only be applied to 481 (65.7%) of the subsets. Of these subsets 180 (37.4%) included only one statistically significant effect size, so the characteristics of the subsets were very challenging for *p*-uniform. The median and interquartile range of the 10% most precise effect size estimates were all equal to one, and estimates of this method were for 676 (92.3%) subsets based on only one effect size.

The median of the average sample size per subset was slightly larger for CDSR (126.6) than for Psychological Bulletin (97.8). The interquartile range of average sample size within subsets from CDSR (68.3; 223.3) was also larger than for subsets from Psychological Bulletin (52.4;173.2). Psychological Bulletin and CDSR showed small differences in the median and interquartile range of the average sample size in subsets from if computed based on only the statistically significant effect sizes (*p*-uniform) or the 10% most precise effect size estimates.

Results of estimating effect size in subsets with random-effects meta-analysis, *p*-uniform, and random-effects meta-analysis based on the 10% most precise observed effect sizes (exploratory analysis) are also included in Table 4.3. To increase interpretability of the results, estimates were grouped depending on whether the effect size estimate of random-effects meta-analysis was positive or negative. The mean and median of the effect size estimates of random-effects meta-analysis and those based on the 10% most precise observed effect sizes were highly similar (difference at most 0.053). However, estimates of *p*-uniform deviated from the other two methods, because *p*-uniform's estimates were in some subsets very positive or negative (i.e., 4 estimates were larger than 10 and 7 estimates were smaller than -10) due to *p*-values of the primary studies' effect sizes close to the α-level. Consequently, the standard deviation and range of the estimates of *p*-uniform were larger than of random-effects meta-analysis and based on the 10% most precise observed effect sizes.

4.4.2 Prevalence of publication bias

Table 4.4 shows the results of applying Egger's regression test, the rank-correlation test, *p*-uniform's publication bias test, and the TES to examine the prevalence of publication bias in the meta-analyses. The panels in Table 4.4 illustrate how often each publication bias test was statistically significant (marginal frequencies and percentages) and also the agreement among the methods (joint frequencies). Agreement among the methods was quantified by means of Loevinger's *H* (bottom-right cell of each panel).

Publication bias was detected in at most 94 subsets (12.9%) by Egger's

Table 4.4. Results of applying Egger's test, rank-correlation test, *p*-uniform's publication bias test, and test of excess significance (TES) to examine the prevalence of publication bias in meta-analyses from Psychological Bulletin and Cochrane Database of Systematic Reviews. *H* denotes Loevinger's *H* to describe the association between two methods.

		Rank-cor.					*p*-uniform		
		Not sig.	Sig.				Not sig.	Sig.	
Egger	Not sig.	600	35	635; 87.1%	Egger	Not sig.	354	34	388; 83.3%
	Sig.	51	43	94; 12.9%		Sig.	70	8	78; 16.7%
	Total	651; 89.3%	78; 10.7%	*H* =.485		Total	424; 91%	42; 9%	*H* =.028

		TES					*p*-uniform		
		Not sig.	Sig.				Not sig.	Sig.	
Egger	Not sig.	609	29	638; 87.2%	Rank-cor.	Not sig.	377	34	411; 88..2%
	Sig.	83	11	94; 12.8%		Sig.	47	8	55; 11.8%
	Total	692; 94.5%	40; 5.5%	*H* =.168		Total	424; 91%	42; 9%	*H* = .082

		TES					TES		
		Not sig.	Sig.				Not sig.	Sig.	
Rank-cor.	Not sig.	620	31	651; 89.3%	*p*-uni-form	Not sig.	393	31	424; 91%
	Sig.	69	9	78; 10.7%		Sig.	33	9	42; 9%
	Total	689; 94.5%	40; 5.5%	*H*=.132		Total	426; 91.4%	40; 8.6%	*H*=.148

Note. The rank-correlation could not be applied to all 732 subsets, because there was no variation in the observed effect sizes in three subsets. All these subsets were part of the meta-analysis by Else-Quest, Hyde, Goldsmith, and Van Hulle (2006) who set effect sizes to zero if the effect size could not have been extracted from a primary study but was reported as not statistically significant.

regression test. The TES and rank-correlation test were statistically significant in 40 (5.5%) and 78 (10.7%) subsets, respectively. In the subsets with at least one statistically significant effect size, *p*-uniform's publication bias test detected publication bias in 42 subsets (9%), which was more than TES (40; 8.6%) and less than both the rank-correlation test (55; 11.8%) and Egger's regression test (78; 16.7%). Since the estimated prevalence values are close to 10%, which equals the significance threshold of each test, we conclude there is at best weak evidence of publication bias on the basis of publication bias tests. Associations among the methods were low ($H < .168$), except for the association between Egger's regression test and the rank-correlation test ($H = .485$).

To answer research question 1b we examined whether publication bias was more prevalent in subsets from Psychological Bulletin than CDSR. Publication bias was detected in 13.4% (Egger's test), 12.8% (rank-correlation test), 11.4% (*p*-uniform), 6.6% (TES) of the subsets from Psychological Bulletin and in 12.2% (Egger's test), 8.5% (rank-correlation test), 5.9% (*p*-uniform), and 4.4% (TES) of the subsets from CDSR. When testing for differences in publication bias we controlled for the number of effect sizes (or for *p*-uniform statistically significant effect sizes) in a meta-analysis. Publication bias was more prevalent in subsets from Psychological Bulletin if the results of *p*-uniform were used as dependent variable (odds ratio=2.226, z=2.217, one-tailed *p*-value=.014), but not for Egger's regression test (odds ratio=1.024, z=0.106, one-tailed *p*-value=.458), rank-correlation test (odds ratio=1.491, z=1.613, one-tailed *p*-value=.054), and TES (odds ratio=1.344, z=0.871, one-tailed *p*-value=.192). Tables with the results of these logistic regression analyses are reported in S1-4 (https://osf.io/wdjy4/). Note, however, that if we control for the number of tests performed (i.e., 4) by means of the Bonferoni correction ($p = .005 < .05/4 = .0125$), the result of *p*-uniform was no longer statistically significant. All in all, we conclude that evidence of publication bias was weak at best and that no clear difference in the extent of publication bias existed between subsets from Psychological Bulletin and CDSR.

4.4.3 Predicting effect size estimation

To answer research question 2, absolute values of the effect size estimates of random-effects meta-analysis and *p*-uniform were predicted based on characteristics of the subsets. One-tailed hypothesis tests were used in case of a directional hypothesis (see Table 4.2 for a summary of our hypotheses). Table 4.5 presents the results of the meta-meta-regression on the absolute value of the effect size estimates of random-effect meta-analysis. The variables in the model explained 67.6% of the variance in the estimates of random-effects meta-analysis (R^2=0.676; $F(5,726)=303$, $p < .001$). The absolute value of the meta-analytic estimate was 0.006 larger for subsets from Psychological Bulletin compared to CDSR, but this effect was not statistically significant and not in line with our hypothesis ($t(726)=0.637$, p=.262, one-tailed). The

I^2-statistic had an unexpected negative association on the absolute value of the meta-analytic estimate (B=-0.001, t(726)=-4.601, p<.001, two-tailed). The harmonic mean of the standard error had, as expected, a positive effect (B=1.185, t(726)=25.514, p<.001, one-tailed). As hypothesized, a positive effect was observed for the proportion of statistically significant effect sizes on the absolute value of the meta-analytic estimate (B=0.489, t(726)=34.269, p<.001, one-tailed).

Table 4.5. Results of meta-meta regression on the absolute value of the random-effects meta-analysis effect size estimate with predictors discipline, I^2-statistic, harmonic mean of the standard error (standard error), proportion of statistically significant effect sizes in a subset (Prop. sig. effect sizes), and number of effect sizes in a subset.

	B (SE)	t-value (p-value)	95% CI
Intercept	-0.144 (0.012)	-11.697 (<.001)	-0.168;-0.12
Discipline	0.006 (0.009)	0.637 (.262)	-0.012;0.023
I^2-statistic	-0.001 (0.0003)	-4.601 (<.001)	-0.002;-0.001
Standard error	1.185 (0.046)	25.514 (<.001)	1.094;1.277
Prop. sig. effect sizes	0.489 (0.014)	34.269 (<.001)	0.461;0.517
Number of effect sizes	-0.0004 (0.0003)	-1.408 (.16)	-0.001;0.0002

Note. CDSR is the reference category for discipline. p-values for discipline, harmonic mean of the standard error, and proportion of significant effect sizes in a subset are one-tailed whereas the other p-values are two-tailed. CI = confidence interval.

Table 4.6 shows the results of meta-meta regressions on the absolute value of p-uniform's estimate as the dependent variable. The proportion explained variance in p-uniform's estimate was R^2=.014 (F(5,475)=1.377, p=.231). None of the predictors was significant. Quantile regression was used as sensitivity analysis to examine whether the results were distorted by extreme effect size estimates of p-uniform (see Table S5 available at https://osf.io/wdjy4/). The results of the predictors discipline and I^2-statistic were also not statistically significant in the quantile regression. The association of the harmonic mean of the standard error was lower in the quantile regression but statistically significant (B=2.037, t(475)=3.739, p<.001, two-tailed) and the predictor "proportion of statistically significant effect sizes" was statistically significant (B=0.196, t(475)=2.353, p=.019, two-tailed).

4.4.4 Overestimation of effect size

Results indicated that the overestimation was less than d=0.06 for subsets from Psychological Bulletin (mean=-0.007, median=0.019, standard deviation = 0.412) and CDSR (mean=0.043, median=0.051, standard deviation = 0.305), and that differences between estimates of subsets from Psychological Bulletin and CDSR were negligible (research question 3a). Table 4.7 presents the results of the meta-meta regression on Y. The predictors explained 11.8% of the variance of Y (F(5,475)=12.76, p < .001). The effect size in subsets from Psychological Bulletin was not significantly larger than from CDSR (B=-0.040, t(475)=-1.651, p=.951, one-tailed). Consistent with the negative effect of the I^2-statistic on the absolute value of the meta-analytic estimate (Table 4.5), we found a negative effect of the I^2-statistic on Y (B=-0.004, t(475)=-5.338, p<.001, one-tailed). The hypothesized relationship between the harmonic mean of the standard error and Y was not statistically significant (B=0.172, t(475)=1.371, p=.086, one-tailed). The proportion of statistically significant effect sizes in a subset was positively associated with Y (B=0.182, t(475)=4.713, p<.001, two-tailed).

Table 4.6. Results of meta-meta-regression on the absolute value of p-uniform's effect size estimate with predictors discipline, I^2-statistic, harmonic mean of the standard error (standard error), proportion of statistically significant effect sizes in a subset (Prop. sig. effect sizes), and number of effect sizes in a subset.

	B (SE)	t-value (p-value)	95% CI
Intercept	0.77 (0.689)	1.118 (0.264)	-0.584;2.124
Discipline	0.001 (0.497)	0.001 (0.999)	-0.975;0.976
I^2-statistic	0.013 (0.014)	0.939 (0.174)	-0.014;0.039
Standard error	3.767 (2.587)	1.456 (0.146)	-1.316;8.851
Prop. sig. effect sizes	-1.287 (0.797)	-1.615 (0.107)	-2.853;0.279
Number of effect sizes	-0.02 (0.015)	-1.363 (0.173)	-0.049;0.009

Note. CDSR is the reference category for discipline. p-value for the I^2-statistic is one-tailed whereas the other p-values are two-tailed. CI = confidence interval.

The estimate of p-uniform was truncated to zero in 136 subsets before computing Y in order to deal with unlikely cases where p-uniform's estimate would be in the opposite direction than the estimate of random-effects meta-analysis. Hence, quantile regression with the median of Y as dependent variable was conducted to examine whether the results of the meta-meta-regression were affected by this

truncation (see Table S6 available at https://osf.io/wdjy4/). The predictor discipline was not statistically significant in the quantile regression. In contrast to the results of the meta-meta-regression, the effects of the I^2-statistic (B=-0.0003, t(475)=-0.2, p=.579, one-tailed) and proportion of statistically significant effect size in a subset (B=-0.002, t(475)=-1.53, p=.127, two-tailed) were no longer statistically significant, whereas the predictor harmonic mean of the standard error was statistically significant (B=0.279, t(475)=1.889, p=.03, one-tailed).

Table 4.7. Results of meta-meta-regression on the effect size overestimation in random-effects meta-analysis when compared to p-uniform (Y) and predictors discipline, I^2-statistic, harmonic mean of the standard error (standard error), proportion of statistically significant effect sizes in a subset (Prop. sig. effect sizes), and number of effect sizes in a subset.

	B (SE)	t-value (p-value)	95% CI
Intercept	-0.017 (0.033)	-0.517 (.605)	-0.083;0.048
Discipline	-0.04 (0.024)	-1.651 (.951)	-0.087;0.008
I^2-statistic	-0.004 (0.001)	-5.338 (<.001)	-0.005;-0.002
Standard error	0.172 (0.126)	1.371 (.086)	-0.074;0.419
Prop. sig. effect sizes	0.182 (0.039)	4.713 (<.001)	0.106;0.258
Number of effect sizes	-0.001 (0.001)	-2.064 (.04)	-0.003;-0.0001

Note. CDSR is the reference category for discipline. p-values for discipline, the I^2-statistic, and the harmonic mean of the standard error are one-tailed whereas the other p-values are two-tailed. CI = confidence interval.

4.5 Conclusion and discussion

Publication bias is a major threat to the validity of meta-analyses. It results in overestimated effect sizes in primary studies which in turn also bias the meta-analytic results (e.g., Lane & Dunlap, 1978; van Assen et al., 2015). Evidence for publication bias has been observed in many research fields (e.g., Fanelli, 2012; Franco et al., 2014; Franco et al., 2016; Sterling et al., 1995), and different methods were developed to examine publication bias in a meta-analysis (for an overview see Rothstein et al., 2005a). We studied the prevalence of publication bias and the overestimation caused by it in a large number of meta-analyses published in Psychological Bulletin and CDSR by applying publication bias methods to homogeneous subsets of these meta-analyses. Homogeneous subsets were created, because publication bias methods have poor statistical properties if the true effect size is heterogeneous (Ioannidis & Trikalinos, 2007a, 2007b; van Aert, Wicherts, et al., 2016; van Assen et al., 2015). The prevalence

of publication bias was studied by means of Egger's test (Egger et al., 1997), the rank-correlation test (Begg & Mazumdar, 1994), TES (Ioannidis & Trikalinos, 2007b), and *p*-uniform's publication bias test (van Assen et al., 2015). We used *p*-uniform and a meta-analysis based on the 10% most precise effect size estimates of a meta-analysis to estimate the effect size corrected for publication bias.

The results of this chapter are not in line with previous research showing that publication bias is omnipresent in science (e.g., Fanelli, 2012; Franco et al., 2014; Franco et al., 2016; Sterling et al., 1995). Only weak evidence for the prevalence of publication bias was observed in our large-scale data set. This weak evidence is not the result of lack of statistical power of these publication bias tests, as correcting effect size estimates for publication bias or focusing on the effect size of the largest study/studies in the meta-analysis also did not result in clear indications of publication bias in our large-scale data set. Another indication for the absence of severe publication bias in the studied meta-analyses is that many statistically nonsignificant effect sizes were included in the meta-analyses. Only 28.9% of the observed effect sizes were statistically significant in the meta-analyses from Psychological Bulletin, and 18.9% of the observed effect sizes were statistically significant in the meta-analyses from CDSR. This is substantially lower than the approximately 90% statistically significant results in the social sciences literature (Fanelli, 2012; Sterling et al., 1995) and the nearly 90% positive results that appear in the clinical medicine literature (Fanelli, 2010b).

A meta-meta-regression on the random-effects meta-analytic estimates revealed, in line with the hypothesis, a negative association of primary studies' precision with a meta-analytic estimate. Since at best only weak evidence for publication bias was found, this association was most likely caused by differences in sample sizes between research fields. For instance, if researchers use statistical power analysis to determine the sample size of their study or if researchers in fields characterized by lower effect sizes use larger sample sizes by habit, larger true effect sizes will be associated with studies using smaller sample sizes (see supplemental materials of Open Science Collaboration, 2015).

The same predictors used for predicting the random-effects meta-analytic effect size estimate were also used in a meta-meta-regression on *p*-uniform's estimate. None of the predictors statistically significantly predicted *p*-uniform's effect size estimate. This was in line with our hypothesis on the relationship with primary studies' precision, but in contrast to the expected positive relationship between the I^2-statistic and *p*-uniform's effect size estimate. The absence of such a positive relationship indicates that *p*-uniform did not overestimate the effect size in the presence of heterogeneity in true effect size. The explained variance in the meta-meta-regression with *p*-uniform's estimate as dependent variable (1.4%) was substantially lower than with the estimate of random-effects meta-analysis as dependent variable (67.6%). This difference in explained variance was mainly caused by the large

variance in *p*-uniform's effect size estimates across subsets. The variance in these estimates was large, because estimates of *p*-uniform were sometimes extremely positive or negative caused by observed effect sizes with *p*-values just below the α-level.

The different publication bias methods were not always in agreement with each other, which is likely caused by the absence of clear publication biases in the meta-analytic subsets. An exception was the association between the results of Egger's test and the rank-correlation, but this association was expected since both methods are very similar in methodology (i.e., testing for publication bias by examining the relationship between observed effect size and some measure of its precision). Substantial differences were also observed among the two methods that we used to correct effect size estimates for publication bias (i.e., *p*-uniform and the 10% most precise observed effect size estimates). Effect size estimates based on the 10% most precise observed effect sizes were close to estimates of the random-effects meta-analysis whereas estimates of *p*-uniform were imprecise and sometimes very different from random-effects meta-analysis and the 10% most precise observed effect sizes. This suggests that *p*-uniform overcorrected for publication bias because of a small number of observed effect sizes in subsets combined with *p*-values of primary studies being close to the α-level.

Publication bias could, however, have gone undetected due to a variety of reasons. First, publication bias is less of an issue if the relationship of interest in a meta-analysis was not the main focus of the primary studies. Statistical significance of the main result in a primary study probably determines whether a result gets published, rather than whether a secondary outcome or supplementary result is significant. For instance, a meta-analysis might be about gender differences where data is extracted from studies that used gender only as a control variable. Second, meta-analysts included many unpublished studies in their meta-analyses, which might have decreased the severity and detectability of publication bias in our selected meta-analyses. Third, publication bias methods were applied to subsets of primary studies' included in the meta-analysis to create more homogeneous subsets. Creating these smaller subsets was necessary because publication bias methods' statistical properties deteriorate if heterogeneity in true effect size is moderate or large. Our selection of homogeneous subsets could have let to the exclusion of subsets with severe publication bias. Imagine a subset with a number of statistically significant effect sizes that were published in a field with considerable publication bias, and a few statistically nonsignificant effect sizes that were obtained from unpublished research. The inclusion of the effect sizes from the unpublished research may cause heterogeneity in true effect size, and therefore a subset with potentially severe publication bias was excluded from our study.

Although no convincing evidence for publication bias was observed in our study, we agree with others (e.g., Banks, Kepes, & McDaniel, 2012; Field & Gillett,

2010; Sutton, 2005) that publication bias should be routinely assessed in every meta-analysis. Moreover, a set of publication bias methods is recommended to be applied and reported in each meta-analysis, because each method assesses publication bias in a different way and one method might detect or correct for publication bias in a meta-analysis whereas another method might not (Coburn & Vevea, 2015; Ferguson & Brannick, 2012; Kepes, Banks, McDaniel, et al., 2012). Future research should focus on developing publication bias methods that are able to examine publication bias in meta-analyses with a small number of effect sizes and heterogeneity in true effect size. We are currently working on extending *p*-uniform such that it can also deal with heterogeneous true effect size. Other promising developments are the recently increased attention for selection models (e.g., Carter et al., 2017; Citkowicz & Vevea, 2017; McShane et al., 2016) and the development of a Bayesian method to correct for publication bias (Guan & Vandekerckhove, 2015). Although meta-analysts will greatly benefit from improved methods to assess publication bias, attention should also be paid to registering studies to make unpublished research readily accessible (e.g., Dickersin, 2005). Such a register enables meta-analysts to also include unpublished research in their meta-analysis and will improve the validity of meta-analytic results.

CHAPTER 5

Correcting for publication bias with the *p*-uniform* method

Abstract

Publication bias is a major threat to the validity of a meta-analysis resulting in overestimated effect sizes in a meta-analysis. The *p*-uniform method is a meta-analysis method that corrects estimates for publication bias, but the method overestimates average effect size in the presence of heterogeneity in true effect sizes (i.e., between-study variance). We propose an extension and improvement of the *p*-uniform method called *p*-uniform*. *P*-uniform* improves upon *p*-uniform in three important ways, as (i) it entails a more efficient estimator, (ii) it eliminates the overestimation of effect size in case of between-study variance in true effect sizes, and (iii) it enables estimating and testing for the presence of the between-study variance in true effect sizes. We compared the statistical properties of *p*-uniform* with the selection model approach of Hedges (1992) as implemented in the R package "weightr" and the random-effects model in both an analytical and a Monte Carlo simulation study. Results revealed that the statistical properties of *p*-uniform* and the selection model approach were generally comparable, and that both methods outperformed the random-effects model if publication bias was present. However, the methods did not perform well when there was extreme publication bias resulting in only statistically significant primary studies' effect sizes in a meta-analysis, but bias in estimates of effect size and between-study variance in true effect size was generally small if there were 10 primary studies included in a meta-analysis. We offer recommendations for correcting meta-analyses for publication bias in practice, and provide an R package and an easy-to-use web application for applying *p*-uniform*.

Independent effect sizes from multiple primary studies can be statistically combined by means of a meta-analysis in order to obtain a quantitative summary of the studied relationship. Meta-analysis is now seen as the "gold standard" for synthesizing evidence from multiple studies (Aguinis et al., 2011; Head et al., 2015). However, a threat to the validity of a meta-analysis is publication bias (e.g., Borenstein et al., 2009; Rothstein et al., 2005a). Publication bias refers to situations where the published literature is not a representative reflection of the population of completed studies (Rothstein et al., 2005a). In its most extreme case this implies that studies with statistically significant results get published and studies with statistically nonsignificant results do not get published. Publication bias is not only caused by reviewers and editors who are reluctant to accept studies without statistically significant results, but also by researchers who do not submit studies with nonsignificant results (Cooper et al., 1997; Coursol & Wagner, 1986). The consequences of publication bias are severe and may hamper scientific progress, because publication bias causes an increase in false positives in the published literature (van Assen et al., 2015) and often results in overestimated effect sizes in primary studies and meta-analyses (e.g., Kraemer et al., 1998; Lane & Dunlap, 1978).

Evidence for publication bias has been observed in multiple research fields. Fanelli (2010b, 2012) studied how often the authors declared to have found support for the tested hypothesis in a random sample of published papers from a variety of research fields. In psychiatry and psychology, 95% of the papers concluded that the hypothesis was supported which was the largest percentage across all included research fields. However, this large percentage is not in line with the on average low statistical power of approximately 50% (or lower) in psychological research (Bakker et al., 2012; Cohen, 1990). This suggests that the published literature is not representative for the population of completed studies, which may have been caused by publication bias. Other more direct evidence of publication bias in psychology was found in Franco et al. (2016). These authors compared the outcomes that were included in grant applications for psychological experiments with the outcomes that were reported in the published paper, and concluded that 70% of the studies did not report all the outcomes. Moreover, the reported effect sizes in the published papers were twice as large as the unreported effect sizes and were approximately three times as likely to be statistically significant. Although the evidence for publication bias in multiple research fields is strong, we have to emphasize that publication bias is not ever-present. For instance, two recent large-scale studies did not observe publication bias in meta-analyses about posttraumatic stress disorder (Niemeyer et al., 2018) and in meta-analyses published in psychology and medicine (van Aert, Wicherts, & van Assen, 2018).

Numerous methods have been developed to assess and test for publication bias in a meta-analysis, including fail-safe *N* (Becker, 2005), funnel plot (Light & Pillemer, 1984), Egger's regression test (Egger et al., 1997), rank-correlation test

(Begg & Mazumdar, 1994), test of excess significance (Ioannidis & Trikalinos, 2007b), *p*-uniform's publication bias test (van Aert, Wicherts, et al., 2016; van Assen et al., 2015). Other methods were developed with the aim to estimate the effect size in a meta-analysis corrected for publication bias: trim and fill (Duval & Tweedie, 2000a, 2000b), PET-PEESE (Stanley & Doucouliagos, 2014), *p*-uniform (van Aert, Wicherts, et al., 2016; van Assen et al., 2015), *p*-curve (Simonsohn et al., 2014a), selection model approaches (for an overview see Hedges & Vevea, 2005; Jin, Zhou, & He, 2014; Sutton, Song, Gilbody, & Abrams, 2000), and methods based on the 10% most precise effect size estimates in a meta-analysis (Stanley et al., 2010). In this chapter, we focus on estimating effect size corrected for publication bias, because publication bias tests have low statistical power especially if the number of effect sizes in a meta-analysis is small (Begg & Mazumdar, 1994; Sterne et al., 2000; van Assen et al., 2015). Furthermore, we believe that from the perspective of an applied researcher it is more relevant to know what the effect size is corrected for publication bias than to know that publication bias distorted the results of a meta-analysis without knowing the consequences on effect size estimation. More specifically, we focus in this chapter on selection model approaches to correct for publication bias, because recently published work suggest that these methods have better statistical properties than other methods (Carter et al., 2017; McShane et al., 2016), and are even nowadays called the state-of-the-art methods to correct for publication bias (McShane et al., 2016).

Selection model approaches combine two models to correct for publication bias: an effect size model and a selection model. The *effect size model* is the distribution of primary studies' effect sizes in the absence of publication bias and the *selection model* determines how the effect size model is affected by publication bias (Hedges & Vevea, 2005). The selection model is actually a set of weights that reflects the likelihood of studies getting published. Effect sizes that are unlikely to be published according to the selection model get more weight compared to effect sizes that are more likely to be published in order to correct for publication bias. Several types of selection model approaches have been proposed, varying from approaches that estimate the selection model, via those that assume a specific selection model and from frequentist to Bayesian approaches (e.g., Cleary & Casella, 1997; Copas & Shi, 2000; Iyengar & Greenhouse, 1988a; Kicinski, 2013; Vevea & Woods, 2005).

Selection model approaches have hardly been used in meta-analyses (Hunter & Schmidt, 2015), because these methods require the user to make sophisticated assumptions and choices (Borenstein et al., 2009), suffer from convergence problems if the number of effect sizes in a meta-analysis is less than 100 (Field & Gillett, 2010), and are often not implemented in user-friendly software for applying these methods. However, easy-to-use software has recently been developed that can be used for applying several types of selection model approaches (R packages "weightr" (Coburn & Vevea, 2016), "selectMeta" (Rufibach, 2015), and "metasens" (Schwarzer, Carpenter, & Rücker, 2016). Convergence problems also seem less of a concern than earlier

stated since two recent simulation studies (i.e., Carter et al., 2017; McShane et al., 2016) that included the selection model approach of Iyengar and Greenhouse (1988a) showed that convergence problems only arose in meta-analyses with (i) extreme publication bias or (ii) only ten studies included in a meta-analysis. However, especially such meta-analyses are quite common in the psychological literature that shows signs of extreme publication bias as evidenced by over 95% articles showing positive outcomes (Fanelli, 2010b). Moreover, the median number of studies in meta-analyses in psychology equals 12 (van Erp, Verhagen, Grasman, & Wagenmakers, 2017). In the medical literature, the positive result rate is also high at approximately 89% (Fanelli, 2010b), while the median number of studies in medical meta-analyses equals 3 (Rhodes et al., 2015; Turner et al., 2015). Hence, we also study in this chapter the statistical properties of selection model approaches for conditions that are realistic for meta-analyses in practice, even though these conditions may at the same time be challenging for applying these methods.

Two recently developed methods to correct effect sizes for publication bias are *p*-uniform (van Aert, Wicherts, et al., 2016; van Assen et al., 2015) and *p*-curve (Simonsohn et al., 2014a). *P*-uniform and *p*-curve are based on the same methodology but slightly differ in implementation (for a comparison of the two methods see van Aert, Wicherts, et al., 2016). These methods use the statistical principle that the *p*-values should be uniformly distributed at the true effect size. Estimation is only based on the statistically significant effect sizes and nonsignificant effect sizes are discarded. For that reason, conditional probabilities (i.e., *p*-values conditional on being statistically significant) are evaluated for being uniformly distributed instead of the traditional *p*-values themselves. The effect size estimate of *p*-uniform and *p*-curve is equal to the effect size where a statistic, which is used for assessing whether these conditional *p*-values are uniformly distributed, equals its expected value.

Three major drawbacks of *p*-uniform and *p*-curve in their current implementation are that (i) the methods only use statistically significant effect sizes which makes the methods inefficient (i.e., estimates often have large variance), (ii) effect size estimates are positively biased in the presence of between-study variance in true effect size (Carter et al., 2017; McShane et al., 2016; van Aert, Wicherts, et al., 2016), and (iii) they do not estimate and test for the presence of this between-study variance. In this chapter, we propose a revised method called *p*-uniform* that solves all three drawbacks: statistically nonsignificant effect sizes are also included in the estimation with *p*-uniform* (i) making it a more efficient estimator than *p*-uniform. (ii) eliminating the overestimation of effect size in case of between-study variance in true effect size, and (iii) enabling estimation and testing for the presence of the between-study variance in true effect size. *P*-uniform* can be seen as a selection model approach where the selection model has one cut-off at the critical value determining whether an effect size is statistically significant or not. An advantage of *p*-uniform* over other selection model approaches is that its selection model does not

have to be estimated or assumed to be known. *P*-uniform* only assumes that the probability that all statistically significant effect sizes get published is the same and also that the probability for publication of the nonsignificant effect sizes is the same. Hence, *p*-uniform* makes fewer assumptions than other selection model approaches.

The goal of this chapter is twofold; we introduce the new method *p*-uniform* and examine the statistical properties of *p*-uniform* and the selection model approach of Hedges (1992) via an analytical study and Monte-Carlo simulations. We compare *p*-uniform* with the selection model approach of Hedges (1992) for four reasons. First, Hedges' method enables estimation of both the effect size as well as the between-study variance in true effect. Second, this selection model approach assumes that the selection model is unknown and has to be estimated which is more realistic than other methods (e.g., Vevea & Woods, 2005) that assume that the selection model is known. Third, easy-to-use software is available for applying this method in the R package "weightr" (Coburn & Vevea, 2016) and this method suffers less from convergence problems than for instance the selection model approach proposed by Copas and colleagues (Copas, 1999; Copas & Shi, 2000, 2001). Finally, statistically significant and nonsignificant effect sizes can be included in this selection model approach whereas other approaches only use the statistically significant effect sizes (e.g., Hedges, 1984).

The remainder of this chapter is structured as follows. We continue with explaining selection model approaches in general and their development. Subsequently, we introduce and explain the extended and improved *p*-uniform* method. Then we present the results of the analytical study and Monte-Carlo simulations for examining the statistical properties of *p*-uniform* and the selection model approach by Hedges (1992). We conclude with a discussion in the final section of this chapter.

5.1 Selection model approaches

All selection method approaches share the common characteristic that they combine an effect size model and selection model to correct for publication bias. The effect size model is usually either the fixed-effect or random-effects model. The random-effects model assumes that *k* independent effect sizes estimates, y_i with *i*=1, …, *k*, are extracted from primary studies. The random-effects model can be written as

$$y_i = \mu + \mu_i + \varepsilon_i$$

where μ is the average true effect size, μ_i is a random effect that denotes the difference between μ and the *i*th primary study's true effect size, and ε_i is the *i*th primary study's sampling error. In the random-effects model, it is commonly assumed that $\mu_i \sim N(0, \tau^2)$ where τ^2 reflects the between-study variance in true effects, and $\varepsilon_i \sim N(0, \sigma_i^2)$ where σ_i^2 is the sampling variance of the *i*th primary study. The μ_i and

ε_i are assumed to be mutually independent of each other, and σ_i^2 is estimated in practice and then assumed to be known. If $\tau^2 = 0$, there is no between-study variance in the true effect size, and the random-effects model collapses to the fixed-effect model.

The selection model is a non-negative weight function that determines the likelihood that a primary study gets published (Hedges & Vevea, 2005). The major difference between the selection model approaches is the weight functions that they use. This weight function can be estimated based on the y_i and its standard error or on the *p*-value of the primary studies. Another option is to assume that a specific weight function is known and use this weight function as selection model. We will describe the different weight functions in more detail when we describe different selection model approaches.

The weight function, $w(y_i, \sigma_i)$, is combined with the effect size model to get a weighted density of y_i,

$$\frac{w(y_i, \sigma_i) f(y_i, \sigma_i)}{\int w(y_i, \sigma_i) f(y_i, \sigma_i) dy_i} \quad (1)$$

where $f(y_i, \sigma_i)$ denotes the (unweighted) density distribution as in the fixed-effect or random-effects model. If $w(y_i, \sigma_i) = 1$ for all y_i, the weighted density is the same as the density of the effect size model (Hedges & Vevea, 2005) and estimates of the selection model approach coincide with those of the fixed-effect or random-effects model. This weighted density can be used to estimate parameters (e.g., μ, τ^2, and parameters in the weight function) in a selection model approach by means of maximum likelihood estimation.

The first selection model approach is proposed in Hedges (1984). This method discards statistically nonsignificant effect sizes and assumes that all significant effect sizes get published. Hence, the weight function of this selection model approach is

$$w(y_i, \sigma_i) = \begin{cases} w_1 = 1, & if\ 1 - \Phi(y_i / \sigma_i) \leq \alpha, \\ w_2 = 0, & if\ 1 - \Phi(y_i / \sigma_i) > \alpha \end{cases},$$

where Φ is the cumulative standard normal distribution function; one-tailed *p*-values of primary studies smaller than α are statistically significant effect sizes and are assigned a weight of 1, whereas nonsignificant effect sizes are assigned a weight of 0. This selection model approach uses the fixed-effect size model, and therefore has only one parameter (μ).

Iyengar and Greenhouse (1988a) extended Hedges' selection model approach by also taking statistically nonsignificant effect sizes into account. They also used the

fixed-effect model, but suggested to extend their model to a random-effects model in the rejoinder of comments on their paper (Iyengar & Greenhouse, 1988b). They suggested two different selection models that, similar to Hedges (1984), both assume that all statistically significant effect sizes get published. One selection model assumes a constant probability of publication for nonsignificant effect sizes

$$w(x;\, \gamma, df) = \begin{cases} w_1 = 1, & if\ |x| > t(df, \alpha), \\ w_2 = e^{-\gamma}, & if\ |x| \leq t(df, \alpha)' \end{cases}$$

where x is the observed *t*-value of a primary study and $t(df, \alpha)$ is the critical *t*-value for a particular α -level and *df* degrees of freedom (i.e., the *t*-value that determines the threshold of an effect size being statistically significant or not). The other selection model assumes that the probability of publication of a nonsignificant effect size increases as the primary study's *t*-value approaches the critical *t*-value

$$w(x;\, \beta, df) = \begin{cases} w_1 = 1, & if\ |x| > t(df, \alpha), \\ w_2 = \frac{|x|^{\beta}}{t(df,\alpha)^{\beta}}, & if\ |x| \leq t(df, \alpha). \end{cases}$$

If γ and β are zero, there is no publication bias and w_1 and w_2 both equal 1. The selection model approach by Iyengar and Greenhouse (1988a) is a two-parameter model (i.e., parameters are μ and either γ or β depending on which selection model is selected) whereas the selection model approach proposed in the rejoinder (Iyengar & Greenhouse, 1988b) is a three-parameter model (i.e., parameters are μ, τ^2, and either γ or β depending on which selection model is selected).

Hedges (1992) generalized the original model of Iyengar and Greenhouse (1988a) to the random-effects model. In contrast to Iyengar and Greenhouse (1988a), this selection model approach does not use a parametric selection model. These approaches use a step function based on primary studies' *p*-values to create a weight function. That is, the steps create intervals of *p*-values, and effect sizes with *p*-values that fall into the same interval get the same weight in the weight function. The probability of publication for each interval of *p*-values is estimated and these probabilities are used in the weight function. The user of the selection model approach of Hedges (1992) has to specify the location of the steps that determine the intervals of the *p*-values to estimate the probabilities that are used in the weight function. Let a_{j-1} denote the left and a_j the right endpoint of an interval of *p*-values where *j* refers to the *j*th interval and $a_0 = 0$ and $a_J = 1$ with *J* reflecting the total number of intervals. The weight function (Hedges, 1992; Hedges & Vevea, 2005) can then be written as

$$w(y_i, \sigma_i) = \begin{cases} w_1, & if -\sigma_i \Phi^{-1}(a_1) < y_i \leq \infty, \\ w_j, & if -\sigma_i \Phi^{-1}(a_j) < y_i \leq -\sigma_i \Phi^{-1}(a_{j-1}), \\ w_J, & if -\infty < y_i \leq -\sigma_i \Phi^{-1}(a_{J-1}) \end{cases}.$$

It is common practice to set w_1 to 1 in order to get relative weights that facilitate the interpretation of the weight function. For example, suppose that the location of only one step is specified at a *p*-value of 0.05 resulting in two intervals of *p*-values (0-0.05 and 0.05-1) and two weights (w_1 and w_2) are estimated. The weight corresponding to the interval 0-0.05 will be set equal to $w_1 = 1$ and if $w_2 = 0.5$ this is interpreted as primary studies with *p*-values in the interval between 0.05-1 being half as likely to be observed as primary studies with *p*-values between 0.05-1.

Hedges and Vevea (1996) conducted a simulation study to examine the statistical properties of this selection model approach and concluded that the weights of the selection model are often poorly estimated, but that the estimates of effect size and between-study variance were then still quite accurate. Furthermore, they assessed whether non-normally distributed μ_i bias the estimates of the selection model approach and concluded that the approach is relatively robust to violations of the normality assumption. The selection model approach by Hedges (1992) was later extended to enable the inclusion of predictor variables (see Vevea & Hedges, 1995) such that the effect size model is a mixed-effects model instead of a random-effects model (Borenstein et al., 2009). This selection model approach of Hedges (1992) is implemented in the R package "weightr". In order to avoid non-convergence of this method, the weight of an interval is set equal to 0.01 in the R package if there are no observations in a particular interval.

Vevea and Woods (2005) proposed not to estimate the selection model as in the approach of Hedges (1992), but rather to assume that the selection model is known. This enables application of the selection model approach of Hedges (1992) in meta-analyses with a small number of primary studies' effect sizes, because 100 to 200 primary studies have to be included in the meta-analysis with at least 10 to 15 *p*-values in each interval of *p*-values to accurately estimate the weights (Hedges & Vevea, 1996, 2005; Vevea & Woods, 2005). Vevea and Woods (2005) proposed four different weight functions that differ in the severity of publication bias and also in whether one or two-tailed tests were conducted in the primary studies. A drawback of the approach of Vevea and Woods is that bias in the estimates of the effect size and between-study variance is introduced if the weights are incorrectly specified.

Another selection model approach assumes that selection occurs through both the effect size and standard error of primary studies (Copas, 1999; Copas & Li, 1997; Copas & Shi, 2000, 2001). This selection model approach makes use of a selection variable

$$z_i = a + b/\sigma_i + \delta_i$$

where a and b are two parameters that are estimated and δ_i is a normally distributed random variable. Parameters a and b determine the probability of publication of an effects size with a determining the minimal probability of a study being published and b the change in a if σ_i increases or decreases. The weights of each primary study's effect size are then determined based on the correlation between z_i and y_i where a non-zero correlation indicates that publication bias occurred. The effect size model of this method is the random-effects model or mixed-effects model if predictor variables are included. This method estimates μ, τ^2, and a and b in the weight function, but it is sometimes not possible to estimate all the parameters in the model since the weight function cannot always be estimated (Hedges & Vevea, 2005). However, parameters in the weight function can also assumed to be known to avoid convergence problems (Copas, 1999; Copas & Shi, 2000, 2001).

Several others have developed selection model approaches based on Bayesian statistics (e.g., Givens, Smith, & Tweedie, 1997; Kicinski, 2013; Larose & Dey, 1998; Silliman, 1997). However, these approaches are considered to be more complicated and require more assumptions or choices of users than previously discussed selection model approaches (Jin et al., 2014; Sutton et al., 2000). A more recently proposed selection model approach by Guan and Vandekerckhove (2015) uses Bayesian model averaging to correct for publication bias. The effect size model of this method is the fixed-effect model and four different selection models that differ in the degree of publication bias are considered. The effect size model and the selection models are combined and subsequently Bayesian model averaging over the four selection models is used to estimate μ. Another recently proposed selection model approach is the Bayesian fill-in meta-analysis (BALM) method (Du, Liu, & Wang, 2017). This method is a Bayesian implementation of the selection model approach of Hedges (1992) using non-informative or weakly informative prior distributions. Estimates for μ, τ^2, and the weight function are obtained using a Gibbs sampling algorithm.

The statistical properties of previously discussed selection model approaches have not been directly compared in for instance Monte-Carlo simulation studies. However, a recent simulation study by McShane et al. (2016) compared the selection model approach of Iyengar and Greenhouse (1988b) with other methods to estimate effect size corrected for publication bias (i.e., trim-and-fill, *p*-uniform, *p*-curve, and PET-PEESE). They concluded that the selection model approach of Iyengar and Greenhouse (1988b) yields the best performance and should be preferred over the other methods, but also noted that the assumptions underlying this selection model approach are idealistic and unlikely to be met in practice. Du et al. (2017) compared their BALM method with trim-and-fill, PET-PEESE, *p*-uniform, and the selection model

of Hedges (1992). They concluded that BALM had better statistical properties than all the included methods except for slightly more bias in the estimate of μ than the selection model approach of Hedges (1992). However, coverage probabilities of BALM were closer to the nominal coverage rate than Hedges' (1992) selection model approach, and the selection model approach suffered from convergence problems. Another simulation study (Terrin et al., 2003) studied the statistical properties of Hedges' (1992) selection model approach and compared it to trim-and-fill. Although this selection model approach outperformed trim-and-fill, it also often failed to converge. Non-convergence of Hedges' (1992) approach happens when there are no *p*-values observed in one of the specified intervals of *p*-values. Although recent studies suggest better performance of selection model approaches, it is currently unknown how these methods perform in meta-analyses with a small number of primary studies' effect sizes in combination with extreme publication bias.

5.2 From *p*-uniform to *p*-uniform*

We continue with explaining *p*-uniform and then explain how we extended the method to *p*-uniform* in order to (i) make the estimator more efficient than *p*-uniform, (ii) eliminate the overestimation of effect size of *p*-uniform in case of between-study variance in true effect size, and (iii) enable estimation and testing for the presence of the between-study variance in true effect size. Estimation in *p*-uniform was done by method of moments estimation, but we implemented maximum likelihood estimation for *p*-uniform*.[14]

5.2.1 *P*-uniform

P-uniform (van Aert, Wicherts, et al., 2016; van Assen et al., 2015) uses the statistical principle that *p*-values are uniformly distributed at the true effect size. The method discards statistically nonsignificant effect sizes and only uses the significant effect sizes to correct for publication bias. Assumptions of the method are that a fixed true effect underlies the primary studies included in the meta-analysis and that all primary studies' effect sizes that are statistically significant using a one-tailed test have an equal probability of getting published. Statistical significance is taken into account -and hence publication bias is corrected- by computing probabilities of observing an effect size or larger conditional on the effect size being statistically significant. This can be written as

[14] In the paper, based on this chapter we will also include the results of method of moments estimators.

$$q_i = \frac{1 - \Phi\left(\frac{y_i - \mu}{\sigma_i}\right)}{1 - \Phi\left(\frac{y_{cv} - \mu}{\sigma_i}\right)} \quad (2)$$

where the numerator is the probability of observing an effect size at the true effect size larger than the effect size in the *i*th primary study and the denominator is the probability of observing a (statistically significant) effect size (i.e., larger than the critical value y_{cv}). If $\mu = 0$, the denominator of Equation (2) is equal to the α-level for a one-tailed test.

P-uniform can also be seen as a selection model approach with the fixed-effect model as effect size model, and a selection model assuming equal weights for statistically significant effect sizes to get published. Discarding nonsignificant effect sizes in *p*-uniform is tantamount to assuming a constant probability to get published, but without estimating and the need to estimate this probability. *P*-uniform is closely related to the selection model approach proposed by Hedges (1984) with two differences. First, Hedges' (1984) approach uses two-tailed *p*-values instead of one-tailed *p*-values in *p*-uniform. Second, Hedges' (1984) approach uses maximum likelihood estimation, whereas *p*-uniform uses method of moments estimation.

P-uniform's effect size estimate is equal to the value of μ where a statistic that is computed based on the q_i that equals its expected value assuming a uniform distribution. Van Assen et al. (2015) proposed to use Fisher's test (Fisher, 1925), $-\sum_{i=1}^{k} \ln(q_i)$, to estimate the effect size in *p*-uniform. Since the distribution of Fisher's test is $\Gamma(k,1)$, $\hat{\mu}$ is the value for μ where the test statistic of the Fisher's test is equal to its expected value (i.e., *k*). A 95% confidence interval for μ is computed by entering different values for μ in Equation (2) (i.e., profiling) until the test statistic of the Fisher's test is equal to the 2.5th and 97th percentile of the gamma distribution. This confidence interval is exact (i.e., 95% coverage probability) if the assumptions of the method hold. The null hypothesis of no effect and the presence of publication bias are tested by examining whether the q_i at $\mu = 0$ and at the estimate of the fixed-effect model deviate from a uniform distribution using the Fisher's test (van Assen et al., 2015). Van Assen et al. (2015) compared *p*-uniform with the Fisher's test as method for estimating the effect size with trim-and-fill (Duval & Tweedie, 2000a, 2000b) to correct for publication bias and concluded that *p*-uniform outperformed trim-and-fill if publication bias exists and between-study variance in true effect size is absent or small. However, they also showed that overestimation of *p*-uniform increased as a function of the between-study variance in true effect size, and van Aert et al. (2016) showed that in case of *p*-hacking (a.k.a. questionable research practices or

opportunistic use of researcher degrees of freedom) *p*-uniform may bias effect size estimation, where size and sign of the bias depend on the type of *p*-hacking. A comparison of *p*-uniform's publication bias test with the test of excess significance (Ioannidis & Trikalinos, 2007b) revealed that *p*-uniform's publication bias test has better statistical properties except for situations with a small amount of publication bias in combination with a zero true effect size.

Another proposed method[15] for estimating the effect size with *p*-uniform is based on the distribution of the sum of independently uniformly distributed random variables, which is called the Irwin-Hall distribution (van Aert, Wicherts, et al., 2016). The expected value of the Irwin-Hall distribution is $0.5k$, so $\hat{\mu}$ is that value of μ for which $\sum_{i=1}^{k} q_i = 0.5k$. An exact 95% confidence interval is again computed by profiling Equation (2) until $\sum_{i=1}^{k} q_i$ is equal to the 2.5th and 97.5th percentile of the Irwin-Hall distribution. The null hypotheses of no effect and no publication bias are rejected if $\sum_{i=1}^{k} q_i$ is larger than the critical value of the Irwin-Hall distribution at $\mu = 0$ and the estimate of the fixed-effect model, respectively. Van Aert et al. (2016) recommended to use the estimator based on the Irwin-Hall distribution as default estimator, because (i) summing q_i is easy to understand, and (ii) it has the nice property that $\hat{\mu}$ is equal, larger, smaller than zero if the average of the statistically significant *p*-values is equal to, smaller, larger than α if one-tailed tests or $\alpha / 2$ if two-tailed tests were used in the primary studies. Moreover, the estimator based on the Irwin-Hall distribution is less susceptible to outlying effect sizes than the estimator using the Fisher's test.

McShane et al. (2016) criticizes *p*-uniform for three reasons: (i) *p*-uniform assumes a fixed true effect underlying all y_i, (ii) it discards statistically nonsignificant effect sizes, and (iii) *p*-uniform uses method of moments estimators instead of maximum likelihood estimation. The first critique is indeed a limitation of *p*-uniform, because assuming that the true effect size is fixed results in overestimated effect size if this assumption is violated (Carter et al., 2017; McShane et al., 2016; van Aert, Wicherts, et al., 2016; van Assen et al., 2015). Hence, we recommended to only interpret *p*-uniform's effect size estimate as the estimate of the average population effect size when heterogeneity is at most moderate (van Aert, Wicherts, et al., 2016). Moreover, methods that do not assume that the true effect size is fixed are favorable,

[15] Effect sizes in *p*-uniform its current implementation can be estimated using six different methods: maximum likelihood, and using the Irwin-Hall distribution, Fisher's test, the Kolmogorov-Smirnov test, Anderson-Darling test, and a variant of Fisher's test, $-\sum_{i=1}^{k} \ln(1 - q_i)$. All methods are implemented in the R package "puniform" that is available on GitHub (https://github.com/RobbievanAert/puniform).

because heterogeneity is often present in meta-analyses (e.g., Higgins, 2008; Higgins et al., 2009; Klein et al., 2014; van Erp et al., 2017).

The second critique by McShane et al. (2016) relates to the loss of efficiency of *p*-uniform's estimators (van Aert, Wicherts, et al., 2016; van Assen et al., 2015) because of discarding nonsignificant effects. Efficiency loss may be limited in many applications, as the vast majority of published results in psychology are statistically significant (e.g., Fanelli, 2010b; Fanelli, 2012). Nevertheless, as meta-analyses including many nonsignificant effects do occur (e.g., van Aert, Wicherts, & van Assen, 2018) and the potential efficiency loss is both particularly prevalent and detrimental in the important cases where average true effect size is (close to) zero, we generalized our methodology and also include statistically nonsignificant effect sizes in *p*-uniform*. Simonsohn, Simmons, and Nelson (2017, December 20), however, argue that nonsignificant effects should not be included in *p*-curve (and therefore also in *p*-uniform). They argue that all nonsignificant effects do not have the same probability of getting published, and it is hard to make assumptions about this probability. However, we contend that the benefits of including statistically nonsignificant effect sizes (more efficient estimator, less biased estimator if the true effect size is heterogeneous, and enabling estimation of the between-study variance in true effect size) outweigh the potential costs (possible bias in the estimator if the assumption of equal probability for publishing nonsignificant effect sizes is violated).

The final critique by McShane et al. (2016) relates to the optimal large-sample properties of the maximum likelihood estimator compared to method of moment estimators (e.g., Casella & Berger, 2002). Although van Aert, Wicherts, et al. (2016) and van Assen et al. (2015) were aware of these large-sample optimal properties, they intentionally selected method of moments estimators because these yield exact confidence intervals even if only one statistically significant effect size is included in a meta-analysis. This is in contrast with the conventional Wald-based confidence interval and hypothesis test that is accompanied by maximum likelihood estimation, because these are accompanied with the requirement that the log-likelihood around the maximum likelihood estimate is regular (Pawitan, 2013). As *p*-uniform uses conditional probabilities (given statistical significance) as likelihoods that are truncated, the log-likelihood is not well approximated by the normal distribution for a large and relevant part of the parameter space (parameter values close to 0). Hence, Wald-based confidence intervals and hypothesis tests are generally inappropriate. To bypass problems of non-normally distributed log-likelihoods, we implemented *p*-uniform* using maximum likelihood estimation and computed confidence intervals of μ and τ^2 by inverting the likelihood-ratio test and using the likelihood-ratio test for testing the null hypothesis $\mu = 0$. These procedures do not make use of asymptotic normality distributions as do the Wald-based confidence intervals (Agresti, 2013; Pawitan, 2013) and are therefore expected to have better statistical properties.

5.2.2 *P*-uniform*

P-uniform* is a selection model approach with the random-effect model as effect size model. The selection model assumes that the probability of publishing a statistically significant effect size as well as a nonsignificant effect size are constant, but these probabilities may be different from each other. Other selection model approaches estimate the probabilities of publication for studies with particular effect sizes and use these probabilities in the weight function for effect size estimation. However, these weight functions are often poorly estimated resulting in bias in the estimates of these selection model approaches (Hedges & Vevea, 1996; Vevea & Woods, 2005). Hence, an advantage of *p*-uniform* over other selection model approaches is that *p*-uniform* does not require estimating these probabilities, but only treats the primary studies' effect sizes differently depending on whether they are statistically significant or not.

Maximum likelihood estimation is used in *p*-uniform* where truncated densities are being used instead of the conditional probabilities in Equation (2). Truncated densities ($q_i^{ML^*}$) are computed for both the statistically significant and nonsignificant effect sizes and are a function of both μ and τ^2,

$$q_i^{ML^*} = \begin{cases} \dfrac{\phi\left(\dfrac{y_i - \mu}{\sqrt{\sigma_i^2 + \tau^2}}\right)}{1 - \Phi\left(\dfrac{y_{cv} - \mu}{\sqrt{\sigma_i^2 + \tau^2}}\right)} & if\ p_i \leq \alpha, \\ \dfrac{\phi\left(\dfrac{y_i - \mu}{\sqrt{\sigma_i^2 + \tau^2}}\right)}{\Phi\left(\dfrac{y_{cv} - \mu}{\sqrt{\sigma_i^2 + \tau^2}}\right)} & if\ p_i > \alpha \end{cases}$$

where ϕ denotes the standard normal probability density function. The likelihood function is the product of the $q_i^{ML^*}$:

$$L(\mu, \tau^2) = \prod_{i=1}^{k} q_i^{ML^*} \cdot \qquad (3)$$

The profile (log-)likelihood functions of Equation (3) can be iteratively optimized until $\hat{\mu}$ and $\hat{\tau}^2$ do not change anymore in consecutive steps. Confidence intervals for μ and τ^2 are obtained by inverting the likelihood-ratio test statistic, and the likelihood-ratio test is used to test the null hypotheses $\mu = 0$ and $\tau^2 = 0$ (Agresti, 2013; Pawitan, 2013).

5.3 Analytical study

We continue by evaluating the statistical properties of *p*-uniform* and the selection model approach of Hedges (1992) implemented in the R package "weightr" (Coburn & Vevea, 2016; henceforth called Hedges1992) by means of an analytical study when there is only one statistically significant and one nonsignificant observed effect size. This seems to be a rather extreme situation for a meta-analysis, but many meta-analyses only contain a small number of primary studies' effect sizes (median number of primary studies' effect sizes in a meta-analysis is 3 in medicine (Rhodes et al., 2015; Turner et al., 2015). Moreover, the statistical properties of these methods have never been examined in such an extreme situation, provided useful insights in the statistical properties of the different methods, and enabled us to evaluate the statistical properties of the methods without the necessity of using Monte-Carlo simulations.

5.3.1 Method

We selected the standardized mean difference as effect size measure for the analytical study, but the included methods can also be applied with other effect size measures. The methods were applied to the joint probability density function (pdf) of one statistically significant and one nonsignificant effect size. The pdf of the statistically significant effect size based on a one-tailed test with $\alpha = .05$ was approximated by selecting 1,000 equidistant cumulative probabilities given that the effect size had a *p*-value smaller than $\alpha = .05$. Hence, the cumulative probabilities ranged from $1 - \pi + \frac{\pi}{1{,}001}$ until $1 - \pi + \frac{1{,}000 \times \pi}{1{,}001}$ where π is the statistical power of the test of no effect. For example, under the null hypothesis ($\mu = 0$) this yielded 1,000 cumulative probabilities ranging from $1 - 0.05 + \frac{0.05}{1{,}001} = 0.95005$ until $1 - 0.05 + \frac{1{,}000 \times 0.05}{1{,}001} = 0.99995$. These cumulative probabilities were then transformed to standardized mean differences using the quantile function of the normal distribution to approximate the pdf of the effect size. Similarly, the pdf of the statistically nonsignificant effect sizes was approximated by selecting 1,000 equidistant cumulative probabilities with the requirement that the *p*-value of these effect sizes was larger than $\alpha = .05$, which were subsequently transformed into standardized mean differences.

The two pdfs were combined to obtain the joint probability density distribution consisting of 1,000 x 1,000 = 1,000,000 combinations of statistically significant and nonsignificant effect sizes. Effect size was estimated for each combination using two methods; *p*-uniform* and Hedges1992.

Two intervals for the weights function of Hedges1992 were imposed with as threshold $\alpha = .05$. This was realistic since the likelihood of publishing a primary study is often determined based on whether a primary study's *p*-value is smaller than $\alpha = .05$. Moreover, it increases the comparability with *p*-uniform*, because *p*-uniform* also treats primary studies' effect sizes differently depending on whether an effect size is statistically significant or not. The estimates of *p*-uniform* were obtained by optimizing the profile log-likelihood function for μ on the interval (μ -4; μ +4) and for τ on the interval (0; τ +1).

We evaluated the statistical properties of the different methods for both μ and τ with respect to average, median, and standard deviation of the estimates, root mean square error (RMSE), and coverage probability (i.e., how frequent μ or τ fell in their respective confidence interval) and average width of the 95% confidence intervals for μ and τ.[16] A two-independent groups design was used for the analytical study with a sample size of 50 per group. Two values for the true effect size ($\mu = 0$ and 0.5) were selected, and two values for the square root of the between-study variance ($\tau = 0$ and 0.346) corresponding to I^2-statistics equal to 0% (no heterogeneity) and 75% (large between-study variance) (Higgins et al., 2003). This analytical study was programmed in R (R Core Team, 2017) and the packages "metafor" (Viechtbauer, 2010) and "weightr" (Coburn & Vevea, 2016) were used for applying the random-effects model and Hedges1992, respectively. R code of the analytical study is available via https://osf.io/qh7tj/.[17]

5.3.2 Results

Table 5.1 presents the results of the average and standard deviation of the estimates, root mean square error (RMSE), and coverage probability of the 95% confidence intervals for μ and τ. The median of the estimates for μ and τ are not reported, because these results were highly comparable to the average of the

[16] Confidence intervals of the random-effects model for μ were computed using the adjustment proposed by Hartung and Knapp (Hartung, 1999; Hartung & Knapp, 2001a, 2001b) and Sidik and Jonkman (Sidik & Jonkman, 2002), because coverage probability after applying this adjustment is closer to the nominal coverage rate (IntHout, Ioannidis, & Borm, 2014; Röver, Knapp, & Friede, 2015; Wiksten, Rücker, & Schwarzer, 2016). Veroniki et al. (2016) reviewed existing methods to compute a confidence interval for τ^2 and concluded that the *Q*-profile method (Viechtbauer, 2007b) and generalized *Q*-statistic method (Jackson, 2013) are the two methods with the best statistical properties. We decided to include the *Q*-profile method in our Monte-Carlo simulations, because this method does not require arbitrary choices with respect to the primary study's weights as compared to the generalized *Q*-statistic method.

[17] In order to verify whether approximating each marginal pdf with 1,000 equidistant cumulative probabilities was sufficient, we also conducted a Monte-Carlo simulation study with the same conditions as in the analytical study, using 10,000 replications (R code https://osf.io/phw7z/). As the results of this Monte-Carlo simulation study were highly similar to those of the analytical study we only report the results of this simulation study in the supplemental materials (https://osf.io/kyv6b/).

estimates for both parameters. The average width of the confidence intervals for μ and τ is also not discussed since coverage probability of the methods often substantially deviated from the nominal coverage rate such that interpreting the width of the confidence intervals was inappropriate. All the omitted results, however, are reported in the supplemental materials (https://osf.io/kyv6b/).

Estimating μ and its confidence interval *P*-uniform* always converged with respect to estimating μ, but Hedges1992 did not converge in at most 194 out of the 1,000,000 combinations (0.02%) of a statistically significant and nonsignificant effect size (condition $\mu = 0.5$ and $\tau = 0.346$). The first four columns of Table 5.1 present the results for estimating μ and computing its confidence interval. Although bias of *p*-uniform* and Hedges1992 was small (at most 0.062), both methods overestimated μ if $\mu = 0$ and underestimated μ if $\mu = 0.5$. Estimates of *p*-uniform* and Hedges1992 were highly similar, but estimates of *p*-uniform were closest to the true effect size if $\mu = 0$ whereas Hedges1992 was slightly less biased if $\mu = 0.5$ in combination with $\tau = 0.346$.

The standard deviation of the estimates of *p*-uniform* for $\mu = 0$ were slightly larger than for Hedges1992. This was caused by some extremely negative estimates of *p*-uniform* when *p*-values of the statistically significant effect size was close to $\alpha = .05$. Consequently, this also affected the RMSE of *p*-uniform* that was also slightly larger than the RMSE of the selection model approach if $\mu = 0$. For other conditions, the standard deviation of the estimates and the RMSE of both methods were highly comparable.

Coverage probability of the 95% confidence interval for μ could always be computed for *p*-uniform*, but not in at most 1.5% of the combinations for Hedges1992. Coverage probabilities of *p*-uniform* were acceptable (.94-.96) if $\tau = 0$, but too low if $\tau = 0.346$ (around .818). Similarly, coverage probabilities of Hedges1992 were acceptable for $\mu = 0.5$, close to acceptable for $\mu = 0$ (.971) if $\tau = 0$, but too low if $\tau = 0.346$ (.84 and .81).

Estimating τ and its confidence interval An estimate of τ could always be computed for *p*-uniform* whereas estimation with Hedges1992 did not converge in at most 0.03% of the combinations (condition $\mu = 0.5$ and $\tau = 0.346$). The last four columns of Table 5.1 show the results of estimating τ and computing a confidence interval for τ. Estimates for τ and the standard deviation of these estimates of *p*-uniform* and Hedges1992 were highly similar to each other except for $\mu = 0$ and $\tau = 0.346$ (*p*-uniform* 0.167 vs. selection model approach 0.185). Both methods yielded accurate estimates for $\tau = 0$, but τ was severely underestimated for $\tau = 0.346$. This underestimation was not surprising since estimating the between-study variance in true effect sizes based on only two primary studies is very

challenging. The RMSEs for estimating τ were comparable for the two methods.

Coverage probabilities could not be computed for the selection model approach in at most 0.02% of combinations (condition $\mu = 0.5$ and $\tau = 0.346$). Surprisingly, results for *p*-uniform* and the selection model approach were very different. While all their coverage probabilities were seriously off, *p*-uniform*'s coverage was too high (≥.995) because of too wide confidence intervals and those of the selection model approach were too low (≤ .55) because of the bias in the estimates and too small confidence intervals.

Conclusion The analytical study demonstrates that convergence for estimating μ and τ and their confidence interval was not a problem for *p*-uniform* and hardly a problem for Hedges1992 in very challenging conditions with only one significant and one nonsignificant effect, invalidating the critique on the selection model approach that at least 100 primary studies' effect sizes are required for estimates to converge (Field & Gillett, 2010; Hedges & Vevea, 2005; Vevea & Woods, 2005). Estimates of μ for both methods were highly comparable and the bias was small, but both methods underestimated τ if $\tau = 0.346$ and provided highly inaccurate confidence intervals for τ (severe over-coverage for *p*-uniform* and severe undercoverage for Hedges1992). We therefore conclude that although both methods hardly suffer from convergence problems and rather accurately estimate average effect size, two studies are (unsurprisingly) not sufficient for estimating τ and its confidence interval.

5.4 Monte-Carlo simulation study

As analytically approximating the statistical properties of the different methods is numerically too intensive for more than two studies, we also conducted Monte-Carlo simulations.

5.4.1 Method

Standardized mean differences were again the effect size measure of interest using a two-independent groups design with a sample size of 50 per group. First, a true effect size θ_i for the *i*th primary study was sampled from $N(\mu, \tau^2)$. Subsequently, this θ_i was used for generating an observed effect size from $N(\theta_i, \frac{2}{50})$. The observed variance for each group was sampled from a χ^2-distribution using $\frac{\chi^2_{df=49}}{49}$. These observed effect size and variances were used for computing the Cohen's *d* standardized mean difference, which were subsequently transformed into Hedges' *g*

Table 5.1. Results of the analytical study with one statistically significant and one nonsignificant effect size for estimating μ and τ with *p*-uniform* and Hedges1992 as a function of μ and τ. Reported outcomes are average and standard deviation (SD) of the estimates, root mean square error (RMSE), and coverage probability of the 95% confidence interval.

		Estimating μ				Estimating τ			
		$\tau = 0$		$\tau = 0.346$		$\tau = 0$		$\tau = 0.346$	
		$\mu = 0$	$\mu = 0.5$	$\mu = 0$	$\mu = 0.5$	$\mu = 0$	$\mu = 0.5$	$\mu = 0$	$\mu = 0.5$
Average (SD) of estimates	*p*-uniform*	0.014 (0.214)	0.486 (0.213)	0.043 (0.404)	0.475 (0.4)	0.031 (0.073)	0.03 (0.072)	0.167 (0.192)	0.165 (0.192)
	Hedges1992	0.029 (0.193)	0.486 (0.213)	0.062 (0.378)	0.477 (0.393)	0.037 (0.076)	0.03 (0.072)	0.185 (0.189)	0.168 (0.191)
RMSE	*p*-uniform*	214.5	213.1	406	400.3	78.8	78.4	262.5	264
	Hedges1992	195.1	213	383.5	393.8	84.9	78.4	248.3	261.6
Coverage	*p*-uniform*	0.958	0.959	0.818	0.821	0.996	0.996	0.995	0.996
	Hedges1992	0.971	0.949	0.84	0.81	0.446	0.379	0.551	0.536

Note. Root mean square error was multiplied with 1,000

by multiplying the Cohen's *d* effect size with $c(98) = \frac{\Gamma\left(\frac{98}{2}\right)}{\sqrt{\frac{98}{2}}\Gamma\left(\frac{98-1}{2}\right)}$ where 98 refers to the degrees of freedom (Hedges, 1981). The unbiased estimate of the sampling variance of Hedges' *g* (see Equation 26 in Viechtbauer [2007a]) was computed with $\frac{1}{25} + \left(1 - \frac{98-2}{98 \times c(98)^2}\right) g_i$ where g_i denotes the Hedges' *g* effect size of the *i*th primary study.

If the effect size was statistically significant based on a one-tailed test with $\alpha = .05$, the *i*th primary study effect size was included in the meta-analysis. Statistically nonsignificant effect sizes were included in the meta-analysis if a randomly drawn number from a uniform distribution ranging from zero to one was smaller than $1 - pub$, where pub represents the probability of a statistically nonsignificant effect size to be included in a meta-analysis with $pub = 1$ referring to extreme publication bias (only statistically significant studies get published). This procedure for generating data of primary studies was repeated until *k* primary studies' effect sizes were included in a meta-analysis.

The following variables were varied in the Monte-Carlo simulations: μ, τ, *k*, and pub. Three different levels were selected for μ (0; 0.2; 0.5) reflecting no, a small, and a medium effect (Cohen, 1988). The square root of the between-study variance in true effect size (τ) was 0, 0.163, or 0.346 representing I^2-statistics equal to 0%, 40%, and 75% (zero, small-medium, large [Higgins & Thompson, 2002]). The number of effect sizes in a meta-analysis (*k*) was equal to 10, 30, 60, and 120; 10 and 30 are close to the median (12) and mean (38.7) number of effect sizes in meta-analyses in psychology (van Erp et al., 2017), respectively, whereas we also included 60 and 120 because previous research (Field & Gillett, 2010; Hedges & Vevea, 2005; Vevea & Woods, 2005) suggests that a large number of effect sizes in a meta-analysis are required in order for selection model approaches to perform well. Four different levels for pub were selected: 0, 0.5, 0.9, and 1. Combining the different levels of these variables resulted in 3 x 3 x 4 x 4 = 144 conditions. For each condition 10,000

replications were conducted.[18]

P-uniform* and Hedges1992 with two intervals and the threshold at $\alpha = .05$ were applied to each simulated meta-analysis. The random-effects model was also included to be able to compare methods that correct for publication bias with the method that is usually applied and does not correct for publication bias. We used the Paule-Mandel estimator (Paule & Mandel, 1982) to estimate the between-study variance in true effect size, because two recent papers reviewing existing estimators of the between-study variance recommend this estimator (Langan et al., 2016; Veroniki et al., 2016). The outcome variables were the average, median, and standard deviation of the estimates, RMSE, and coverage probability and average width of the 95% confidence intervals for μ and τ. Moreover, we also studied the Type I error rate and statistical power for the test of no effect with $\alpha = .05$.

The Monte-Carlo simulation study was programmed in R (R Core Team, 2017) and the packages "metafor" (Viechtbauer, 2010) and "weightr" (Coburn & Vevea, 2016) were used for the random-effects model and the selection model approach, respectively. Similar to the analytic study, the estimates of *p*-uniform* were obtained by optimizing the profile log-likelihood function for μ on the interval (μ -5; μ +5) and for τ on the interval (0; τ +2). Other R packages that were used to decrease the computing time of the simulations were the "parallel" package (R Core Team, 2017) for parallelizing the simulations and the "Rcpp" package (Eddelbuettel, 2013) for executing C++ functions. R code of this Monte-Carlo simulation study is available via https://osf.io/79k3p/.

5.4.2 Results

The medians of the estimates of μ and τ are not presented, because they were highly similar to their means. The width of the confidence intervals are not presented, because coverage probabilities often substantially deviated from the nominal coverage rate, thereby decreasing the usefulness of assessing the width of confidence intervals. Finally, we only present the results for *k*=10 and 60 in this section, because these conditions already illustrate how the methods' performances

[18] We conducted an additional Monte-Carlo simulation study to examine whether varying the primary study's sample size influenced the results (R code https://osf.io/ms5kn/). The same variables were varied in this simulation study as in the one described above except for that *k* was now fixed to 30. Sample sizes of these 30 primary study's effect sizes were varied such that the median sample size per group of the primary studies included in a meta-analysis was equal to 50. That is, ten of the 30 primary studies' effect sizes were based on a sample size of 25 per group, eight on a sample size of 50 per group, six on a sample size of 100 per group, four on a sample size of 150 per group, and two on a sample size of 300 per group. These sample sizes were chosen in such a way that the median sample size per group of the primary studies included in a meta-analysis was 50. We only report the results of this Monte-Carlo simulation study in the supplemental materials (https://osf.io/kyv6b/), because the results of this simulation study were not remarkably different from the one with fixed sample size.

increase in *k*; the condition *k*=120 was omitted because the methods' performance in that condition was not remarkably different from that in *k*=60. All omitted results are included in the supplemental materials (https://osf.io/kyv6b/).

Average estimates of μ *P*-uniform* and Hedges1992 did not always converge whereas the random-effects model always obtained an estimate of μ. The reason for the non-convergence of *p*-uniform* was that the estimate of *p*-uniform* was equal to one of the boundaries of the parameter space (i.e., μ -5 or μ +5). This non-convergence was most severe for the condition $\mu = 0$, $\tau = 0$, *k*=10, and $pub = 1$ (12.6%). Hedges1992 failed to converge in at most 15.8% of the replications for the condition $\mu = 0$, $\tau = 0$, *k*=10, and $pub = 0$. Both methods' non-convergence rate was close to zero if both statistically significant and nonsignificant primary studies' effect sizes were included in a meta-analysis.

Figures 5.1 and 5.2 show the average of the estimates of μ when μ (columns of the figures), τ (rows of the figures), and pub (*x*-axis of the figures) were varied for *k*=10 and *k*=60, respectively. All the figures are centered at the true effect size μ (dashed gray line) to facilitate comparability of the different subfigures as we varied μ. We first describe the results of *k*=10 and then illustrate how the results change if *k*=60.

Highlighting common issues with a lack of correction or publication bias, the random-effects model overestimated μ under publication bias and this overestimation decreased in μ and increased in τ and pub. Hedges1992 and *p*-uniform* were less biased than the random-effects model if $pub > 0$ with no (i.e., $pub = 0$) or negligible bias (i.e., $pub = 0.5$). For $pub = 0.9$, Hedges1992 provided accurate average estimates (maximum bias 0.056). For $pub = 0.9$, *p*-uniform* also provided accurate average estimates for $\mu = 0$ and $\mu = 0.2$, but slightly underestimated μ when $\mu = 0.5$ in combination with $\tau = 0.346$ (bias = -0.07). Hedges1992 was severely positively biased in case of extreme publication bias (i.e., $pub = 1$; maximum bias 0.329), and this bias decreased in μ. In case of extreme publication bias, *p*-uniform* generally showed less severe bias than Hedges1992 (maximum bias -0.144), and tended to underestimate μ.

While bias in the random-effects model was unaffected by increasing the number of studies to 60 (see Figure 5.2), bias of the two other methods decreased slightly. Bias was negligible for $pub < 1$ with maximum bias equal to 0.032 and 0.031 for *p*-uniform* and Hedges1992, respectively. For $pub = 1$, Hedges1992 still provided strongly overestimated estimates of μ if $\mu = 0$ (bias at most 0.301), whereas *p*-uniform* showed slight underestimation (maximum bias -0.110). Because these biases for $pub = 1$ also hold for *k*=120, revealing systematic bias, these results suggest

that the selection model approach should not be used when a meta-analysis only consists of statistically significant effect sizes, particularly when the true effect size can be expected to be small.

RMSE for estimating μ Figure 5.3 and 4 show the RMSE for estimating μ for *k*=10 and 60. The RMSE for the random-effects model followed the patterns observed for its bias; RMSE increased in publication bias and τ, and decreased in μ. For $pub = 0$ or $pub = 0.5$, the random-effects model had a lower RMSE than the two other methods, because its bias was zero ($pub = 0$) or small ($pub = 0.5$) while at the same time the standard deviation of its estimates was lower than for the other methods (see supplemental materials available at https://osf.io/kyv6b/). For severe publication bias ($pub \geq 0.9$), RMSE of the other methods was often smaller because the contribution of the higher bias of the random-effects model outweighed its higher precision. Comparing Hedges1992 with *p*-uniform* shows a highly similar RMSE for $pub = 0$ and $pub = 0.5$ except for $\mu = 0$ where *p*-uniform* had a higher RMSE. For $pub = 0.9$, both methods have similar RMSE if $\mu = 0$, but *p*-uniform* had a higher RMSE for nonzero true effect size. For $pub = 1$, *p*-uniform* had a much higher RMSE than Hedges1992, and even higher than the very biased random-effects model. The differences between Hedges1992 and *p*-uniform* are not explained by bias (bias of *p*-uniform* is generally smaller), but were caused by a considerably larger standard deviation of the estimates of *p*-uniform* (see supplemental materials available at https://osf.io/kyv6b/). This was a consequence of primary studies with *p*-values close to the α -level resulting in highly negative effect size estimates of *p*-uniform*.

As the standard deviation of estimates decreased in *k*, the RMSE decreased in *k* for all methods in all conditions (see Figure 5.4). Performance of the three methods became more similar for $pub = 0$ and $pub = 0.5$. As bias mainly determined the RMSE for larger values of *k*, the RMSE of Hedges1992 (less bias) was generally smaller than of the random-effects model (most bias). The RMSE of *p*-uniform* even exceeded that of the random-effects model if $pub = 1$ because of a considerably larger standard deviation of the estimates. To conclude, if the severity of publication bias is unknown it is ill-advised to interpret estimates of the random-effects model. Additionally, although *p*-uniform* is generally less biased than the other methods if true effect size is small and $pub = 1$, its estimates are more variable than of Hedges1992, particularly for a small number of studies.

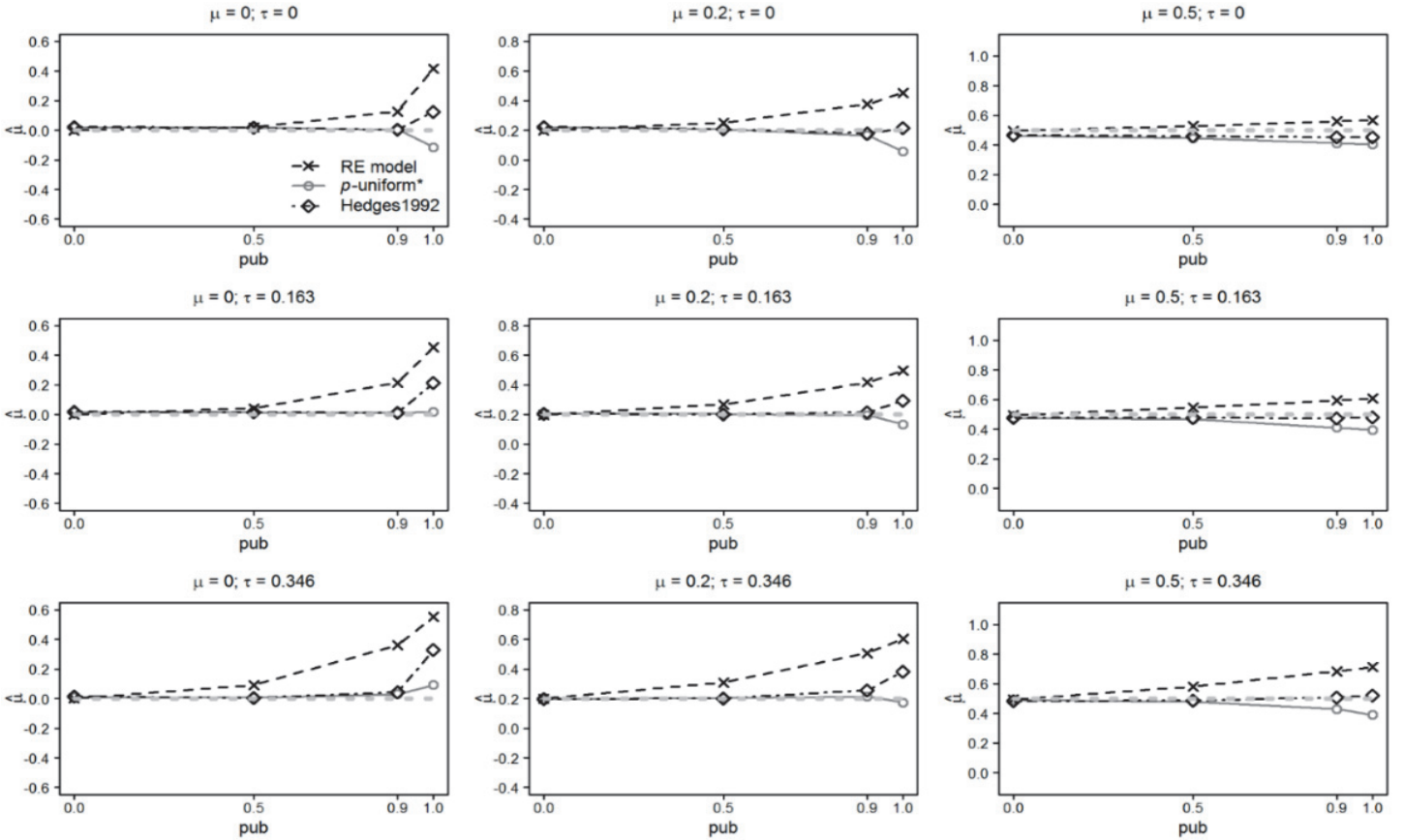

Figure 5.1. Average of the estimates of μ for the random-effects model (RE), *p*-uniform*, and Hedges1992 as a function of μ, τ, and the severity of publication bias (pub) with the number of primary studies' observed effect sizes (*k*) equal to 10.

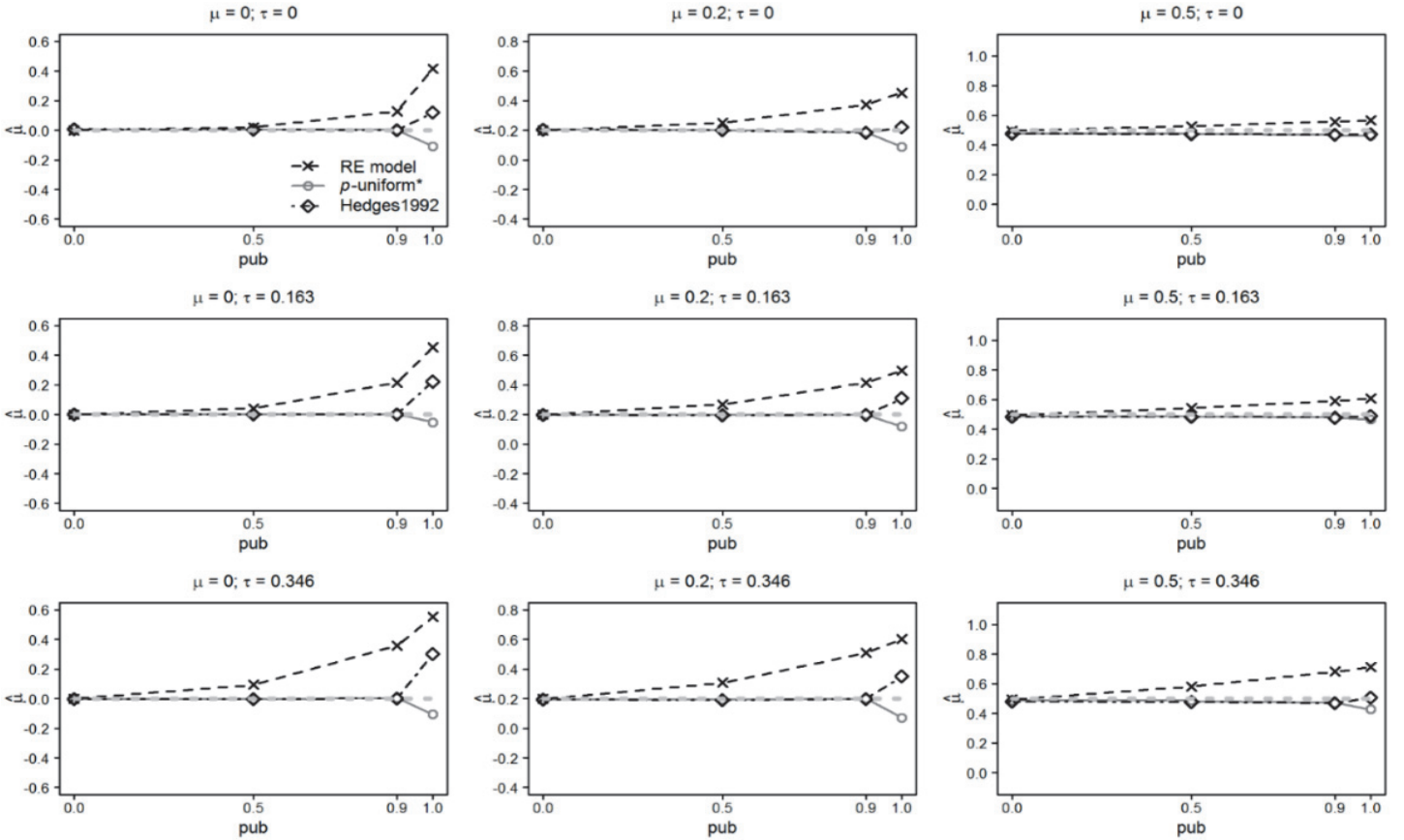

Figure 5.2. Average of the estimates of μ for the random-effects model (RE), *p*-uniform*, and Hedges1992 as a function of μ, τ, and the severity of publication bias (pub) with the number of primary studies' observed effect sizes (*k*) equal to 60.

Coverage probability of confidence interval for μ A confidence interval for μ could always be computed with the random-effects model, but not with *p*-uniform* and Hedges1992. Non-convergence was at most 12.7% for *p*-uniform*'s confidence interval (condition $\mu = 0$, $\tau = 0.346$, *k*=10, $pub = 1$), and 29.3% for Hedges1992 (condition $\mu = 0$, $\tau = 0$, *k*=10, $pub = 1$).

Table 5.2 presents the coverage probability of the 95% confidence interval for μ if *k*=10 and *k*=60. Coverage probabilities of the random-effects model were equal to 0.95 in the absence of publication bias and decreased as a function of pub with coverage probabilities approaching 0. *P*-uniform*'s coverage probabilities were close to 0.95 for $pub \leq 0.5$ and $\tau = 0$, and decreased as a function of pub and τ. However, the undercoverage of *p*-uniform* was less severe than for the random-effects model. Coverage probabilities of Hedges1992 were close to 0.95 in the absence of publication bias, but also decreased as a function of pub and τ. Undercoverage was, in general, more extreme for Hedges1992 than for *p*-uniform*.

If *k* was increased, undercoverage of the random-effects model became more severe, as detrimental effects of its bias were more pronounced for a larger number of studies; for *k*=120, the coverage probability of the random-effects model was at most 0.021 if $pub \geq 0.9$. Coverage probabilities of *p*-uniform* and Hedges1992 got closer to the nominal coverage rate if *k* increased except for $pub = 1$ where their undercoverage was severe. These results confirm that estimates of the random-effects model should not be interpreted if publication bias is present, and that performance of *p*-uniform* and Hedges1992 is not acceptable if only statistically significant results are present in the meta-analysis.

Testing null hypothesis of no effect Table 5.3 presents the Type I error rate and statistical power for testing the null hypothesis of no effect. The first four columns ($\mu = 0$) refer to the Type I error rate whereas the other columns illustrate the statistical power of the methods. For *k*=10, the Type I error rate of the random-effects model was close to 0.05 in the absence of publication bias, but it increased as a function of pub with Type I error equal to 1 for $pub = 1$. These large Type I error rates were caused by the overestimation of effect size due to publication bias. *P*-uniform* better controlled the Type I error rate than the random-effects model with the Type I error rate being close to 0.05 except for conditions with extreme publication bias and $\tau = 0.346$. However, Type I error rate of *p*-uniform* also increased as a function of pub, which followed a similar pattern as the bias for estimating μ (see Figure 5.1). Statistical power of *p*-uniform* decreased as pub was increased, and was at most 0.702 (for condition $pub = 1$, $\mu = 0.5$, $\tau = 0$). Type I error rate of Hedges1992 was larger than of *p*-uniform* if there was extreme publication

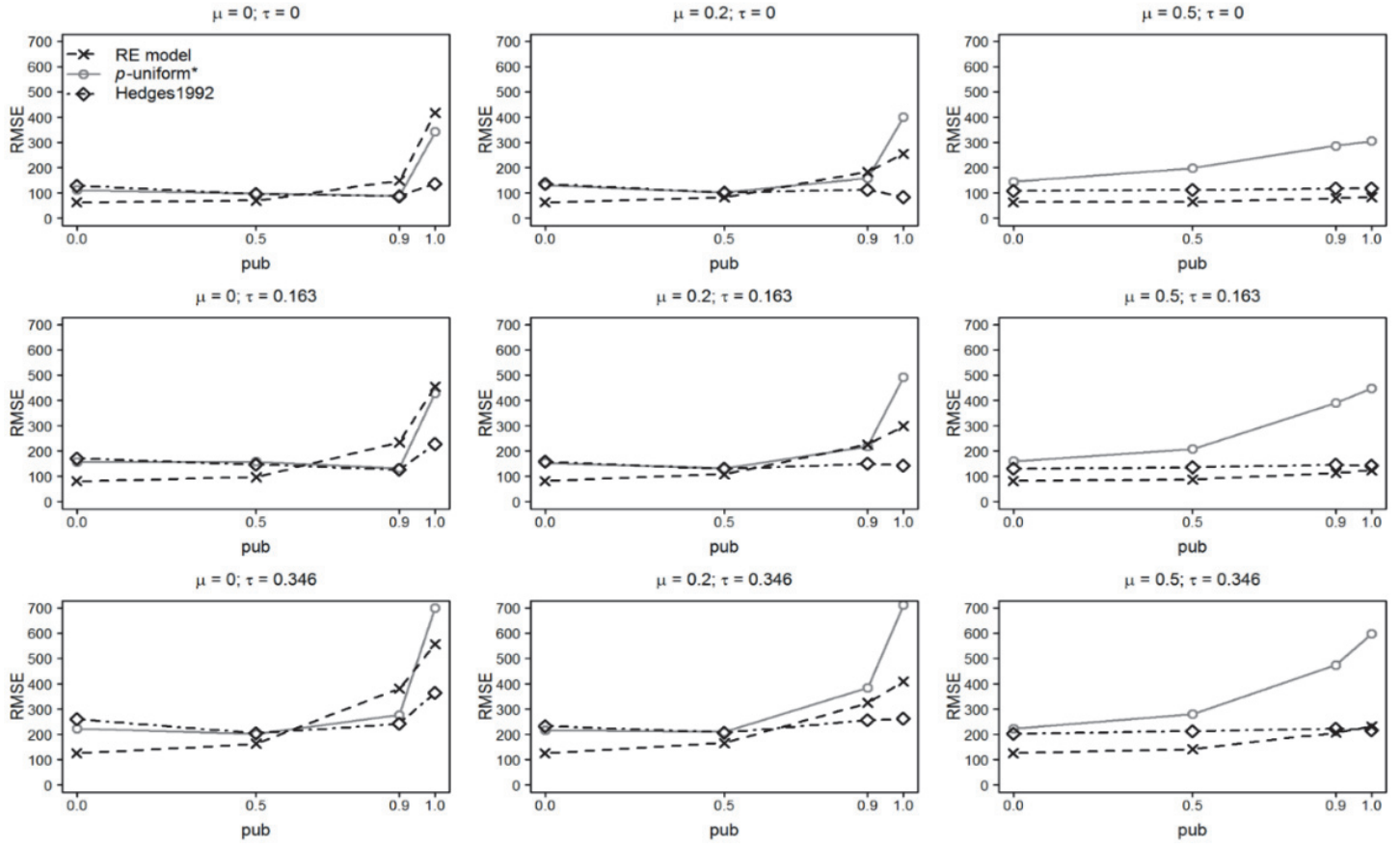

Figure 5.3. Root mean square error (RMSE) of estimating μ with the random-effects model (RE), *p*-uniform*, and Hedges1992 as a function of μ, τ, and the severity of publication bias (pub) with the number of primary studies' observed effect sizes (k) equal to 10.

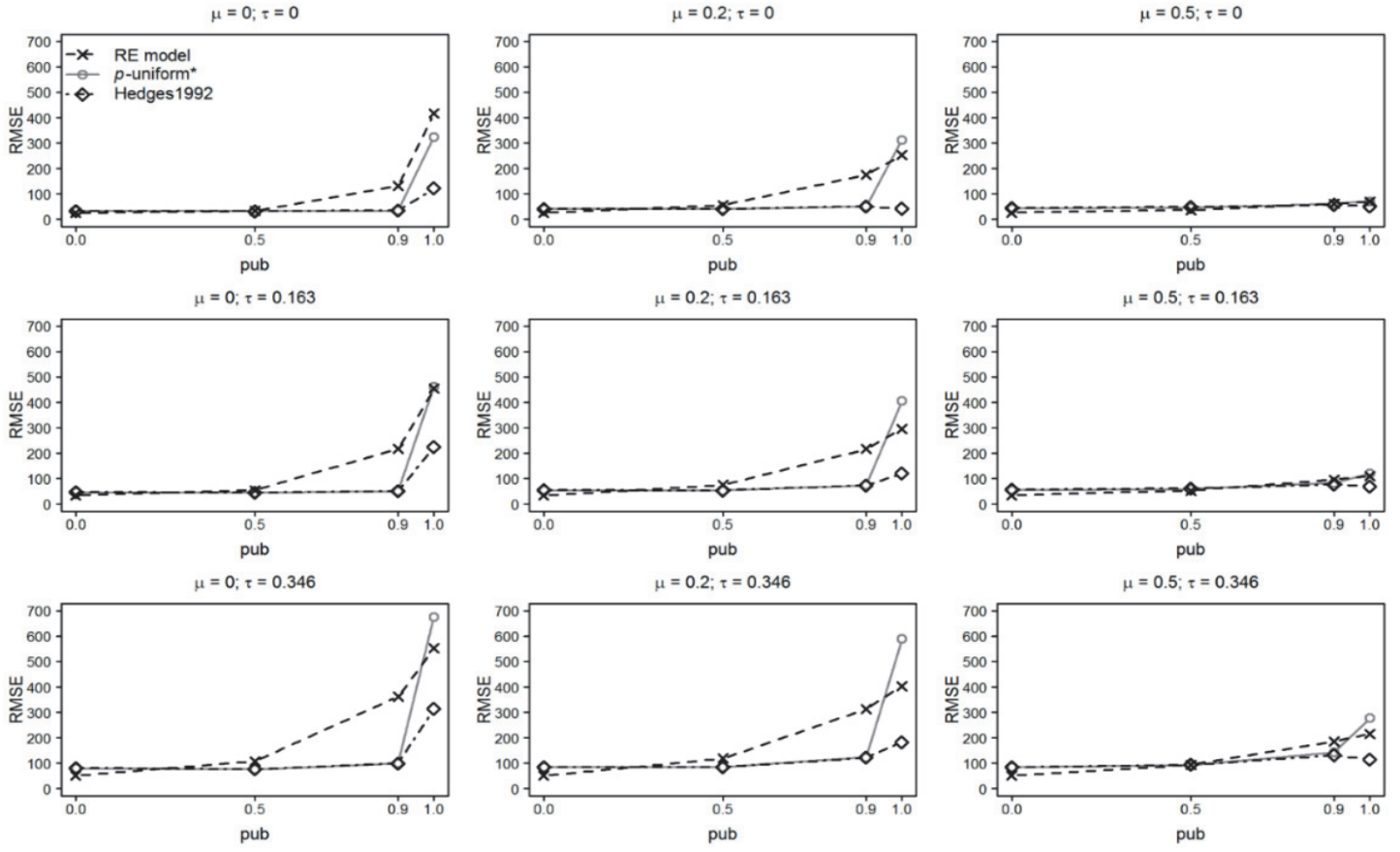

Figure 5.4. Root mean square error (RMSE) of estimating μ with the random-effects model (RE), *p*-uniform*, and Hedges1992 as a function of μ, τ, and the severity of publication bias (pub) with the number of primary studies' observed effect sizes (k) equal to 60.

Table 5.2. Coverage probability of the confidence interval for μ computed with the random-effects model (RE), *p*-uniform*, and Hedges1992 as a function of μ, τ, the severity of publication bias (pub), and the number of primary studies' observed effect sizes (*k*).

		k=10											
		$\mu = 0$				$\mu = 0.2$				$\mu = 0.5$			
	pub	0	0.5	0.9	1	0	0.5	0.9	1	0	0.5	0.9	1
τ=0	RE model	0.947	0.946	0.732	0	0.953	0.889	0.269	0	0.951	0.933	0.854	0.823
	p-uniform*	0.946	0.95	0.956	0.915	0.954	0.956	0.95	0.879	0.946	0.933	0.904	0.898
	Hedges1992	0.958	0.96	0.96	0.68	0.956	0.956	0.951	0.808	0.928	0.84	0.661	0.596
τ=0.163	RE model	0.953	0.933	0.498	0	0.953	0.878	0.253	0	0.95	0.917	0.775	0.688
	p-uniform*	0.92	0.925	0.921	0.749	0.918	0.926	0.892	0.773	0.921	0.911	0.861	0.853
	Hedges1992	0.943	0.945	0.943	0.283	0.923	0.929	0.891	0.503	0.927	0.87	0.657	0.559
τ=0.346	RE model	0.95	0.902	0.291	0	0.95	0.866	0.26	0	0.951	0.904	0.587	0.39
	p-uniform*	0.884	0.898	0.84	0.394	0.877	0.896	0.746	0.499	0.893	0.866	0.767	0.721
	Hedges1992	0.907	0.91	0.869	0.16	0.887	0.901	0.759	0.333	0.905	0.865	0.604	0.424

Table 5.2. Continued

		k=60											
		$\mu = 0$				$\mu = 0.2$				$\mu = 0.5$			
	pub	0	0.5	0.9	1	0	0.5	0.9	1	0	0.5	0.9	1
τ=0	RE model	0.954	0.898	0.037	0	0.951	0.571	0	0	0.947	0.802	0.206	0.06
	p-uniform*	0.946	0.949	0.956	0.829	0.951	0.955	0.943	0.779	0.932	0.914	0.864	0.845
	Hedges1992	0.978	0.973	0.96	0.346	0.975	0.961	0.974	0.738	0.97	0.987	0.889	0.528
τ=0.163	RE model	0.951	0.803	0.001	0	0.947	0.52	0	0	0.943	0.726	0.056	0.002
	p-uniform*	0.915	0.933	0.944	0.361	0.921	0.941	0.873	0.484	0.923	0.905	0.817	0.766
	Hedges1992	0.945	0.95	0.945	0	0.935	0.941	0.939	0.108	0.953	0.964	0.901	0.443
τ=0.346	RE model	0.951	0.618	0	0	0.952	0.472	0	0	0.952	0.608	0.004	0
	p-uniform*	0.911	0.938	0.916	0.355	0.931	0.94	0.838	0.45	0.934	0.901	0.776	0.661
	Hedges1992	0.937	0.946	0.942	0.009	0.937	0.943	0.935	0.203	0.951	0.954	0.915	0.352

bias (at most 0.867 for $pub = 1$, $\mu = 0$, $\tau = 0.346$). Statistical power of Hedges1992 was also generally larger than of *p*-uniform* and increased as a function of pub.

If *k* was increased, the Type I error rate of *p*-uniform* became closer to the α-level whereas the Type I error rate of the random-effects model for $pub \geq 0.9$ and Hedges1992 for $pub = 1$ converged to 1 if *k*=120 (see supplemental material available at https://osf.io/kyv6b/). Statistical power of all methods naturally increased in *k*. To conclude, while the random-effects model only provided an accurate Type I error rate if no publication bias was present, *p*-uniform* better controlled the Type I error rate than the random-effects model and Hedges1992 if publication bias was present. Because of the large Type I error rate of Hedges1992 in conditions with extreme publication bias (e.g., only statistically significant effect sizes), we advise not to use Hedges1992 when the goal is the test the null hypothesis. Additionally, *p*-uniform* had low statistical power if only statistically significant effect sizes were present, and is therefore also not advised to be used in these situations.

Average estimates of τ Estimates of τ could always be obtained with the random-effects model, but *p*-uniform* and the Hedges1992 did not always converge with respect to estimating τ. While non-convergence rates for *p*-uniform* were equal to those for estimating μ (i.e., at most 12.6%), it was at most 15.8% for Hedges1992 (in condition $\mu = 0$, $\tau = 0$, *k*=10, $pub = 0$). Figures 5.5 and 5.6 show the average estimates of τ for *k*=10 and 60, respectively. For *k*=10 and $pub = 0$, the random-effects model overestimated τ if $\tau = 0$ (maximum bias 0.052) and underestimated it if $\tau > 0$ (maximum bias -0.028). If $pub = 1$ and $\tau > 0$, the random-effects model underestimated τ for all conditions, because meta-analyses in this condition only consisted of statistically significant observed y_i resulting in hardly any variability in y_i and difficulties for estimating τ. If $\tau = 0$, there was a small positive bias in *p*-uniform* (maximum bias 0.067) and Hedges1992 (maximum bias 0.05) for all levels of pub, and this bias was the largest for $pub = 1$ in combination with $\mu = 0.5$ (bias = 0.067 for *p*-uniform* and 0.05 for Hedges1992). *P*-uniform* and Hedges1992 underestimated τ if $\tau > 0$ for all levels of pub with *p*-uniform* being less negatively biased than Hedges1992.

Increasing *k* resulted in less bias of *p*-uniform* and Hedges1992, but did not affect the bias of the random-effects model. Bias of *p*-uniform* and Hedges1992 was negligible for *k*=30 (maximum bias 0.058 for *p*-uniform* and 0.047 for Hedges1992), for *k*=60 (maximum bias 0.053 for *p*-uniform* and 0.044 for Hedges1992) and for *k*=120 (maximum bias 0.042 for *p*-uniform* and 0.039 for Hedges1992). Hence, these results imply that either *p*-uniform* or the Hedges1992 should be used for estimating τ instead of the random-effects model in the presence of publication bias, although a substantial number of studies was needed for accurate estimation.

Table 5.3. Type I error rate and statistical power for testing the null hypothesis of no effect with the random-effects model (RE), *p*-uniform*, and Hedges1992 as a function of μ, τ, the severity of publication bias (*pub*), and the number of primary studies' observed effect sizes (*k*).

		k=10											
		$\mu = 0$				$\mu = 0.2$				$\mu = 0.5$			
	pub	0	0.5	0.9	1	0	0.5	0.9	1	0	0.5	0.9	1
τ=0	RE model	0.05	0.08	0.402	1	0.896	0.958	0.999	1	1	1	1	1
	p-uniform*	0.042	0.039	0.039	0.155	0.609	0.498	0.163	0.094	0.702	0.495	0.199	0.1
	Hedges1992	0.026	0.026	0.035	0.401	0.518	0.553	0.361	0.793	0.939	0.927	0.965	0.99
τ=0.163	RE model	0.048	0.111	0.638	1	0.723	0.89	0.999	1	1	1	1	1
	p-uniform*	0.049	0.054	0.057	0.127	0.392	0.331	0.138	0.096	0.671	0.479	0.226	0.141
	Hedges1992	0.016	0.022	0.063	0.776	0.385	0.437	0.415	0.924	0.914	0.888	0.927	0.969
τ=0.346	RE model	0.05	0.161	0.808	1	0.426	0.711	0.988	1	0.975	0.996	1	1
	p-uniform*	0.059	0.069	0.065	0.14	0.186	0.162	0.123	0.142	0.5	0.364	0.203	0.158
	Hedges1992	0.023	0.043	0.156	0.867	0.235	0.281	0.452	0.864	0.75	0.728	0.819	0.909

Table 5.3. Continued

		k=60											
		$\mu = 0$				$\mu = 0.2$				$\mu = 0.5$			
	pub	0	0.5	0.9	1	0	0.5	0.9	1	0	0.5	0.9	1
τ=0	RE model	0.048	0.167	0.984	1	1	1	1	1	1	1	1	1
	p-uniform*	0.045	0.043	0.042	0.091	1	0.999	0.901	0.018	1	0.998	0.908	0.773
	Hedges1992	0.025	0.029	0.037	0.756	0.998	0.999	0.938	1	1	1	0.993	1
τ=0.163	RE model	0.054	0.297	1	1	1	1	1	1	1	1	1	1
	p-uniform*	0.056	0.048	0.053	0.059	0.986	0.966	0.664	0.077	1	0.997	0.827	0.566
	Hedges1992	0.022	0.029	0.051	1	0.991	0.983	0.82	1	1	1	0.983	1
τ=0.346	RE model	0.05	0.505	1	1	0.986	1	1	1	1	1	1	1
	p-uniform*	0.054	0.05	0.053	0.079	0.682	0.613	0.299	0.084	0.998	0.972	0.619	0.305
	Hedges1992	0.026	0.036	0.068	0.994	0.76	0.729	0.511	0.995	1	0.993	0.884	0.998

RMSE for estimating τ Figures 5.7 and 5.8 present the RMSE for estimating τ of the different methods for *k*=10 and 60, respectively. For *k*=10, the RMSE of the random-effects model increased in pub if $\tau > 0$. If $\tau = 0$, the RMSE was very small for $pub = 1$, but this was because the random-effects model severely underestimated τ in this condition (see Figures 5.5 and 5.6). *P*-uniform* and Hedges1992 had similar RMSEs, except that *p*-uniform*'s RMSE exceeded that of Hedges1992 if $pub = 1$; this was generally not caused by higher bias (see Figure 5.5 and 6), but due to higher variability of *p*-uniform*'s estimates for τ (see supplemental materials available at https://osf.io/kyv6b/).

While the RMSE did not substantially decrease for the random-effects model when increasing *k* to 60, it did decrease for *p*-uniform* and Hedges1992. The RMSEs of *p*-uniform* and Hedges1992 were quite similar, although lower for Hedges1992 if $pub = 1$. The RMSEs of *p*-uniform* and Hedges' selection model approach were both considerably lower than that of the random-effects model if $pub \geq 0.9$ and $\tau > 0$. These results implied that the larger bias in the random-effects model compared to *p*-uniform* and Hedges1992 was compensated with the smaller standard deviation of the estimates resulting in a lower RMSE for the random-effects model if $pub \leq 0.5$ and $\tau > 0$. **Coverage probability of confidence interval for τ** A confidence interval for τ could always be computed with the random-effects model but not always with *p*-uniform* or the selection model approach. Non-convergence of *p*-uniform*'s confidence interval was the same as for estimating μ and τ, while non-convergence of Hedges1992 was at most 15.8% ($\mu = 0$, $\tau = 0$, *k*=10, $pub = 0$).

Table 5.4 presents the coverage probabilities of the three methods. Coverage probabilities of the random-effects model were close to 0.95 for $pub = 0$ but decreased as a function of pub. Undercoverage of the random-effects model was most severe (0.072) for $pub = 1$ in combination with $\mu = 0$ and $\tau = 0$. Coverage probabilities of *p*-uniform* were close to 0.95 if $pub = 0$ and $\tau < 0.346$, but generally decreased as pub and τ were increased. Undercoverage was most severe for $pub = 1$ (minimum coverage (0.346). There was undercoverage for Hedges1992 for $pub = 0$, (minimum coverage (0.346). There was undercoverage for Hedges1992 for $pub = 0$, but coverage of Hedges1992 was, in contrast to *p*-uniform*, too high even up to 1 for $pub = 1$.

Coverage probabilities of the random-effects model decreased if *k* was increased. For *k*=60, the coverage probability of the random-effects model was even equal to zero for $pub = 1$ in combination with $\mu = 0$ and $\mu = 0.2$. Coverage probabilities of *p*-uniform* and Hedges1992 became closer to the nominal coverage rate except for $pub = 1$. Hence, researchers are advised against interpreting

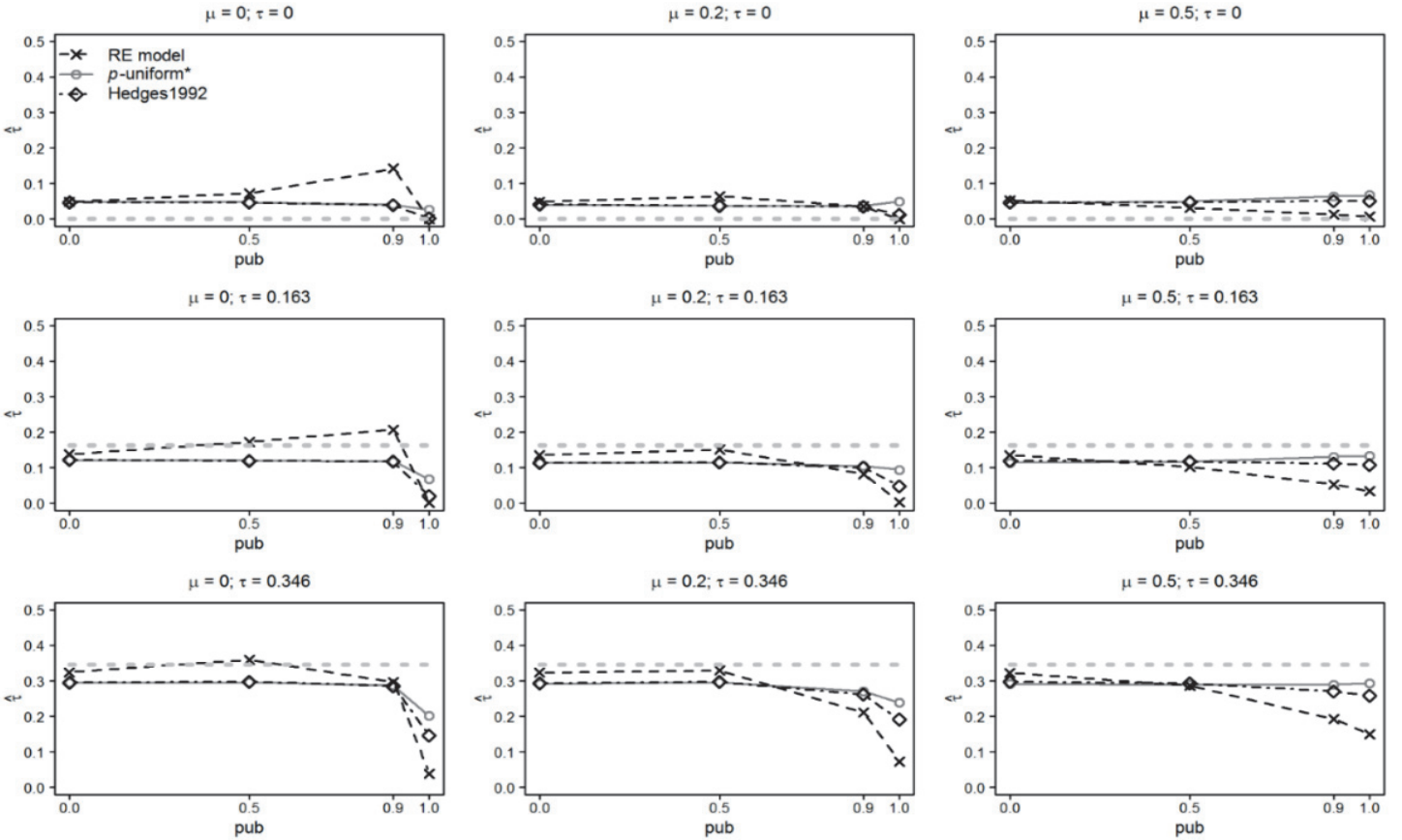

Figure 5.5. Average of the estimates of τ for the random-effects model (RE), *p*-uniform*, and Hedges1992 as a function of μ, τ, and the severity of publication bias (pub) with the number of primary studies' observed effect sizes (*k*) equal to 10.

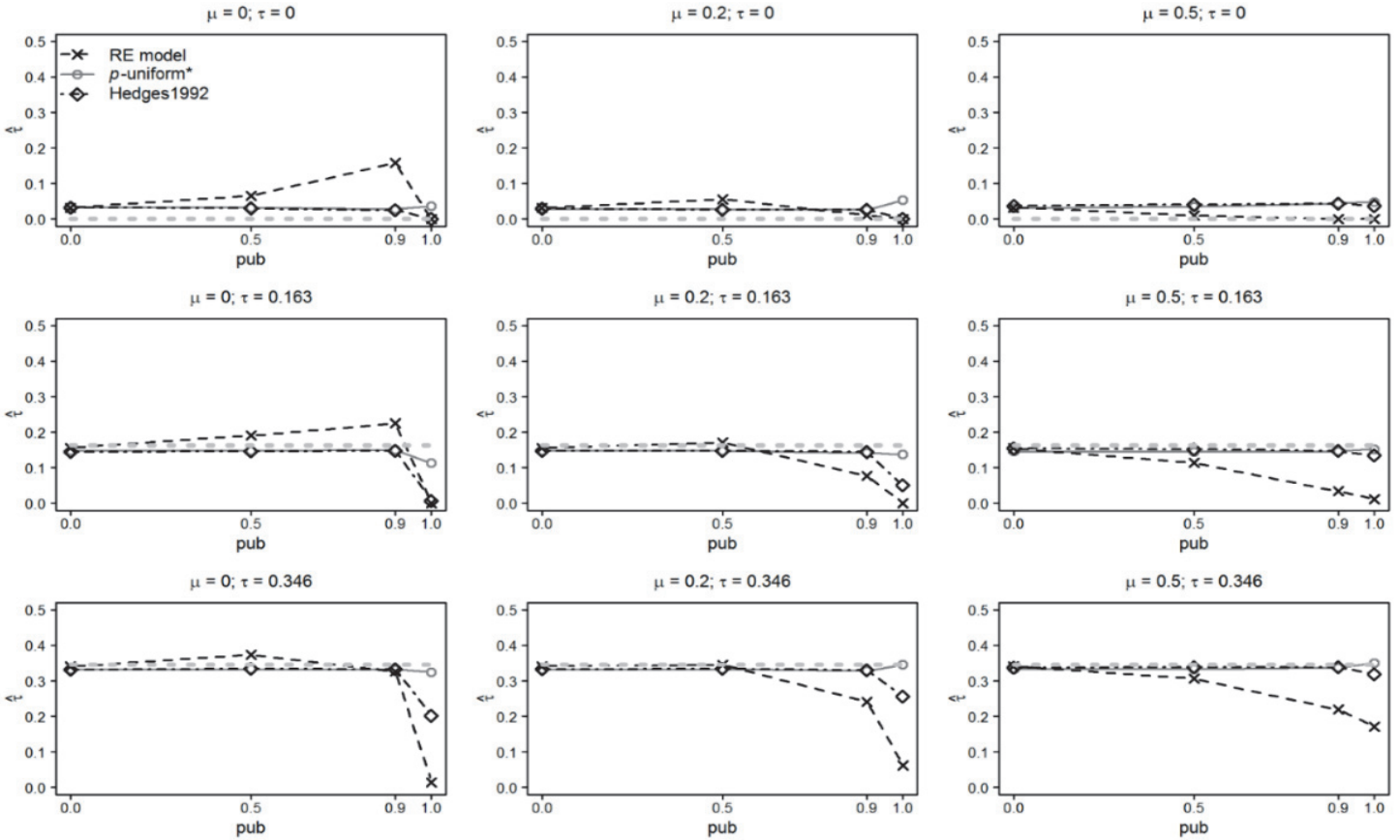

Figure 5.6. Average of the estimates of τ for the random-effects model (RE), *p*-uniform*, and Hedges1992 as a function of μ, τ, and the severity of publication bias (pub) with the number of primary studies' observed effect sizes (k) equal to 60.

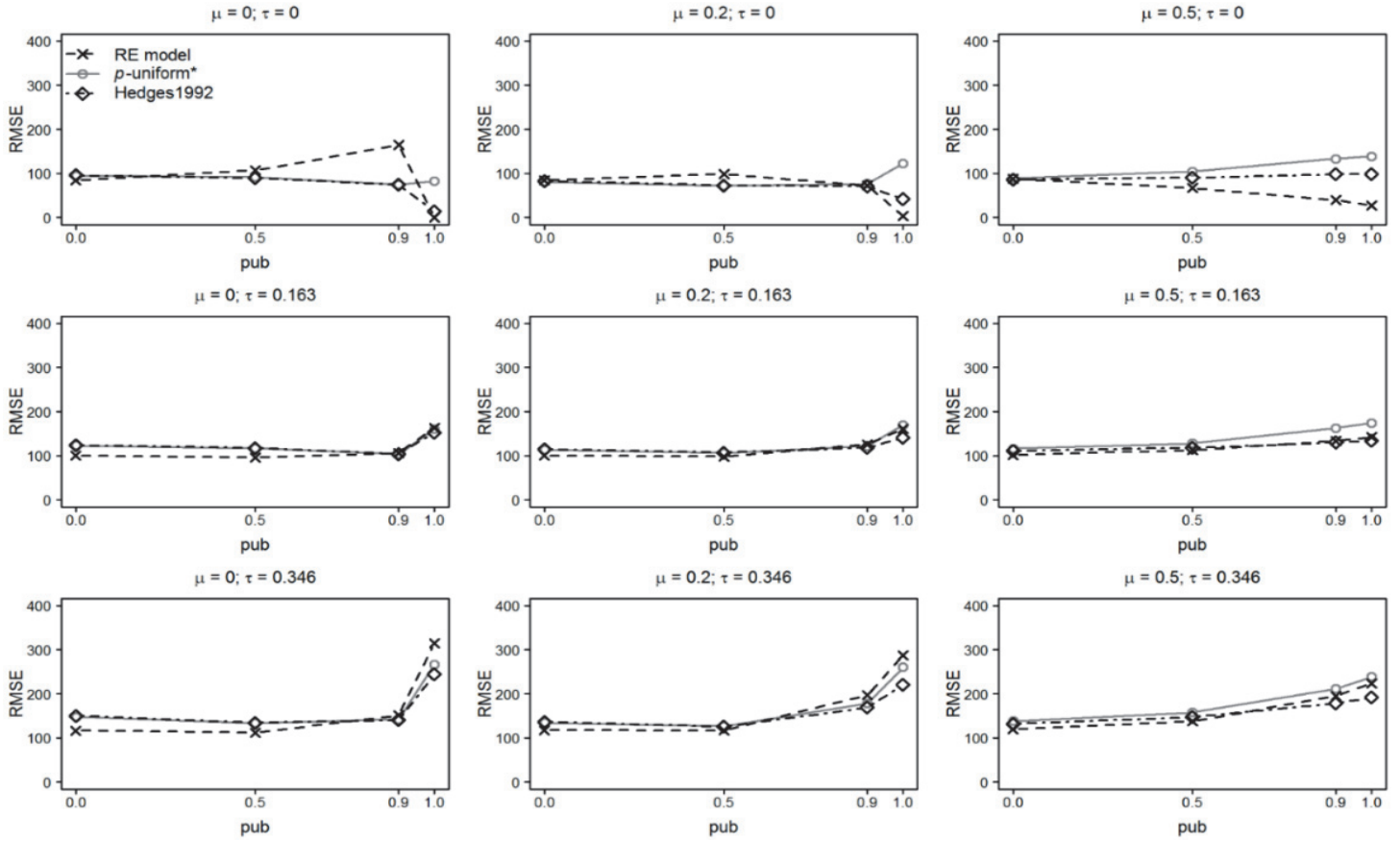

Figure 5.7. Root mean square error (RMSE) of estimating τ with the random-effects model (RE), *p*-uniform*, and Hedges1992 as a function of μ, τ, and the severity of publication bias (pub) with the number of primary studies' observed effect sizes (k) equal to 10.

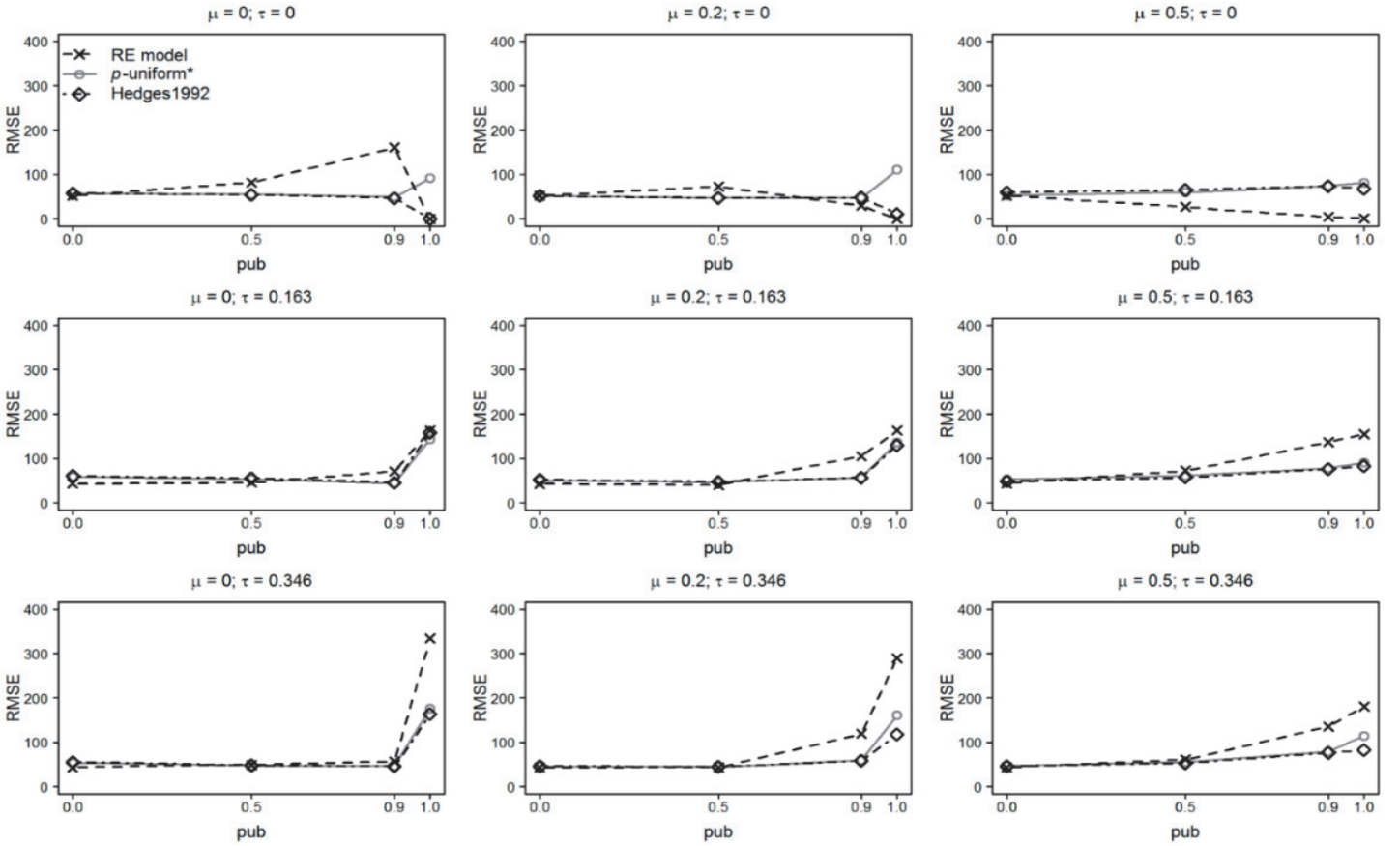

Figure 5.8. Root mean square error (RMSE) of estimating τ with the random-effects model (RE), *p*-uniform*, and Hedges1992 as a function of μ, τ, and the severity of publication bias (pub) with the number of primary studies' observed effect sizes (k) equal to 60.

confidence intervals of *p*-uniform* and Hedges1992 if there they suspect extreme publication bias.

Conclusion None of the methods outperformed the other methods for all the studied conditions and outcome variables in the Monte-Carlo simulation study. However, some general recommendations can be made. Although the random-effects model had the best statistical properties in the absence of publication bias, we do usually not know the severity of publication bias. Hence, we recommend to always accompany traditional fixed-effect or random-effects meta-analysis with either *p*-uniform* or the selection model approach.

P-uniform* and Hedges1992 outperformed the random-effects model if publication bias was present. However, statistical properties of *p*-uniform* and Hedges1992 were not good in case of extreme publication bias with only statistically significant primary studies' effect sizes in a meta-analysis. As increasing the number of studies to even 120 did not always improve the statistical properties, we recommend not to put much trust in the estimates of any of the methods when there is extreme publication bias with a meta-analysis only consisting of statistically significant studies.

The selection model approach and *p*-uniform* were highly comparable which makes it impossible to recommend one method over the other. However, some recommendations can still be made based on the results of our Monte-Carlo simulations. First, we recommend to use *p*-uniform* if a researcher's main emphasis is on estimating the average effect size and between-study variance, as *p*-uniform* had no systematic bias in estimating the average effect size and hardly suffers from convergence problems. However, estimates of *p*-uniform* can be highly negative, which resulted in a larger RMSE than that of Hedges1992 and sometimes even larger than of the random-effects model. These highly negative estimates of *p*-uniform* were caused by primary studies' *p*-values close to the α -level. Hence, we recommend to set *p*-uniform*'s estimate of the average effect size to zero if this occurs, which is in line with our recommendation for *p*-uniform and *p*-curve (van Aert, Wicherts, et al., 2016). This adjustment is defensible, because it is unlikely that the average effect size estimate is (strongly) negative if statistically significant positive primary studies' effect sizes are observed. Second, researchers are recommended not to interpret confidence intervals for μ and τ of *p*-uniform* and the selection model approach if there is extreme publication bias with a meta-analysis consisting of only statistically significant studies, because our results indicated that confidence intervals substantially deviated from the nominal coverage rate in these conditions. Although coverage probabilities of *p*-uniform* and the selection model approach were generally closer to nominal coverage than the random-effects model for $pub = 0.9$, coverage of *p*-uniform* and the selection model approach was not close to the nominal coverage rate, especially if the between-study variance in true effect sizes was large.

Table 5.4. Coverage probability of the confidence interval for τ computed with the random-effects model (RE), *p*-uniform*, and Hedges1992 as a function of μ, τ, the severity of publication bias (pub), and the number of primary studies' observed effect sizes (*k*).

		k=10											
		$\mu = 0$				$\mu = 0.2$				$\mu = 0.5$			
	pub	0	0.5	0.9	1	0	0.5	0.9	1	0	0.5	0.9	1
τ=0	RE model	0.95	0.933	0.793	0.072	0.951	0.95	0.845	0.296	0.952	0.93	0.883	0.852
	p-uniform*	0.962	0.968	0.987	0.94	0.975	0.986	0.975	0.893	0.97	0.947	0.903	0.895
	Hedges1992	0.716	0.738	0.795	0.707	0.754	0.809	0.799	0.952	0.833	0.843	0.945	1
τ=0.163	RE model	0.949	0.939	0.883	0.12	0.953	0.952	0.805	0.336	0.951	0.931	0.849	0.809
	p-uniform*	0.959	0.971	0.978	0.926	0.977	0.983	0.963	0.894	0.976	0.954	0.899	0.88
	Hedges1992	0.918	0.927	0.951	0.953	0.93	0.951	0.922	0.995	0.949	0.94	0.967	0.992
τ=0.346	RE model	0.953	0.947	0.877	0.232	0.95	0.949	0.776	0.413	0.948	0.927	0.809	0.726
	p-uniform*	0.896	0.91	0.863	0.479	0.903	0.913	0.797	0.577	0.903	0.872	0.78	0.732
	Hedges1992	0.78	0.788	0.766	0.346	0.788	0.802	0.702	0.455	0.798	0.774	0.656	0.574

Table 5.4. Continued

		k=60											
		$\mu = 0$				$\mu = 0.2$				$\mu = 0.5$			
	pub	0	0.5	0.9	1	0	0.5	0.9	1	0	0.5	0.9	1
τ=0	RE model	0.951	0.886	0.137	0	0.948	0.927	0.743	0	0.95	0.85	0.386	0.186
	p-uniform*	0.972	0.978	0.986	0.866	0.977	0.987	0.978	0.813	0.977	0.957	0.907	0.881
	Hedges1992	0.986	0.988	0.996	0.951	0.993	0.997	0.996	1	0.996	0.989	0.956	0.952
τ=0.163	RE model	0.949	0.872	0.554	0	0.95	0.938	0.587	0	0.949	0.834	0.286	0.09
	p-uniform*	0.91	0.927	0.938	0.308	0.924	0.936	0.866	0.425	0.922	0.892	0.796	0.729
	Hedges1992	0.889	0.897	0.902	0.053	0.9	0.905	0.894	0.312	0.933	0.926	0.91	0.668
τ=0.346	RE model	0.951	0.905	0.876	0	0.955	0.947	0.396	0	0.955	0.849	0.227	0.027
	p-uniform*	0.913	0.937	0.915	0.364	0.936	0.938	0.837	0.454	0.934	0.897	0.765	0.657
	Hedges1992	0.891	0.904	0.901	0.178	0.907	0.902	0.896	0.421	0.919	0.916	0.888	0.642

5.5 Discussion

Publication bias distorts the results of meta-analyses yielding overestimated effect sizes (e.g., Kraemer et al., 1998; Lane & Dunlap, 1978) and false positives (van Assen et al., 2015). Multiple methods were developed to correct for publication bias in a meta-analysis, and selection model approaches are seen as the state-of-the-art methods (Carter et al., 2017; McShane et al., 2016). The *p*-uniform method (van Aert, Wicherts, et al., 2016; van Assen et al., 2015) can also be seen as a selection model approach, and we extended and improved this method in this chapter. The new method, *p*-uniform*, does not only use statistically significant primary studies' effect sizes for estimation as is done in *p*-uniform, but also uses the nonsignificant effect sizes. Including the statistically nonsignificant primary studies' effect sizes results in three major improvements of *p*-uniform* over *p*-uniform: (i) it makes *p*-uniform* a more efficient estimator than *p*-uniform, (ii) overestimation of effect size by *p*-uniform in case of between-study variance in true effect is eliminated, and (iii) it enables estimation and testing for the presence of the between-study variance in true effect size.

The aim of this chapter was to introduce *p*-uniform* and compare the statistical properties of the method with those of the selection model approach of Hedges (1992) and the random-effects model that is commonly used but does not correct for publication bias. We assessed the statistical properties of the different methods by means of an analytical study where a meta-analysis only consisted of one statistically significant and one nonsignificant primary study's effect size. Moreover, we studied their relative performance under different levels of publication bias using a Monte-Carlo simulation study where we selected conditions that are representative for meta-analyses in practice. Statistical properties of the random-effects model were better than of *p*-uniform* and the selection model approach of Hedges (1992) if publication bias did not affect the probability of publishing a primary study. If publication bias was present, the random-effects model performed worse than the two other methods, confirming previous research showing that it overestimates effect size (e.g., Kraemer et al., 1998; Lane & Dunlap, 1978) and yields unpredictable bias in estimating between-study variance when publication bias operates (Augusteijn et al., 2017; Jackson, 2006, 2007). Statistical properties of *p*-uniform* and the selection model approach of Hedges (1992) were generally comparable. *P*-uniform* showed slightly larger RMSE than the selection model approach, but *p*-uniform* suffered less often from convergence problems when estimating the effect size and between-study variance and computing confidence intervals. Moreover, coverage probabilities of *p*-uniform* were closer to the nominal coverage rate than the selection model approach of Hedges (1992). Statistical properties of both methods were, however, not good in case of extreme publication bias with only statistically significant primary studies' effect sizes in a meta-analysis. To conclude, bias in estimates of effect size and

between-study variance was generally small except for conditions with meta-analyses only consisting of statistically significant effect sizes due to publication bias. Bias was small even if only 10 primary studies were included in a meta-analysis, but coverage probabilities of confidence intervals deviated in this condition from the nominal coverage rate especially if the between-study variance in true effect sizes was large.

The comparable statistical properties of *p*-uniform* and the selection model approach were caused by the similarities between the two methods. Both methods use maximum likelihood estimation, the random effects model as effect size model, and a selection model with one threshold at the α -level. However, there are also differences between the two methods explaining why the statistical properties were not exactly the same. First, weights in the selection model have to be estimated or assumed to be known in the selection model approach by Hedges (1992) which is in contrast with *p*-uniform*. *P*-uniform* only assumes that these probabilities are the same for the statistically significant and nonsignificant primary studies' effect sizes and there is no need for estimating these probabilities. Another difference is that the selection model approach relies on asymptotic normality distributions for creating Wald-type confidence intervals around estimates of effect size and between-study variance. This is in contrast with *p*-uniform*'s confidence intervals, because these are computed by inverting the likelihood-ratio tests which may explain why the coverage probabilities of *p*-uniform's confidence intervals were closer to the nominal coverage rate than of the selection model approach.

We provide recommendations for meta-analysts in practice based on the results of our analytical study and Monte-Carlo simulation study. These recommendations are built upon guidelines for conducting meta-analyses as the Meta-Analytic Reporting Standards (MARS; American Psychological Association, 2010), and the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA; Moher, Liberati, Tetzlaff, Altman, & The Prisma Group, 2009). First, we recommend to apply both the fixed-effect and random-effects meta-analysis models since differences in estimates of the average effect size may already signal the presence of publication bias (Greenhouse & Iyengar, 2009). The fixed-effect model weighs the primary studies by the inverse of their sampling variance whereas the random-effects model weighs primary studies by the inverse of the sampling variance and the between-study variance. Consequently, the primary studies with small sample sizes that are more prone to publication bias get more weight in the random-effects than fixed-effect model.

Second, we recommend to not solely rely on the fixed-effect and random-effects model if publication bias may have affected the meta-analysis, but to supplement these models by either *p*-uniform* or the selection model approach of Hedges (1992) or by both methods. As publication bias is most likely affecting many meta-analyses, supplementing the fixed-effect or random-effects model with methods that correct for publication bias is of utmost importance especially if there are

indications for publication bias in a meta-analysis such as many underpowered but statistically significant primary studies in a meta-analysis. Examining publication bias in a meta-analysis is in agreement with the MARS and PRISMA that both recommend to assess the risk and consequences of bias in any meta-analysis. We advise researchers to use so-called triangulation where researchers do not rely on one particular publication bias method, but use multiple publication bias methods (Coburn & Vevea, 2015; Kepes, Banks, McDaniel, et al., 2012). This is necessary because research, including ours, has shown that there is no single publication bias methods that outperforms all the other available publication bias methods.

Third, we cannot generally recommend using *p*-uniform* over the selection model approach of Hedges (1992), or the other way around. *P*-uniform* has the advantage over the selection model approach that the method does not require that the weights in the selection model are estimated or assumed to be known. A disadvantage of *p*-uniform*, however, is that estimates of the average effect size have a larger variance and may be highly negative if many primary studies' *p*-values are close to the α -level. Hence, we offer the same recommendation as for *p*-uniform and *p*-curve (van Aert et al., 2016) to set the effect size estimate to zero if it is negative since it is unlikely that the average effect size estimate is (strongly) negative if statistically significant positive primary studies' effect sizes are observed. An additional advantage of the selection model approach of Hedges (1992) is its flexibility, because covariates can be included (Vevea & Hedges, 1995) and its selection model can be extended to more than two intervals. *P*-uniform* currently does not offer such extensions.

Fourth, we do not recommend using either *p*-uniform* or the selection model approach of Hedges (1992) if the meta-analysis only contains statistically significant effect sizes due to extreme publication bias, as our simulation study showed bad performance for both methods in that condition. If there is no between-study variance in true effect sizes, *p*-uniform can then better be applied since this method provides estimates close to the true effect size and exact confidence intervals in these situations (van Aert, Wicherts, et al., 2016; van Assen et al., 2015). Moreover, we also suspect good performance by *p*-uniform* and the selection model approach in this condition when all statistically significant effect sizes are accompanied by very small *p*-values (say < .001), suggesting that these significant effects are not caused by publication bias but by high power of the original studies. This condition was not examined in the simulation study.

P-uniform* can be easily applied by meta-analysts using the function "puni_star" in the R package "puniform" that is available via https://github.com/RobbievanAert/puniform. Users can currently analyze primary studies based on a two-samples or one-sample *t*-test and correlation coefficient or can supply the function directly with standardized effect sizes. Meta-analysts who are not familiar with R can also apply *p*-uniform* to their data via an easy-to-use web

application (https://rvanaert.shinyapps.io/p-uniformstar/). This web application can be applied if the primary study's effect size measure is a one-sample mean, two-independent means or correlation coefficient. The selection model approach of Hedges (1992) can be applied using the "weightr" package (Coburn & Vevea, 2016) or the web application (https://vevealab.shinyapps.io/WeightFunctionModel/).

A limitation of *p*-uniform* and the selection model approach of Hedges (1992) is that the probability of publishing a statistically nonsignificant primary study's effect size is assumed to be the same for all nonsignificant effect sizes in a meta-analysis. This assumption may be violated in practice causing biased estimates of the methods (see Simonsohn et al., 2017, December 20 for a discussion about this assumption). To counteract a violation of this assumption, the flexibility of the selection model approach of Hedges (1992) can be used by creating more than two intervals of the method's selection model. However, information needs then to be available to select appropriate thresholds for these intervals. A threshold at the α -level seems logical since it determines statistical significance, but selecting another appropriate threshold is probably not as straightforward in practice. Nevertheless, future research may study to what extent statistical properties of *p*-uniform* and the selection model approach of Hedges (1992) with one or multiple thresholds for the selection model deteriorate if this assumption is violated. Future research may also study the effect of publication bias on the meta-analytic results if covariates are included in a meta-analysis model.

Another limitation of the methods is that their results will be distorted if *p*-hacking (a.k.a. questionable research practices or researcher degrees of freedom, see Simmons et al., 2011; Wicherts et al., 2016) are used in the primary studies that are included in a meta-analysis, but this limitation applies to any meta-analysis method. A limitation that only applies to the selection model of Hedges (1992) is that it is not supposed to converge according to the underlying statistical theory if there are no primary studies' *p*-values observed in each interval of the selection model. Weights of these intervals cannot be computed causing the non-convergence of the method. This non-convergence is circumvented in the R package "weightr" (Coburn & Vevea, 2016) used in our study by fixing weights to 0.01 if these cannot be estimated. Although previous research has shown that the weights are often poorly estimated and that the selection model approach is quite robust to misestimated weights (Hedges & Vevea, 1996), future research may scrutinize the effects of fixing weights and whether these weights are better fixed to another value than 0.01. The value .01 seems unrealistically small, implying that 99% of the statistically nonsignificant effect sizes end up in the file drawer. Additionally, this weight may be estimated by other evidence in the field of the meta-analysis.

As our study did not address the performance of tests of publication bias, we also recommend further research on developing and examining publication bias tests. Multiple tests for publication bias are currently used; rank-correlation test (Begg &

Mazumdar, 1994), Egger's test (Egger et al., 1997), *p*-uniform's publication bias test (van Assen et al., 2015), and the test of excess significance (Ioannidis & Trikalinos, 2007b). A publication bias test was also proposed in Hedges (1992) based on the selection model approach, and a publication bias test in the framework of *p*-uniform* can also be developed. Future research may study the statistical properties of these publication bias tests. Future research may also consider to implement *p*-uniform* in a Bayesian framework. This Bayesian version of *p*-uniform* can then together with *p*-uniform* and the selection model approach of Hedges (1992) be compared with other selection model approaches that were developed in a Bayesian framework. Two recently developed Bayesian methods that deserve attention in the future research are the BALM method (Du et al., 2017) and the Bayesian model averaging method proposed by Guan and Vandekerckhove (2015).

To conclude, scientific progress can best be achieved by using meta-analysis (Cumming, 2008), but this progress is hampered by publication bias causing false positive (Bakker et al., 2012; van Assen et al., 2015) and overestimated (e.g., Kraemer et al., 1998; Lane & Dunlap, 1978) effect sizes. Hence, there is a need for methods that can accurately estimate the effect size and between-study variance in a meta-analysis in the presence of publication bias. In line with others (e.g., Borenstein et al., 2009; Rothstein et al., 2005a), we recommend to routinely assess the impact and presence of publication bias in each meta-analysis, and not apply a single publication bias method but to use triangulation (Coburn & Vevea, 2015; Kepes, Banks, McDaniel, et al., 2012). The *p*-uniform* method is an extension and substantial improvement over *p*-uniform and showed promising results in an analytical study and Monte-Carlo simulations. *P*-uniform* can easily be applied by meta-analysts via R or the web application (https://rvanaert.shinyapps.io/p-uniformstar/).

CHAPTER 6

Examining reproducibility in psychology: A hybrid method for combining a statistically significant original study and a replication

Abstract

The unrealistic high rate of positive results within psychology increased the attention for replication research. Researchers who conduct a replication and want to statistically combine the results of their replication with a statistically significant original study encounter problems when using traditional meta-analysis techniques. The original study's effect size is most probably overestimated because of it being statistically significant and this bias is not taken into consideration in traditional meta-analysis. We developed a hybrid method that does take statistical significance of the original study into account and enables (a) accurate effect size estimation, (b) estimation of a confidence interval, and (c) testing of the null hypothesis of no effect. We analytically approximate the performance of the hybrid method and describe its good statistical properties. Applying the hybrid method to the data of the Reproducibility Project Psychology (Open Science Collaboration, 2015) demonstrated that the conclusions based on the hybrid method are often in line with those of the replication, suggesting that many published psychological studies have smaller effect sizes than reported in the original study and that some effects may be even absent. We offer hands-on guidelines for how to statistically combine an original study and replication, and developed a web-based application (https://rvanaert.shinyapps.io/hybrid) for applying the hybrid method.

This chapter is published as van Aert, R. C. M., & van Assen, M. A. L. M. (2017). Examining reproducibility in psychology: A hybrid method for combining a statistically significant original study and a replication. Behavior Research Methods. doi:10.3758/s13428-017-0967-6

There is increased attention for replication research in psychology mainly due to the unrealistic high rate of positive results within the published psychological literature. Approximately 95% of the published psychological research contain statistically significant results in the predicted direction (Fanelli, 2012; Sterling et al., 1995). This is not in line with the average amount of statistical power which has been estimated at .35 (Bakker et al., 2012) and .47 (Cohen, 1990) in psychological research and .21 in neuroscience (Button et al., 2013), and indicates that statistically nonsignificant results do often not get published. This suppression of statistically nonsignificant results from being published is called publication bias (Rothstein et al., 2005a). Publication bias causes the population effect size to be overestimated (e.g., Lane & Dunlap, 1978; van Assen et al., 2015), and raises the question whether a particular effect reported in the literature actually exists. Other research fields also show an excess of positive results (e.g., Ioannidis, 2011; Kavvoura et al., 2008; Renkewitz et al., 2011; Tsilidis, Papatheodorou, Evangelou, & Ioannidis, 2012), so publication bias and overestimation of effect size by published research is not only an issue within psychology.

Replication research can help to identify whether a particular effect in the literature is probably a false positive (Murayama, Pekrun, & Fiedler, 2014), and to increase accuracy and precision of effect size estimation. The Open Science Collaboration carried out a large-scale replication study to examine the reproducibility of psychological research (Open Science Collaboration, 2015). In this so-called Reproducibility Project Psychology (RPP), articles were sampled from the 2008 issues of three prominent and high-impact psychology journals and a key effect of each article was replicated according to a structured protocol. The results of the replications were not in line with the results of the original studies for the majority of replicated effects. For instance, 97% of the original studies reported a statistically significant effect for a key hypothesis, whereas only 36% of the replicated effects were statistically significant (Open Science Collaboration, 2015). Moreover, the average effect size of replication studies was substantially smaller (r = 0.197) compared to those of original studies (r = 0.403). Hence, the results of the RPP confirm both the excess of significant findings and overestimation of published effects within psychology.

The larger effect size estimates in the original studies compared to their replications can be explained by the expected value of a statistically significant original study being larger than the true mean (i.e., overestimation). The observed effect size of a replication, which has not (yet) been subjected to selection for statistical significance, will usually be smaller. This statistical principle of an extreme score on a variable (in this case a statistically significant effect size) being followed by a score closer to the true mean is also known as regression to the mean (e.g., Straits & Singleton, 2011). Regression to the mean occurs if simultaneously (i) selection occurs on the first measure (in our case, only statistically significant effects), and (ii) both of

the measures are subject to error (in our case, sampling error).

It is crucial to realize that the expected value of statistically significant observed effects of the original studies will be larger than the true effect size *irrespective of the presence of publication bias*. That is, conditional on being statistically significant, the expected value of the original effect size will be larger than the true effect size. The distribution of the statistically significant original effect size is actually a truncated distribution at the critical value, and these effect sizes are larger than the nonsignificant observed effects. Hence, the truncated distribution of statistically significant effects has a larger expected value than the true effect size. Publication bias only determines how often statistically nonsignificant effects get published, and therefore it does not influence the expected value of the statistically significant effects. Consequently, statistical analyses based on an effect that was selected for replication because of its significance should correct for the overestimation in effect size irrespective of the presence of publication bias.

Estimating effect size and determining whether an effect does truly exist based on an original published study and a replication is important. This is not only relevant for projects such as the RPP. Because replicating published research is often the starting point for new research where the replication is the first study of a multi-study paper (Neuliep & Crandall, 1993), it is also relevant for researchers who carry out a replication and want to aggregate the results of the original study and their replication. Cumming (2012) emphasized that combining two studies by means of a meta-analysis has added value over interpreting two studies in isolation. Moreover, researchers in the field of psychology also started to use meta-analysis to combine the studies within a single paper with what is called an *internal* meta-analysis (Ueno, Fastrich, & Murayama, 2016). Additionally, the proportion of published replication studies will increase in the near future due to the widespread attention for replicability of psychological research nowadays. Finally, we must note that the estimate of Makel, Plucker, and Hegarty (2012) of 1% of published studies in psychology being replications is a gross underestimation. They searched for the word "replication" and variants thereof in psychological articles. However, researchers do not label studies as replications to increase the likelihood of publication (Neuliep & Crandall, 1993), even though many of them carry out a replication before starting their own variation of the study. To conclude, making sense of and combining the results of original study and replication is a common and important problem.

The main difficulty with combining an original study and a replication is *how* to aggregate a likely overestimated effect size in the published original study with the unpublished and probably unbiased replication. For instance, what should a researcher conclude when the original study is statistically significant and the replication is not? This situation often arises, e.g., of the 100 effects examined in the RPP, in 62% of the cases the original study was statistically significant while the replication was not. To examine the main problem in more detail, consider the

following hypothetical situation. Both the original study and replication consist of two independent groups of equal size, with the total sample size in the replication being twice as large as in the original study (80 vs. 160). The researcher may encounter the following standardized effect sizes (Hedges' g)[19], t-values, and two-tailed p-values: g = 0.490; $t(78)$ = 2.211; p = .03 for the original study and g = 0.164; $t(158)$ = 1.040; p = .3 for the replication. A logical next step for interpreting these results would be to combine the observed effect sizes of both the original study and replication by means of a fixed-effect meta-analysis. The results of such a meta-analysis suggest that there is indeed an effect in the population after combining the studies with meta-analytic effect size estimate $\hat{\theta}$ = 0.270; z = 2.081; p = .0375 (two-tailed). However, the researcher may not be convinced that the effect really exists and does not know how to proceed since the original study is probably biased and the meta-analysis does not take this bias into account.

The aim of this chapter is threefold. First, we will develop a method (i.e., the hybrid method of meta-analysis, hybrid for short) that combines a statistically significant original study and replication and does correct for the likely overestimation in the original study's effect size estimate. The hybrid method yields (a) an accurate estimate of the underlying population effect based on the original study and the replication, (b) a confidence interval around this effect size estimate, and (c) a test of the null hypothesis of no effect for the combination of the original study and replication. Second, we will apply the hybrid method and also traditional meta-analysis methods to the data of the RPP to examine the reproducibility of psychological research. Third, to assist practical researchers in assessing effect size using an original and replication study, we formulate guidelines on which method to use under what conditions, and explain a newly developed web-based application for estimation based on these methods.

The remainder of this chapter is structured as follows. We explain traditional meta-analysis and propose the new hybrid method for combining an original study and a replication while taking into account statistical significance of the original study's effect. We adopt a combination of the frameworks of Fisher and Neyman-Pearson that is nowadays commonly used in practice to develop and examine our procedures for testing and estimating effect size. Next, we analytically approximate the performance of meta-analysis and the hybrid method in a situation where an original study and its replication are combined. The performance of meta-analysis and the hybrid method are compared to each other, and to estimation using only the

[19] Hedges' g is an effect size measure for a two-independent groups design that corrects for the small positive bias in Cohen's d by multiplying the Cohen's d effect sizes with correction factor $J = 1 - \frac{3}{4df - 1}$, where df refers to the degrees of freedom (Hedges, 1981). Note that different estimators for the effect size in a two-independent groups design exist, and that Hedges' g and Cohen's d are just two of these estimators (for other estimators see Viechtbauer [2007a] and Hedges [1981]).

replication. Based on the performance of the methods, we formulate guidelines on which method to use under what conditions. Subsequently, we describe the RPP and apply meta-analysis and the hybrid method to these data. This chapter concludes with a discussion and an illustration of a web-based application (https://rvanaert.shinyapps.io/hybrid) allowing straightforward application of the hybrid method to researchers' applications.

6.1 Methods for estimating effect size

The statistical technique for estimating effect size based on multiple studies is meta-analysis (Borenstein et al., 2009). The advantage of meta-analysis over interpreting the studies in isolation is that the effect size estimate in a meta-analysis is more precise. Two meta-analysis methods are often used: fixed-effect meta-analysis and random-effects meta-analysis. Fixed-effect meta-analysis assumes that there is one common population effect size underlying the studies in the meta-analysis while random-effects meta-analysis assumes that each study has its own population effect size. The studies' population effect sizes in random-effects meta-analysis are assumed to be a random sample from a normal distribution of population effect sizes and one of the aims of random-effects meta-analysis is to estimate the mean of this distribution (e.g., Borenstein et al., 2009).

Fixed-effect rather than random-effects meta-analysis is the recommended method to aggregate the findings of an original study and an exact or direct replication, assuming both studies assess the same underlying population effect. Note also that statistically combining two studies by means of random-effects meta-analysis is practically infeasible since the amount of heterogeneity among a small number of studies cannot be accurately estimated (e.g., Borenstein et al., 2010; IntHout et al., 2014). After discussing fixed-effect meta-analysis, we introduce the hybrid method as an alternative meta-analysis method that takes into account the statistical significance of the original study.

6.1.1 Fixed-effect meta-analysis

Before the average effect size with a meta-analysis can be computed, studies' effect sizes and sampling variances have to be transformed to one common effect size measure (see Borenstein, 2009; Fleiss & Berlin, 2009). The true effect size (θ) is estimated in each study with sampling error (ε_i). This model can be written as

$$y_i = \theta + \varepsilon_i,$$

where y_i reflects the effect size in the ith study and it is assumed that the ε_i is normally and independently distributed, $\varepsilon_i \sim N(0, \sigma_i^2)$ with σ_i^2 being the sampling variance in the population for each study. These sampling variances are assumed to be known in

meta-analysis.

The average effect size is computed by weighting each y_i with the reciprocal of the estimated sampling variance ($w_i = \frac{1}{\hat{\sigma}_i^2}$). For k studies in a meta-analysis, the weighted average effect size estimate ($\hat{\theta}$) is computed by

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i y_i}{\sum_{i=1}^{k} w_i} \quad (1)$$

with variance $v_{\hat{\theta}} = \frac{1}{\sum_{i=1}^{k} w_i}$. A 95% confidence interval around $\hat{\theta}$ can be obtained by $\hat{\theta} \pm 1.96\sqrt{v_{\hat{\theta}}}$ with 1.96 being the 97.5th percentile of the normal distribution and a z-test can be used to test H_0: $\theta = 0$, $z = \frac{\hat{\theta}}{\sqrt{v_{\hat{\theta}}}}$. Applying fixed-effect meta-analysis to the example as presented in the introduction, we first have to compute the sampling variance of the Hedges' g effect size estimates for the original study and replication. An unbiased estimator of the variance of y is computed by $\hat{\sigma}^2 = \frac{1}{n_1} + \frac{1}{n_2} + \left[\frac{1-(n_1+n_2-4)}{(n_1+n_2-2)J^2}\right] g^2$ where n_1 and n_2 are the sample sizes for group 1 and 2 (Viechtbauer, 2007). This yields weights 19.390 and 39.863 for the original study and replication, respectively. Computing the fixed-effect meta-analytic estimate (Equation 1) with y_i being the Hedges' g observed effect size estimates gives

$$\hat{\theta} = \frac{19.390 \times 0.490 + 39.863 \times 0.164}{19.390 + 39.863} = 0.270$$

with corresponding variance is $v_{\hat{\theta}} = \frac{1}{(19.390 + 39.863)} = 0.017$. The 95% confidence interval of the fixed-effect meta-analytic estimate ranges from 0.016 to 0.525, and the null hypothesis of no effect is rejected (z = 2.081, two-tailed p-value = 0.0375). Note that the t-distribution was used as reference distribution for testing the original study and replication individually whereas a normal distribution was used in the fixed-effect meta-analysis. The use of a normal distribution as reference distribution in fixed-effect meta-analysis is a consequence of the common assumptions in meta-analysis of

known sampling variances and normal sampling distributions of effect size (Raudenbush, 2009).

6.1.2 Hybrid Method

Like fixed-effect meta-analysis, the hybrid method estimates the common effect size of an original study and replication. By taking into account that the original study is statistically significant, the proposed hybrid method corrects for the likely overestimation in the effect size of the original study. The hybrid method is based on the statistical principle that the distribution of *p*-values at the true effect size is uniform. A special case of this statistical principle is that the *p*-values are uniformly distributed under the null hypothesis (e.g., Hung et al., 1997). This principle is also underlying the recently developed meta-analytic techniques *p*-uniform (van Aert, Wicherts, et al., 2016; van Assen et al., 2015) and *p*-curve (Simonsohn et al., 2014a, 2014b). These methods discard statistically nonsignificant effect sizes, and only use the statistically significant effect sizes in a meta-analysis to examine publication bias. *P*-uniform and *p*-curve correct for publication bias by computing probabilities of observing a study's effect size conditional on the effect size being statistically significant. The effect size estimate of *p*-uniform and *p*-curve equals that effect size for which the distribution of these conditional probabilities is best approximated by a uniform distribution. Both methods yield accurate effect size estimates in the presence of publication bias if heterogeneity in true effect size is at most moderate (Simonsohn et al., 2014a; van Aert, Wicherts, et al., 2016; van Assen et al., 2015). In contrast to *p*-uniform and *p*-curve, which assume that all included studies are statistically significant, only the original study is assumed to be statistically significant in the hybrid method. This assumption hardly restricts the applicability of the hybrid method since approximately 95% of the published psychological research contains statistically significant results (Fanelli, 2012; Sterling et al., 1995).

In order to deal with bias in the original study, its *p*-value is transformed by computing the probability of observing the effect size or larger conditional on the effect size being statistically significant and at the population effect size (θ).[20] This can be written as

$$q_O = \frac{P(y \geq y_O; \theta)}{P(y \geq y_O^{CV}; \theta)} \tag{2}$$

where the numerator refers to the probability of observing a larger effect size than in the original study (y_O) at effect size θ, and the denominator denotes the probability of

[20] Without loss of generality we assume the original study's effect size is positive. If the original effect size is negative, the direction of the original study, the replication, and the resulting combined estimated effect size should be reversed to obtain the required results.

observing an effect size larger than its critical value (y_O^{CV}) at effect size θ. Note that y_O^{CV} is independent of θ. The conditional probability q_O at true effect size θ is uniform whenever y_O is larger than y_O^{CV}. These conditional probabilities are also used in *p*-uniform for estimation and testing for an effect while correcting for publication bias (van Aert et al., 2016; van Assen et al., 2015). The replication is not assumed to be statistically significant, so we compute the probability of observing a larger effect size than in the replication (q_R) at effect size θ

$$q_R = P(y \geq y_R; \theta) \tag{3}$$

with the observed effect size of the replication denoted by y_R. Both q_O and q_R are calculated under the assumption that the sampling distributions of y_O and y_R are normally distributed, which is the common assumption in meta-analysis (Raudenbush, 2009).

Testing of H_0: $\theta = 0$ and estimation is based on the principle that each (conditional) probability is uniformly distributed at the true value θ. Different methods exist for testing whether a distribution deviates from a uniform distribution. The hybrid method uses the distribution of the sum of independently uniformly distributed random variables (i.e., the Irwin-Hall distribution)[21], $x = q_O + q_R$, because this method is intuitive, showed good statistical properties in the context of *p*-uniform, and can also be used for estimating a confidence interval (van Aert et al., 2016). The probability density function of the Irwin-Hall distribution for x based on two studies is

$$f(x) = \begin{cases} x & 0 \leq x \leq 1 \\ 2 - x & 1 \leq x \leq 2 \end{cases}$$

and its cumulative distribution function is

$$F(x) = \begin{cases} \frac{1}{2}x^2 & 0 \leq x \leq 1 \\ -\frac{1}{2}x^2 + 2x - 1 & 1 \leq x \leq 2 \end{cases}. \tag{4}$$

Two-tailed *p*-values of the hybrid method can be obtained with $G(x)$,

$$G(x) = \begin{cases} x^2 & 0 \leq x \leq 1 \\ 2 - (-x^2 + 4x - 2) & 1 \leq x \leq 2 \end{cases}. \tag{5}$$

[21] Estimation was based on the Irwin-Hall distribution instead of maximum likelihood. The distribution of the likelihood is typically highly skewed if true effect size is close to zero and sample size of the original study is small (as is currently common in psychology), making the asymptotic standard errors of maximum likelihood inaccurate. The probability density function and the cumulative distribution function of the Irwin-Hall distribution are available in the software Mathematica (Wolfram Research Inc., 2015).

The null hypothesis H_0: $\theta = 0$ is rejected if $F(x \mid \theta = 0) \leq .05$ in case of a one-tailed test, and $G(x \mid \theta = 0) \leq .05$ in case of a two-tailed test. The 2.5th and 5th percentiles of the Irwin-Hall distribution are 0.224 and 0.316, respectively. Effect size θ is estimated as $F(x \mid \theta = \hat{\theta}) = .5$, or equivalently, that value of θ for which $x = 1$. The 95% confidence interval of θ, ($\hat{\theta}_L, \hat{\theta}_H$), is calculated as $F(x \mid \theta = \hat{\theta}_L) = .975$ and $F(x \mid \theta = \hat{\theta}_H) = .025$.

We will now apply the hybrid method to the example presented in the introduction. The effect size measure of the example in the introduction is Hedges' *g*, but the hybrid method can also be applied to an original study and replication where another effect size measure (e.g., correlation coefficient) is computed. Figure 6.1 illustrates the computation of q_O and q_R for $\theta = 0$ (Figure 6.1a.) and for $\theta = \hat{\theta}$ (Figure 6.1b.) based on the example presented in the introduction. The steepest distribution in both panels refers to the effect size distribution of the replication, which has the largest sample size. The conditional probability q_O for $\theta = 0$ (Figure 6.1a) equals the area larger than y_O^{CV} (intermediate gray color) divided by the area larger than y_O (dark gray): $q_O = \frac{0.015}{0.025} = 0.6$. The probability q_R equals the one-tailed *p*-value (0.3/2 = 0.15) and is indicated by the light gray area.[22] Summing these two probabilities gives x = 0.75, which is lower than the expected value of the Irwin-Hall distribution suggesting that effect size exceeds 0. The null hypothesis of no effect is not rejected, with a two-tailed *p*-value equal to .558 as calculated by Equation (5). Shifting θ to hybrid's estimate $\hat{\theta}$ = 0.103 yields x =1, as depicted in Figure 6.1b, with q_O = 0.655 and q_R = 0.345. Estimates of the lower and upper bound of a 95% confidence interval can also be obtained by shifting $\hat{\theta}$ till x equals the 2.5th and 97.5th percentile for the lower and upper bound of the confidence interval. The confidence interval of the hybrid method for the example ranges from $\hat{\theta}_L$ = -1.109 to $\hat{\theta}_H$ = 0.428.

[22] The probabilities q_O and q_R are not exactly equal to 0.6 and 0.15 due to transforming the effect sizes from Cohen's *d* to Hedges' *g*. The conditional probabilities based on the transformed effect sizes are $q_O = \frac{0.0156}{0.0261} = 0.596$ and q_R = 0.151. Transforming the effect sizes from Cohen's *d* to Hedges' *g* may bias effect size estimates of the hybrid method. We studied to what extent q_O and q_R are influenced by this transformation of effect size. This distributions of q_O and q_R based on the transformed effect sizes were analytically approximated by means of numerical integration (see supplementary material available at https://osf.io/9e3qd/ for more information and the results), and these distributions should closely follow a uniform distribution according to the theory underlying the hybrid method. Results show that distributions of q_O and q_R after the transformation are accurate approximations of uniform distributions. Hence, the transformation from Cohen's *d* to Hedges' *g* will hardly bias the estimates of the hybrid method.

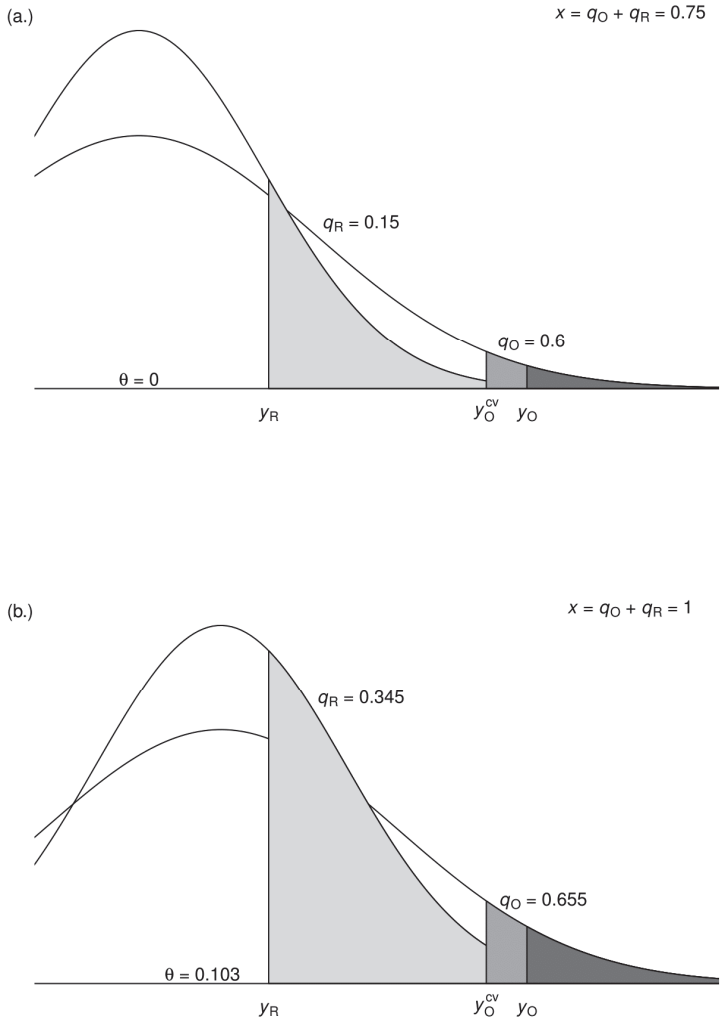

Figure 6.1. Effect size distributions of the original study and replication for the example presented in the introduction. Panels (a) and (b) refer to the effect size distributions for $\theta = 0$ and $\theta = \hat{\theta} = 0.103$. y_O and y_R denote the observed effect sizes in the original study and replication, and y_O^{CV} denotes the critical value of the original study based on a two-tailed hypothesis test of H_0: $\theta = 0$ with $\alpha = .05$. The shaded regions refer to probabilities larger than y_R, y_O, and y_O^{CV}. The (conditional) probabilities of the original study and replication are indicated by q_O and q_R, and its sum by x.

The results of applying fixed-effect meta-analysis and the hybrid method to the example are summarized in Table 6.1. The original study suggests that the effect size is medium and statistically significant different from zero (first row), but the

effect size in the replication is small at best and not statistically significant (second row). Fixed-effect meta-analysis (third row) is usually seen as the best estimator of the effect size in the population and suggests that the effect size is small to medium (0.270) and statistically significant (p = .0375). However, the hybrid's estimate is small (0.103) and not statistically significant (p = 0.558) (fourth row). Hybrid's estimate is lower than the estimate of fixed-effect meta-analysis because it corrects for the first study being statistically significant. Hybrid's estimate is even lower than the estimate of the replication because, when taking the significance of the original study into account, the original study suggest a zero or even negative effect, which pulls the estimate to zero.

Table 6.1. Effect size estimate (Hedges' g), 95% confidence interval (CI), and two-tailed p-value of the original study and replication in the hypothetical situation and results of fixed-effect meta-analysis and the hybrid, hybrid^0, and hybrid^R method when applied to the hypothetical situation.

Method	$\hat{\theta}$ (95% CI) [p-value]
Original study (y_O)	0.490 (.044; .935) [.0311]
Replication (y_R)	0.164 (-.147; .474) [.302]
Fixed-effect meta-analysis	0.270 (.016; .525) [.0375]
Hybrid	0.103 (-1.109; .428) [.558]
Hybrid^0	0.103 (-1.109; .429) [.558]
Hybrid^R	0.164 (-.147; .474) [.302]

Van Aert et al. (2016) showed that not only the lower bound of a 95% confidence interval, but also the estimated effect sizes by p-uniform can become highly negative if the effect size is estimated based on a single study and its p-value is close to the alpha level.[23] The effect size estimates can be highly negative because conditional probabilities like q_O are not sensitive to changes in θ when the (unconditional) p-value is close to alpha. Applying p-uniform to a single study where a one-tailed test is conducted with α=.05 yields an effect size estimate of p-uniform equal to zero if the p-value is .025, a positive estimate if the p-value is smaller than .025, a negative estimate if the p-value is larger than .025, and a highly negative

[23] In case of a two-tailed hypothesis test, the alpha level has to be divided by 2 because it is assumed that all observed effect sizes are statistically significant in the same direction.

estimate if the *p*-value is close to .05. Van Aert et al. (2016) recommended to set the effect size estimate equal to zero if the mean of the primary studies' *p*-values is larger than half the alpha level, because *p*-uniform's effect size estimate will then be below zero. Setting the effect size to 0 is analogous to testing a one-tailed null hypothesis where the observed effect size is in the opposite direction than expected. Computing a test statistic and *p*-value is redundant in such a situation because the test statistic will be negative and the one-tailed *p*-value will be above .5.

The hybrid method can also yield highly negative effect size estimates because, like *p*-uniform, it uses a conditional probability for the original study's effect size. In line with the proposal in van Aert et al. (2016), we developed two alternative hybrid methods, hybrid^0 and hybrid^R, to avoid highly negative estimates. The hybrid^0 method is a direct application of the *p*-uniform method as recommended in van Aert et al. (2016), which recommends setting the effect size estimate to 0 if the studies' combined evidence points to a negative effect. Applied to the hybrid^0 method, this translates to setting the effect size equal to 0 if $x > 1$ under the null hypothesis, and equal to that of hybrid otherwise. Consequently, hybrid^0 will, in contrast to hybrid, never yield an effect size estimate which is below zero. Applied to the example, hybrid^0 equals hybrid's estimate because $x = 0.75$ under the null hypothesis.

The other alternative hybrid method, hybrid^R (R in hybrid^R refers to replication), addresses the problem of highly negative estimates in a different way. The estimate of hybrid^R is equal to hybrid's estimate if the original study's two-tailed *p*-value is smaller than .025 and is equal to the effect size estimate of the replication if the original study's two-tailed *p*-value is larger than .025. A two-tailed *p*-value of .025 in the original study is used because this results in a negative effect size estimate, which is not in line with both the theoretical expectation as well as the observed effect size in the original study. Hence, if the original study's just statistically significant effect size (i.e., $.025 < p < .05$) points to a negative effect, evidence of the original study is discarded and only the results of the replication are interpreted. The estimate of hybrid^R (and also of hybrid) is not restricted to be in the same direction as the original study as is the case for hybrid^0. The results of applying hybrid^R to the example are presented in the last row of Table 6.1. Hybrid^R only uses the observed effect size in the replication because the *p*-value in the original study, .03, exceeds .025, and hence yields the same results as the replication study as reported in the second row.

Since all discussed methods may yield different results it is important to examine their statistical properties. The next section describes the performance of the methods evaluated using an analytical approximation of these methods' results.

6.2 Performance of estimation methods: Analytical comparison

6.2.1 Method

We used the correlation coefficient as effect size measure because our application discussed later, the RPP, also used correlations. However, all methods can also deal with other effect size measures as for instance standardized mean differences. We analytically compared the performance of five methods; fixed-effect meta-analysis, estimation using only the replication (maximum likelihood), and the hybrid, hybrid^0, and hybrid^R method.

We evaluated the methods' statistical properties by using a procedure analogous to the procedure described in van Aert and van Assen (2017). The methods were applied to the joint probability density function (pdf) of statistically significant original effect size and replication effect size. This joint pdf was a combination of the marginal pdfs of the statistically significant original effect size and the replication effect size, and was approximated by using numerical integration. Both marginal pdfs depended on the true effect size and the sample size in the original study and replication. The marginal pdf of statistically significant original effect sizes was approximated by first creating 1,000 evenly distributed cumulative probabilities or percentiles P_i^O of this distribution given true effect size and sample size in the original study, with $P_i^O = 1-\pi+\frac{(i\times\pi)}{1{,}001}$. Here, π denotes the power of the null hypothesis test of no effect, i.e. the probability that effect size exceeds the critical value. We used the Fisher-*z* test, with α=.025 corresponding to common practice in psychological research where two-tailed hypothesis tests are conducted and only results in the predicted direction get published. For instance, if the null hypothesis is true the cumulative probabilities P_i^O are evenly distributed and range from $1-0.025+\frac{(1\times.025)}{1{,}001}=0.975025$ to $1-0.025+\frac{(1{,}000\times.025)}{1{,}001}=0.999975$. Finally, the 1,000 P_i^O values were converted by using a normal distribution to the corresponding 1,000 (statistically significant) Fisher-transformed correlation coefficients.

The marginal pdf of the replication was approximated by selecting another 1,000 equally spaced cumulative probabilities given true effect size and sample size of the replication with $P_i^R = \frac{i}{1{,}001}$. These cumulative probabilities range from $\frac{1}{1{,}001}=0.000999001$ to $\frac{1{,}000}{1{,}001}=0.999001$, and were subsequently also transformed to Fisher-transformed correlation coefficients by using a normal distribution. The joint pdf was obtained by multiplying the two statistically independent marginal pdfs, and

yielded 1,000×1,000=1,000,000 different combinations of statistically significant original effect size and replication effect size. The methods were applied to each of the combination of effect sizes in the original study and replication. For presenting the results, Fisher-transformed correlations were transformed to correlations.[24]

Statistical properties of the different methods were evaluated based on average effect size estimate, median effect size estimate, standard deviation of effect size estimate, root mean square error (RMSE), coverage probability (i.e., the proportion describing how often the true effect size falls inside the confidence interval), and statistical power and alpha for testing the null hypothesis of no effect. Population effect size (ρ) and sample size in the original study (N_O) and replication (N_R) were varied. Values for ρ were chosen to reflect no (0), small (0.1), medium (0.3), and a large (0.5) true effect as specified by Cohen (1988). Representative sample sizes within psychology were used for the computations by selecting the first quartile, median, and third quartile of the original study's sample size in the RPP: 31, 55, and 96. These sample sizes were used for the original study and replication. A sample size of 783 was also included for the replication to reflect a recommended practice where the sample size is determined with a power analysis to detect a small true effect with a statistical power of 0.8. The computations were conducted in R, using the parallel package for parallel computing (R Core Team, 2017). The root-finding bisection method (Adams & Essex, 2013) was used to estimate the effect size and the confidence interval of the hybrid method. R code of the analyses is available via https://osf.io/tzsgw/.

6.2.2 Results

A consequence of analyzing Fisher-transformed correlations instead of correlations is that the estimator of true effect size becomes slightly underestimated. However, this underestimation is negligible under the selected conditions for sample

[24] The variance of 1,000 equally spaced probabilities (.08325), which were used to generate the observed effect sizes in the replication, was not exactly equal to the variance in the population (.08333). In order to examine whether this smaller variance would bias the effect size estimates of the methods, we also computed the effect size estimates for 5,000 equally spaced probabilities for both the original study and replication (i.e., based on 25 instead of 1 million points). These effect size estimates were almost equal to the estimates based on 1,000 equally spaced probabilities (i.e., difference less than .0002). Therefore, we continued using 1,000 equally spaced probabilities for both marginal densities in our analyses.

size and true effect size.[25] The results of using only the replication data are the reference because the expected value of the replication's effect size is equal to the population effect size if no *p*-hacking or questionable research practices have been used. Both fixed-effect meta-analysis and the hybrid methods also use the data of the original study. In describing the results, we will focus on answering the question under which conditions these methods will improve upon estimation and testing using only the replication data.

Mean and median of effect size estimate Table 6.2 shows the methods' expected values as a function of the population effect size (ρ) and sample sizes in the original study (N_O) and replication (N_R). Expected values of the methods' estimators at N_R=783 are presented in Table 6.A1 because their bias is very small in those conditions. We also present the median effect size estimate (Figure 6.2[26]) since the expected value of the hybrid method is negative because hybrid's estimate becomes highly negative if the conditional probability is close to one (in other words, the probability distribution of hybrid's estimate is skewed to the left). Note that the median effect size estimates of the replication, hybrid, and hybrid^0 are all exactly equal to each other and therefore coincide in Figure 6.2.

The expected values based on the replication are exactly equal to the population effect size for ρ = 0, but are slightly smaller than the true value for larger population effect sizes. This underestimation is caused by transforming the Fisher-*z*-values to correlation coefficients.[27] The median estimate of the replication is exactly equal to the population effect size in all conditions (solid lines with filled bullets in Figure 6.2). Fixed-effect meta-analysis generally yields too high estimates when there is no or a small effect in the population, particularly if sample sizes are small (bias equal to 0.215 and 0.168 for no and small effect). However, its bias is small for a very large sample size of the replication (at most .026 for zero true effect size, and N_O=96 and N_R=783, see Table 6.A1). Bias decreases in population effect size and sample size,

[25] We examined the underestimation caused by transforming the correlations to Fisher-transformed correlations by computing the expected value and variance of the exact probability density distribution of the correlation (Hotelling, 1953) and the probability density distribution of the correlation that is obtained by applying the Fisher-transformation. This procedure for computing the expected value and variance is analogous to the one described in Schulze (2004, pp. 119-123). Of the conditions for sample size and true effect size (ρ) included in our study, bias in expected value and variance is largest for a sample size of 31 and true effect size of ρ=0.5. For this condition, the expected value and variance of the exact probability density distribution are 0.494 and 0.0260, respectively, and 0.487 and 0.0200 for the probability density distribution after applying the Fisher-transformation. In other conditions, bias was less than 0.004 and 0.002 for the expected value and variance, respectively.

[26] A line for each method was drawn through the points in Figures 2-5 to improve their interpretability. The lines do not reflect extrapolated estimates of the performance of the different methods for true effect sizes that were not included in our analytical approximation.

[27] Observed effect sizes were first transformed from Fisher-*z*-values to correlation coefficients before the average effect size was calculated. This caused a slight underestimation in the effect size estimate based on the replication study.

and is .037 or smaller if the population effect size is at least medium and both sample sizes are at least 55.

The estimator of the hybrid method has slight negative bias compared to the replication (never more than -0.021, Table 6.2) caused by the highly negative estimates if *x* is close to 2 under the null hypothesis. However, its median (dashed lines with filled squares in Figure 6.2) is exactly equal to the population effect size. Hybrid^0, which was developed to correct for the negative bias of hybrid's estimator, overcorrects and yields overestimated effect size for $\rho = 0$, with bias equal to 0.072 and 0.04 for small and large sample sizes, respectively. The positive bias of hybrid^0's estimator is small for small effect size (at most .027, for small sample sizes), whereas there is a small negative bias for medium and large effect size. Hybrid^0's median estimate is exactly equal to the population effect size (dashed lines with asterisks in Figure 6.2). Results of estimator hybrid^R parallel those of hybrid^0, but with less positive bias for no effect (0.049 and 0.027 for small and large sample sizes, respectively), more bias for small effect size (at most .043) and medium effect size (at most .023). The median estimate of hybrid^R slightly exceeds population effect size (dashed lines with triangles in Figure 6.2) because the data of the original study are omitted only if they indicate a negative effect.

To conclude, the negative bias of the hybrid's estimator is small, whereas the estimators of hybrid^R and hybrid^0 overcorrect this bias for no and small population effect size. The fixed-effect meta-analytic estimator yields severely overestimated effect sizes for no and small population effect size, but yields approximately accurate estimates for large effect size. Bias of all methods decreases if sample sizes increase and all methods yield accurate effect size estimates for large population effect size.

Precision Table 6.2 also presents the standard deviation of the effect size estimates reflecting the precision of these estimates. Standard deviations of the effect size estimates for $N_R = 783$ are presented in Table 6.A1, and are substantially smaller than the standard deviations of the other conditions for N_R. The fixed-effect meta-analytic estimator yields the most precise estimates. Precision of hybrid's estimator increases relative to the precision of the replication's estimator in population effect size and the ratio of original to replication sample size. For zero and small population effect size, the estimator of hybrid has lower precision than the replication's estimator if the replication sample size is equal or lower than the original sample size. For medium and large population effect size, the estimator of hybrid generally has higher precision, except when the sample size in the original study is much smaller than replication's sample size. Estimators of Hybrid^0 and hybrid^R have higher precision than hybrid's estimator because they deal with possible strongly negative estimates of hybrid, with hybrid^0's estimator in general being most precise for zero and small population effect size, and the estimator of hybrid^R being most precise for medium

Table 6.2. Effect size estimate and standard deviation of this estimate in brackets for the estimators of fixed-effect meta-analysis, replication study and hybrid, hybrid^0, and hybrid^R method as a function of population effect size ρ and sample size of the original study (N_O) and replication (N_R).

		N_R=31			N_R=55			N_R=96		
	ρ	N_O=31	N_O=55	N_O=96	N_O=31	N_O=55	N_O=96	N_O=31	N_O=55	N_O=96
FE	0	0.215 (.094)	0.207 (.069)	0.184 (.049)	0.152 (.089)	0.16 (.071)	0.154 (.053)	0.101 (.079)	0.115 (.067)	0.12 (.053)
	0.1	0.268 (.093)	0.248 (.07)	0.217 (.053)	0.219 (.088)	0.215 (.071)	0.198 (.055)	0.179 (.078)	0.183 (.067)	0.177 (.054)
	0.3	0.381 (.09)	0.349 (.076)	0.318 (.068)	0.357 (.084)	0.337 (.072)	0.315 (.065)	0.338 (.073)	0.327 (.065)	0.312 (.059)
	0.5	0.516 (.086)	0.499 (.079)	0.497 (.068)	0.511 (.076)	0.499 (.071)	0.498 (.062)	0.507 (.064)	0.5 (.06)	0.498 (.055)
Replica-tion	0	0 (.182)	0 (.182)	0 (.182)	0 (.135)	0 (.135)	0 (.135)	0 (.102)	0 (.102)	0 (.102)
	0.1	0.097 (.18)	0.097 (.18)	0.097 (.18)	0.098 (.134)	0.098 (.134)	0.098 (.134)	0.099 (.101)	0.099 (.101)	0.099 (.101)
	0.3	0.291 (.167)	0.291 (.167)	0.291 (.167)	0.295 (.124)	0.295 (.124)	0.295 (.124)	0.297 (.093)	0.297 (.093)	0.297 (.093)
	0.5	0.487 (.141)	0.487 (.141)	0.487 (.141)	0.493 (.103)	0.493 (.103)	0.493 (.103)	0.496 (.077)	0.496 (.077)	0.496 (.077)

Table 6.2. Continued

		N_R=31			N_R=55			N_R=96		
	ρ	N_0=31	N_0=55	N_0=96	N_0=31	N_0=55	N_0=96	N_0=31	N_0=55	N_0=96
Hybrid	0	-0.013 (.195)	-0.016 (.182)	-0.019 (.168)	-0.007 (.155)	-0.01 (.146)	-0.012 (.136)	-0.004 (.122)	-0.006 (.117)	-0.007 (.11)
	0.1	0.083 (.189)	0.081 (.173)	0.078 (.155)	0.09 (.15)	0.088 (.139)	0.086 (.126)	0.094 (.119)	0.092 (.112)	0.091 (.103)
	0.3	0.279 (.164)	0.28 (.14)	0.285 (.112)	0.287 (.131)	0.287 (.114)	0.29 (.094)	0.292 (.105)	0.292 (.093)	0.293 (.079)
	0.5	0.483 (.123)	0.491 (.094)	0.496 (.072)	0.489 (.099)	0.494 (.079)	0.497 (.063)	0.493 (.08)	0.496 (.066)	0.498 (.055)
Hybrid^0	0	0.072 (.101)	0.065 (.09)	0.057 (.079)	0.058 (.083)	0.054 (.075)	0.048 (.067)	0.047 (.067)	0.044 (.062)	0.04 (.057)
	0.1	0.127 (.127)	0.12 (.115)	0.112 (.102)	0.117 (.11)	0.112 (.101)	0.107 (.092)	0.11 (.094)	0.107 (.088)	0.104 (.081)
	0.3	0.285 (.149)	0.284 (.13)	0.287 (.106)	0.289 (.126)	0.288 (.111)	0.29 (.092)	0.292 (.103)	0.292 (.092)	0.293 (.078)
	0.5	0.483 (.122)	0.491 (.093)	0.496 (.072)	0.489 (.099)	0.494 (.079)	0.497 (.063)	0.493 (.08)	0.496 (.066)	0.498 (.055)

Table 6.2. Continued

		N_R=31			N_R=55			N_R=96		
	ρ	N_0=31	N_0=55	N_0=96	N_0=31	N_0=55	N_0=96	N_0=31	N_0=55	N_0=96
Hybrid^R	0	0.049 (.172)	0.043 (.164)	0.038 (.157)	0.04 (.133)	0.036 (.128)	0.032 (.122)	0.032 (.104)	0.03 (.1)	0.027 (.096)
	0.1	0.143 (.164)	0.136 (.153)	0.128 (.142)	0.136 (.128)	0.131 (.12)	0.125 (.112)	0.13 (.1)	0.126 (.095)	0.122 (.089)
	0.3	0.323 (.139)	0.312 (.123)	0.302 (.102)	0.321 (.11)	0.312 (.099)	0.303 (.085)	0.319 (.088)	0.312 (.08)	0.304 (.071)
	0.5	0.501 (.107)	0.495 (.089)	0.496 (.072)	0.503 (.087)	0.497 (.076)	0.497 (.063)	0.504 (.071)	0.498 (.064)	0.498 (.055)

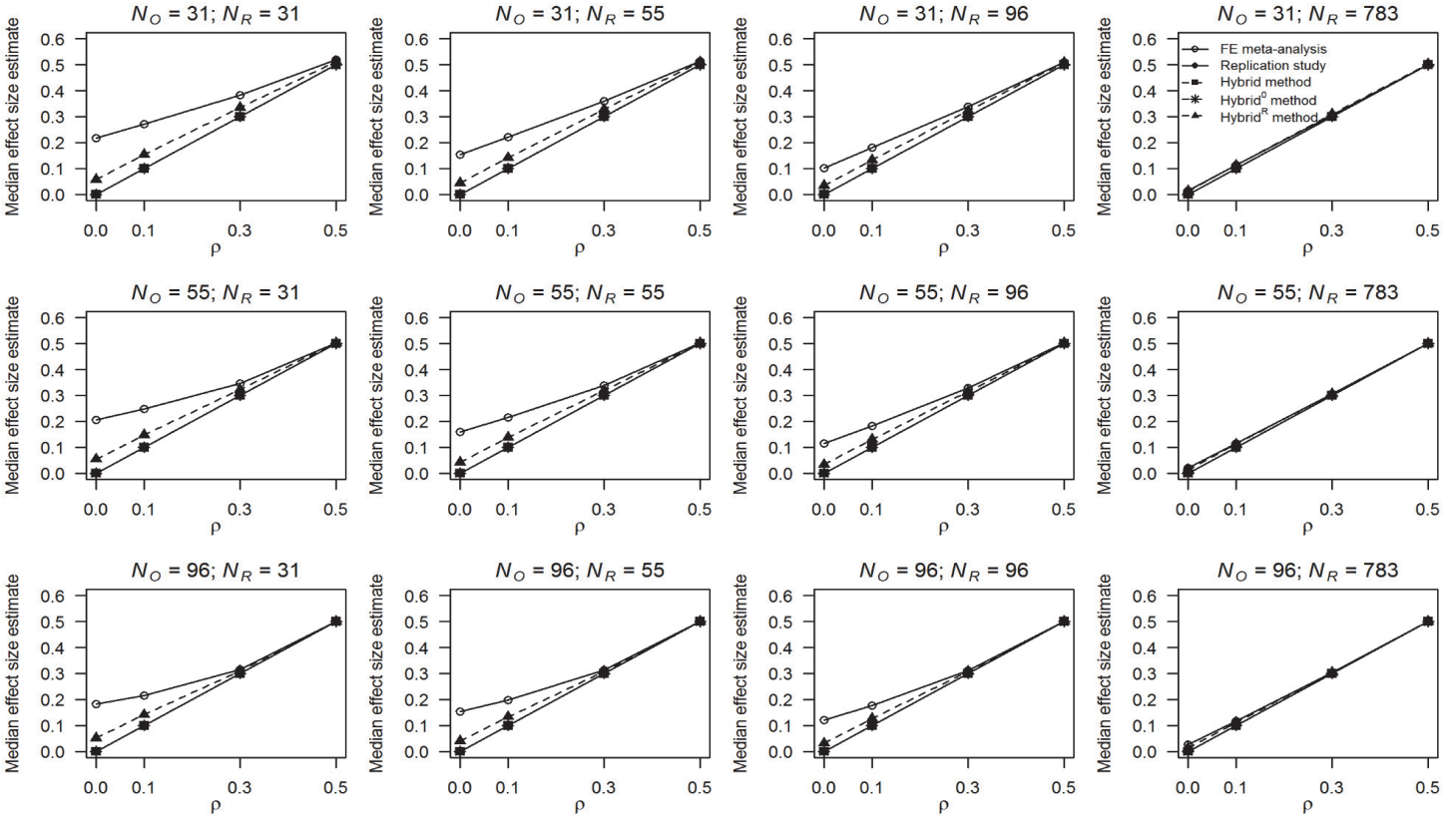

Figure 6.2. Median effect size estimate of the estimators of fixed-effect meta-analysis (solid line with open bullets), replication study (solid line with filled bullets) and hybrid (dashed line with filled squares), hybrid^0 (dashed line with asterisks), and hybrid^R method (dashed line with filled triangles) as a function of population effect size ρ and sample size of the original study (N_O) and replication (N_R). Median effect size estimates of the replication study, hybrid, and hybrid^0 are exactly equal to the population effect size and therefore coincide.

and large population effect size. They also have higher precision compared to the estimator of the replication, but not when replication's sample size is larger than the sample size of the original study and at the same time the effect size is medium or large in the population (hybrid^0; N_O = 31/55 and N_R = 96) or zero (hybrid^R; N_O = 31 and N_R = 96).

RMSE The RMSE combines two important statistical properties of an estimator: bias and precision. A slightly biased and very precise estimator is often preferred over an unbiased but very imprecise estimator. The RMSE is an indicator of this trade-off between bias and precision and is displayed in Figure 6.3. Compared to the replication's estimator, the RMSE of the fixed-effect meta-analytic estimator is higher for no effect in the population and smaller for medium and large effect size. For small population effect size, RMSE of the estimators of the replication and fixed-effect meta-analysis are roughly the same for equal sample sizes while RMSE of the replication's estimator was higher for $N_O > N_R$ and lower for $N_O < N_R$. Comparing the estimators of hybrid to the replication for equal sample sizes of both studies, hybrid's RMSE is higher for zero and small population effect size, but lower for medium and large population effect size. However, performance of hybrid's estimator relative to the estimator of the replication depends on sample sizes, and increases in ratio N_O/N_R. RMSE of the estimators of hybrid^0 and hybrid^R are always lower than that of hybrid's estimator. They are also lower than RMSE of the replication, except for N_O = 31 and N_R = 96 and at the same time zero or small population effect size (hybrid^R), and medium or large population effect size (hybrid^0). RMSE of the estimators of hybrid^0 and hybrid^R are lower than of the fixed-effect meta-analytic estimator for zero or small population effect size, and higher for medium or large population effect size. For N_R = 783, RMSE of all estimators were close to each other (see figures in last column of Figure 6.3).

Statistical properties of the test of no effect Figure 6.4 presents Type I error and statistical power of all methods' testing procedures. The Type I error rate is exactly .025 for the replication, hybrid, and hybrid^0 method. Type I error rate is slightly too high for hybrid^R (.037 in all conditions) and substantially too high for fixed-effect meta-analysis (increases in N_O/N_R, up to 0.551 for N_O = 96 and N_R = 31). Concerning statistical power, fixed-effect meta-analysis has by far the highest power because of its overestimation in combination with high precision. With respect to statistical power of the other methods, we first consider the cases with equal sample sizes of both studies. Here, hybrid^R has highest statistical power, followed by the replication. Hybrid and hybrid^0 have about equal statistical power compared to the replication for zero and small population effect size, but lower statistical power for medium and large population effect size. For $N_O > N_R$, all hybrid methods have higher power than the replication. For $N_O < N_R$ and $N_R < 783$, hybrid^R has higher statistical power than the replication for zero or small population effect size, but lower statistical power for medium or large population effect size; hybrid and hybrid^0 have lower statistical

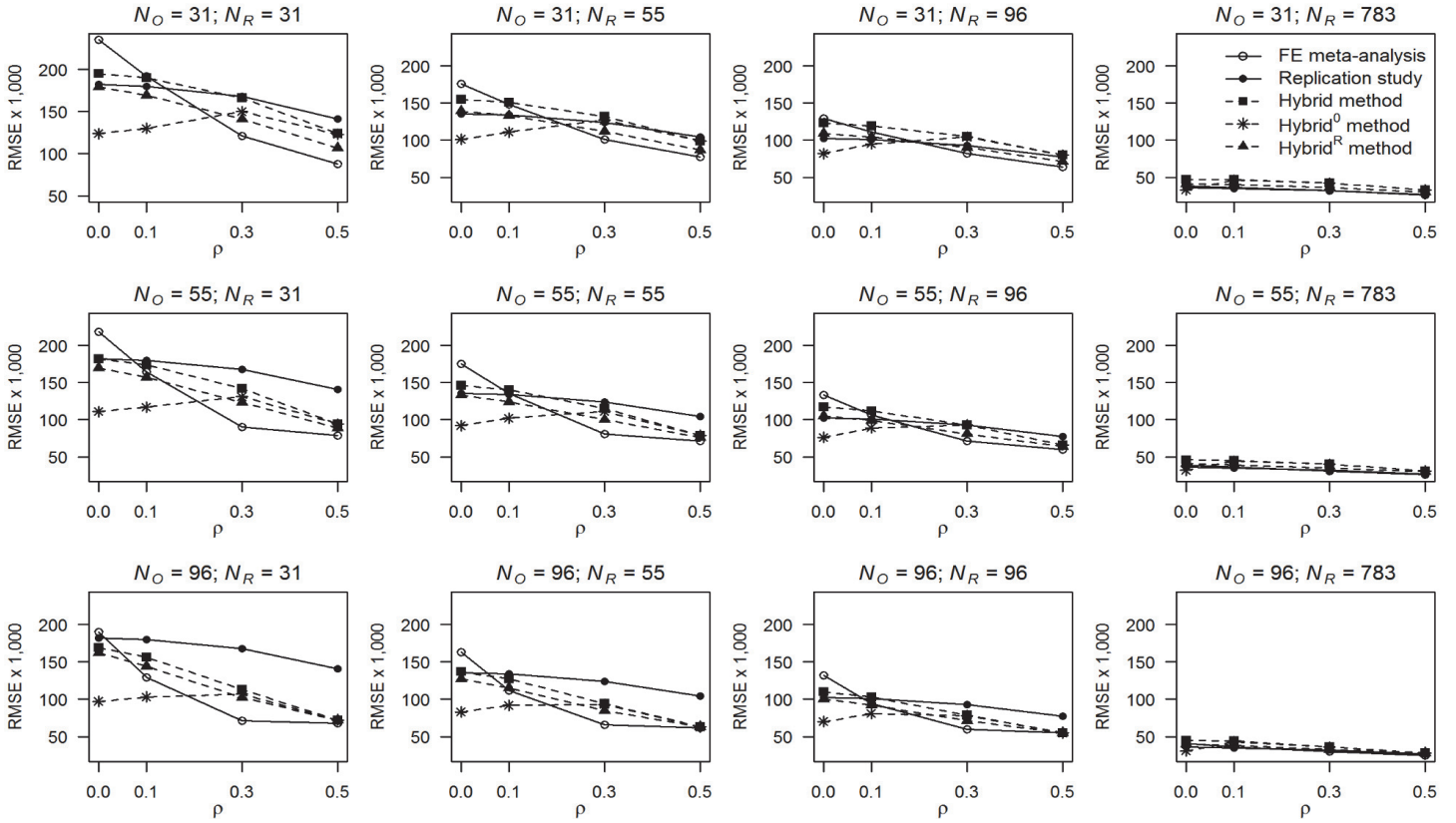

Figure 6.3. Root mean square error (RMSE) of the estimators of fixed-effect meta-analysis (solid line with open bullets), replication study (solid line with filled bullets) and hybrid (dashed line with filled squares), hybrid^0 (dashed line with asterisks), and hybrid^R method (dashed line with filled triangles) as a function of population effect size ρ and sample size of the original study (N_O) and replication (N_R).

power than the replication in this case. Statistical power of the replication is 0.8 for ρ=0.1 and N_R=783 because sample size was determined to obtain a power of 0.8 in this condition and 1 for ρ>0.1 and N_R=783.

Coverage is presented in Figure 6.5.[28] The replication and hybrid yield coverage probabilities exactly equal to 95% in all conditions. Coverage probabilities of fixed-effect meta-analysis are substantially too low for ρ = 0 and ρ = 0.1 due to the overestimation in average effect size; generally, its coverage improves in effect size and ratio N_R/N_O. Coverage probabilities of hybrid^0 and hybrid^R are close to .95 in all conditions.

Guidelines for applying methods Using the methods' statistical properties we attempted to answer the essential question which method to use, under what conditions. Answering this question is difficult because an important condition, population effect size, is unknown and in fact has to be estimated and tested. We will present guidelines (Table 6.3) that take this uncertainty into account. Each guideline is founded on and explained by using the previously described results.

The hybrid method and its variants have good statistical properties when testing the hypothesis of no effect, i.e., both Type I error and coverage are equal or close to .025 and 95%, respectively. Although the methods show similar performance, we recommend using hybrid^R over the hybrid and hybrid^0 method. Hybrid^R's estimator has small positive bias, but this bias is less than hybrid^0's estimator if the population effect size is zero. Moreover, hybrid^R's estimator has lower RMSE than hybrid, and has higher power than the testing procedures of hybrid and hybrid^0. Hence, in the guidelines below we consider when to use only the replication, fixed-effect meta-analysis, and hybrid^R.

If the magnitude of the effect size in the population is uncertain, fixed-effect meta-analysis has to be discarded because it generally yields highly overestimated effect size and a too high Type I error rate when the population effect size is zero or small (Guideline 1, Table 6.3). If the replication's sample size is larger than the original study, we recommend using only the replication (Guideline 1a) because then the replication outperforms hybrid^R with respect to power and provides accurate estimates. Additionally, the RMSE of the replication relative to hybrid^R gets more favorable for increasing N_R/N_O.

In case of uncertainty about the magnitude of the population effect size and the sample size in the replication is smaller than in the original study, we recommend using hybrid^R (Guideline 1b) because the estimator of hybrid^R

[28] The hybrid^0 method was omitted from Figure 6.5 illustrating the coverage probabilities because the average effect size estimate was set to zero if the p-value of the original study was larger than .0125. This made the confidence interval meaningless since the average effect size estimate could not be included in the confidence interval.

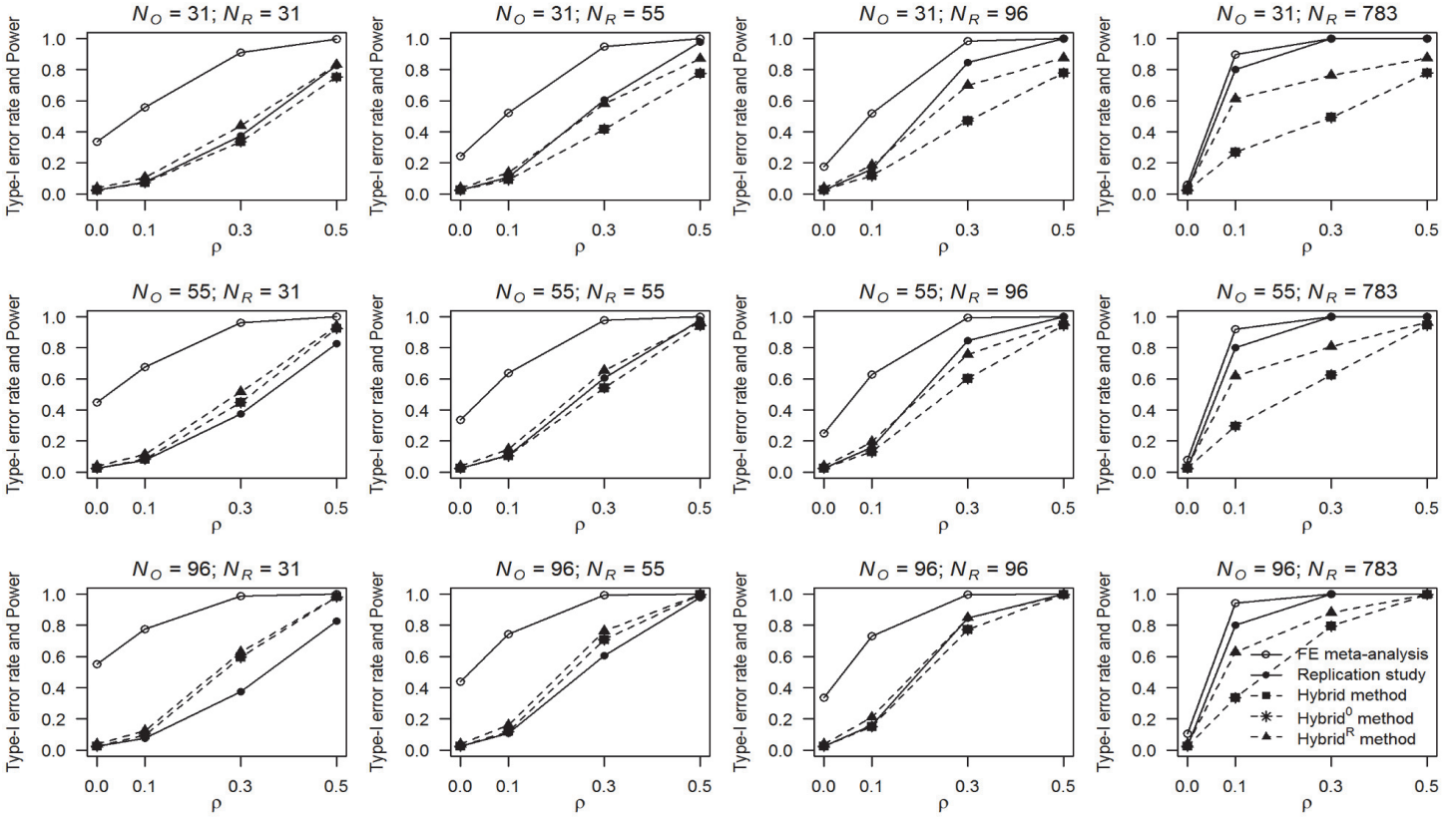

Figure 6.4. Type I error rate and statistical power of the testing procedures of fixed-effect meta-analysis (solid line with open bullets), replication study (solid line with filled bullets) and hybrid (dashed line with filled squares), hybrid^0 (dashed line with asterisks), and hybrid^R method (dashed line with filled triangles) as a function of population effect size ρ and sample size of the original study (N_O) and replication (N_R).

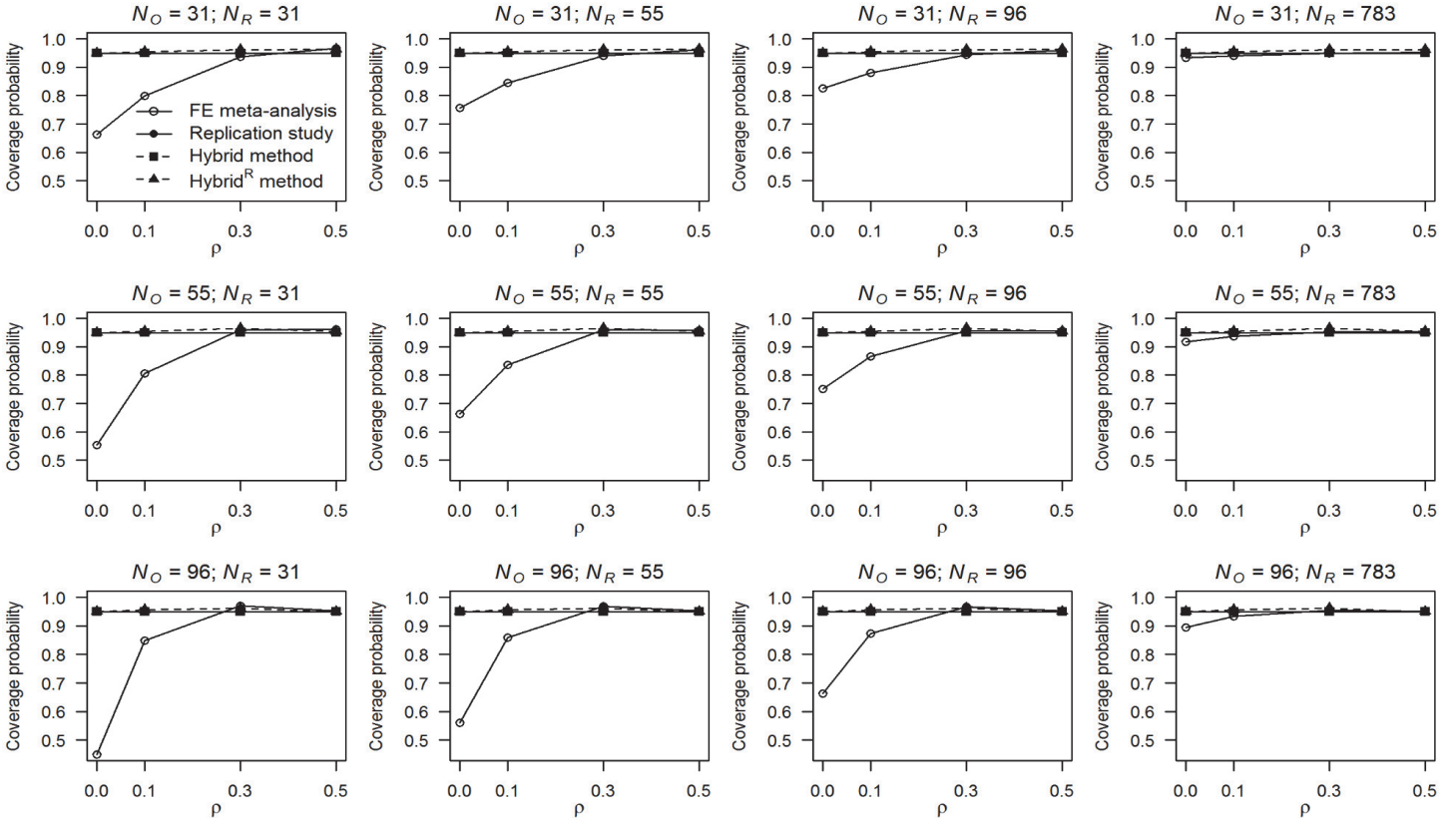

Figure 6.5. Coverage probability of fixed-effect meta-analysis (solid line with open bullets), replication study (solid line with filled bullets) and hybrid (dashed line with filled squares), and hybrid^R method (dashed line with filled triangles) as a function of population effect size ρ and sample size of the original study (N_O) and replication (N_R).

Table 6.3. Guidelines for applying which method to use when statistically combining an original study and replication

(1a) *When uncertain about population effect size* and sample size in the replication is larger than in the original study ($N_R > N_O$), use only the replication data.
(1b) *When uncertain about population effect size* and the sample size in the replication is equal or smaller than in the original study ($N_R \leq N_O$), use hybrid^R.
(2) *When suspecting zero or small population effect size*, use hybrid^R
(3) *When suspecting medium or larger population effect size*, use fixed-effect meta-analysis.

outperforms the replication's estimator with respect to RMSE and the testing procedure of hybrid^R yields larger statistical power than the procedure of the replication. For this situation, including the original data is beneficial since it contains sufficient information to improve estimation of effect size compared to using only the replication data. A drawback of using the hybrid^R method is that its Type I error rate is slightly too high (.037 vs. .025), but a slightly smaller alpha level can be selected to decrease the probability of falsely concluding that an effect exists. If information on the population effect size is known based on previous research, it is valuable to include this information in the analysis (akin to using an informative prior distribution in Bayesian analyses). If the population effect size is suspected to be zero or small we also recommend using hybrid^R (Guideline 2), because its estimator then has lower RMSE and only a small positive bias, and its testing procedure has higher statistical power than the replication. Fixed-effect meta-analysis should be abandoned in this case because its estimator overestimates zero and small population effects.

Fixed-effect meta-analysis is recommended if a medium or larger population effect size is expected (Guideline 3). Bias of the fixed-effect meta-analytic estimator is minor in this case, but its RMSE is smaller and the testing procedure has larger statistical power than of any other method. An important qualification of this guideline is the sample size of the original study because bias is a decreasing function of N_O. If N_O is small, statistical power of the original study's testing procedure is small when population effect size is medium, and consequently the original's effect size estimate is generally too high. Hence, to be on the safe side, if expecting a medium population effect size in combination with a small sample size in the original study, one can also decide to use only the replication data (if $N_R > N_O$) or hybrid^R (if $N_R \leq N_O$). When expecting a large population effect size and the main focus is not only on effect size estimation but also on testing, fixed-effect meta-analysis is the optimal choice.

However, if the ultimate goal of the analysis is to get an unbiased estimate of the effect size only the replication data should be used for the analysis. The replication is not published and its effect size estimate is therefore not affected by publication bias. Of course, the replication only provides an unbiased estimate if the research is well conducted, and for instance no questionable research practices were used.

6.3 Reproducibility Project Psychology (RPP)

The RPP was initiated to examine the reproducibility of psychological research (Open Science Collaboration, 2015). Articles from three high-impact psychology journals (Journal of Experimental Psychology: Learning, Memory, and Cognition [JEP: LMC], Journal of Personality and Social Psychology [JPSP], and Psychological Science [PSCI]) published in 2008 were selected to be replicated. The key effect of each article's last study was replicated according to a structured protocol, with authors of the original study being contacted for study materials and reviewing the planned study protocol and analysis plan to ensure the quality of the replication.

A total of 100 studies were replicated in the RPP. One requirement for inclusion in our analysis was that the correlation coefficient and its standard error could be computed for both the original study and replication. This was not possible for 27 study-pairs.[29] Moreover, transforming the effect sizes to correlation coefficients may have biased the estimates of the hybrid method since q_O and q_R may not exactly be uniformly distributed at the true effect size due to the transformation. We examined the influence of transforming effect sizes to correlation coefficients on the distributions of q_O and q_R, and concluded that the transformation of effect size will hardly bias the effect size estimates of the hybrid method (see supplemental materials available at https://osf.io/9e3qd/).

Another requirement for including a study-pair in the analysis was that the original study had to be statistically significant, which was not the case for six studies. Hence, fixed-effect meta-analysis and the hybrid methods could be applied on 67 study-pairs. The effect sizes of these study-pairs and the results of applying fixed-effect meta-analysis and the hybrid methods are available in Table 6.B1 in the Appendix. For completeness, we present the results of all three hybrid methods. The results in Table 6.B1 show that hybrid^0 sets the effect size to zero in 11 study-pairs (16.4%), i.e. where hybrid's effect size is negative, and hybrid^R also yielded 11 studies with results different from hybrid (16.4%); in five studies (7.5%) all three hybrid variants yielded different estimates.

Table 6.4 summarizes the results on effect size estimates of replications, fixed-effect meta-analysis, and the hybrid methods. For each method, the mean and

[29] If test statistics of the original study or replication were, for instance, $F(df1 >1, df2)$ or χ^2, the standard error of the correlation coefficient using Fisher transformation could not be computed and fixed-effect meta-analysis and the hybrid methods could not be applied to these study-pairs.

standard deviation of the estimates and the percentage of statistically significant results (i.e., $p < .05$) are presented. The columns in Table 6.4 refer to the overall results or results grouped per journal. Since PSCI is a multi-disciplinary journal, original studies published in PSCI were classified as belonging to cognitive or social psychology as in Open Science Collaboration (2015).

The estimator of fixed-effect meta-analysis yielded the largest average effect size estimate (0.322) and the highest proportion of statistically significant results (70.1%). We learned from the previous section to distrust these high numbers when being uncertain about true effect size, particularly in combination with a small sample size in the original study. The estimator of the replication yielded on average the lowest effect size estimates (0.199), with only 34.3% of the cases where the null hypothesis was rejected. The estimators of the hybrid variants yielded a higher average estimate (0.250-0.268), with an equal (hybrid^R) or lower proportion of rejecting the null hypothesis of no effect (hybrid and hybrid^0). The lower proportion of rejections of the null hypothesis by the hybrid methods is not only caused by the generally lower effect size estimates, but also by the much higher uncertainty in these estimates. The methods' uncertainty, expressed by the average width of confidence intervals, was: 0.328 (fixed-effect meta-analysis), 0.483 (replication), 0.648 (hybrid), 0.615 (hybrid^0), 0.539 (hybrid^R). The higher uncertainty of the hybrid methods than the replication demonstrates that controlling for the significance of the original study may come at high costs (i.e. an increase in uncertainty relative to estimation by the replication only), particularly when the ratio of the replication's to the original's sample size gets larger.

If we apply our guidelines to the data of the RPP and suppose that we are uncertain about the population effect size (Guidelines 1a and 1b in Table 6.3), only the replication data is interpreted 43 times because $N_R > N_O$ and hybrid^R is 24 times applied ($N_O \geq N_R$). The average effect size estimate of the replication's estimator with $N_R > N_O$ is lower than of the fixed-effect meta-analytic estimator (0.184 vs. 0.266), and the number of statistically significant pooled effect sizes is also lower (34.9% vs. 55.8%). The average effect size estimate of hybrid^R's estimator applied to the subset of 24 studies with $N_O \geq N_R$ is also lower than that of the fixed-effect meta-analytic estimator (0.375 vs. 0.421), and the same holds for the number of statistically significant results (54.2% vs. 95.8%).

The results per journal showed higher effect size estimates and more rejections of the null hypothesis of no effect for cognitive psychology (JEP: LMC and PSCI: cog.) than social psychology (JPSP and PSCI: soc.), independent of the method. The estimator of fixed-effect meta-analysis yielded higher estimates and the null hypothesis was more often rejected compared to the other methods. Estimates of the replication were always lower than of the hybrid methods. The number of statistically significant results of hybrid and hybrid^0 were equal or lower than of the replication, whereas the number of statistically significant results of hybrid^R was equal or higher

Table 6.4. Summary results of effect size estimates and percentage of times the null hypothesis of no effect was rejected of fixed-effect meta-analysis (FE), replication, hybrid, hybrid^R, and hybrid^0 method to 67 studies of the Reproducibility Project Psychology.

		Overall	JEP: LMC	JPSP	PSCI: cog.	PSCI: soc.
Number of study-pairs		67	20	18	13	16
Mean (SD)	FE	0.322 (.229)	0.416 (.205)	0.133 (.083)	0.464 (.221)	0.300 (.241)
	Replication	0.199 (.280)	0.291 (.264)	0.026 (.097)	0.289 (.365)	0.206 (.292)
	Hybrid	0.250 (.263)	0.327 (.287)	0.071 (.087)	0.388 (.260)	0.245 (.275)
	Hybrid^0	0.266 (.242)	0.353 (.237)	0.080 (.075)	0.400 (.236)	0.257 (.259)
	Hybrid^R	0.268 (.254)	0.368 (.241)	0.083 (.093)	0.394 (.272)	0.247 (.271)
%Significant results (i.e., *p*-value < .05)	FE	70.1%	90%	44.4%	92.3%	56.2%
	Replication	34.3%	50%	11.1%	46.2%	31.2%
	Hybrid	28.4%	45%	11.1%	30.8%	25%
	Hybrid^0	28.4%	45%	11.1%	30.8%	25%
	Hybrid^R	34.3%	55%	16.7%	38.5%	25%

Note. % Significance was based on two-tailed *p*-values; JEP: LMC = Journal of Experimental Psychology: Learning, Memory, and Cognition; JPSP = Journal of Personality and Social Psychology; PSCI: cog. = Psychological Science cognitive psychology; PSCI: soc. = Psychological Science social psychology

than of hybrid and hybrid^0. Particularly striking is the low number of statistically significant results for JPSP: 16.7% (hybrid^R) or 11.1% (replication, hybrid, and hybrid^0).

We also computed a measure of association to examine how often the methods yielded the same conclusions with respect to the test of no effect for all study-pairs together and grouped per journal. Since this resulted in a dichotomous variable, we used Loevinger's *H* (Loevinger, 1948) as the measure of association. Table 6.5 shows Loevinger's *H* of the replication with each other method for all 67 study-pairs. Associations between fixed-effect meta-analysis, hybrid, hybrid^0, and

Table 6.5. Loevinger's *H* across all 67 studies of all methods' results of hypothesis testing.

	FE	Hybrid	Hybrid^0	Hybrid^R
Replication	1	0.519	0.519	0.603
FE		1	1	1
Hybrid			1	1
Hybrid^0				1
Hybrid^R				

Note. JEP: LMC = Journal of Experimental Psychology: Learning, Memory, and Cognition; JPSP = Journal of Personality and Social Psychology; PSCI: cog. = Psychological Science cognitive psychology; PSCI: soc. = Psychological Science social psychology

hybrid^R were perfect (H = 1), implying that a hybrid method only rejected the null hypothesis if fixed-effect meta-analysis did as well. Associations of the replication with hybrid, hybrid^0, and hybrid^R were 0.519, 0.519, and 0.603, respectively.

To conclude, when correcting for the statistical significance of the original study, estimators of the hybrid methods on average provided smaller effect size estimates than the fixed-effect meta-analytic estimator. Uncertainty of the hybrid estimators (width of the confidence interval) was invariably larger than that of fixed-effect meta-analytic estimator, which together with its lower estimates explain the hybrids' lower proportion of rejection of the null hypothesis of no effect. If a hybrid method rejected the null hypothesis, this hypothesis was also rejected by fixed-effect meta-analysis, but not the other way around. This suggests that the testing procedures of the hybrid methods are primarily more conservative than the testing procedure of fixed-effect meta-analysis. Compared to the replication, the hybrid methods' estimators on average provided somewhat larger effect sizes, but higher uncertainty and a similar percentage reflecting how often the null hypothesis of no effect was rejected. The results of the hybrid methods were more in line with those of only the replication than fixed-effect meta-analysis or only the original study.

6.4 Discussion

One of the pillars of science is replication; does a finding withstand replication in similar circumstances, or can the results of a study generalized across different settings and people, and do the results persist over time? According to Popper (1959/2005), replications are the only way to convince ourselves that an effect really exists and is not a false positive. The replication issue is particularly relevant in psychology, which shows an unrealistically high rate of positive findings (e.g., Fanelli,

2012; Sterling et al., 1995). The RPP (Open Science Collaboration, 2015) replicated 100 studies in psychology and confirmed these unrealistic findings; less than 40% of original findings were statistically significant. This chapter examined several methods for estimating and testing effect size combining a statistically significant effect size of the original study and effect size of a replication. By approximating analytically the joint probability density function of original study and replication effect size we show that the estimator of fixed-effect meta-analysis yields overestimated effect size, particularly if the population effect size is zero or small, and yields a too high Type I error rate. We developed a new method, called hybrid, which takes into account that the expected value of the statistically significant original study is larger than the population effect size, and enables point and interval estimation, and hypothesis testing. The statistical properties of hybrid and two variants of hybrid are examined and compared to fixed-effect meta-analysis and to using only replication data. On the basis of this comparison, we formulated guidelines for when to use which method to estimate effect size. All methods were also applied to the data of the RPP.

The hybrid method is based on the statistical principle that the distribution of *p*-values at the population effect size has to be uniform. Since positive findings are overrepresented in the literature, the method computes probabilities at the population effects size for both the original study and replication where likely overestimation of the original study is taken into account. The hybrid method showed good statistical properties (i.e., Type I error rate equal to alpha level, coverage probabilities matching the nominal level, and median effect size estimate equal to the population effect size) when its performance was analytically approximated. However, hybrid's estimator is slightly negatively biased if the mean of the (conditional) probabilities was close to 1. This negative bias was also observed in another meta-analytic method (*p*-uniform) using conditional probabilities. To correct for this bias, we developed two alternative methods (hybrid^0 and hybrid^R) that do not suffer from these highly negative estimates and have the same desirable statistical properties as the hybrid method. We recommend using the hybrid^R method among the three hybrid variants because its estimator is least biased, its RMSE is lower than hybrid's estimator, and hybrid^R's testing procedure has the most statistical power.

We formulated guidelines (see Table 6.3) to help researchers select the most appropriate method when combining an original study and replication. The first two guidelines suppose that a researcher does not have knowledge about the magnitude of the population effect size. In this case, we advise to use only the replication data if the original study's sample size is smaller than of the replication and to use the hybrid^R method if the sample size in the original study is larger or equal to the sample size of the replication. The hybrid^R method is also recommended to be used if the effect size in the population is expected to be either absent or small. Fixed-effect meta-analysis has the best statistical properties and is advised to be used if the expected population effect size is medium or large. To prevent researchers selecting a method based on its

results ('*p*-hacking'), we recommend selecting the method using our guidelines *before* analyzing the data.

Applying the hybrid methods to studies of RPP largely confirmed the results of only the replication study as reported by the Open Science Collaboration (2015). Average effect size and proportion of statistically significant effects was considerably larger for fixed-effect meta-analysis than for the other methods, providing indirect evidence of overestimation by fixed-effect meta-analysis. The results suggest that many findings published in the three included psychology journals have smaller effect sizes than reported and that some effects may even be absent. In addition, uncertainty of the estimates of the hybrid methods was generally high, meaning that discarding the original studies generally made effect size estimates more precise. We draw two general conclusions from our re-analysis of the RPP. First, estimates of only the replication and the hybrid methods are generally more accurate than both the original study and fixed-effect meta-analysis which tend to overestimate because of publication bias. Second, most estimates of the replication and the hybrid methods were too uncertain to draw strong conclusions on the magnitude of the effect size, i.e. sample sizes were too small to provide precise estimates. These two conclusions are in line with a Bayesian re-analysis of the RPP (Etz & Vandekerckhove, 2016).

The effect size estimates of the hybrid methods can also be used to estimate the power of the original study, based on hybrid's effect size estimate. This alternative calculation of so-called 'observed power' has the advantage that it is based on evidence of both the original study and the replication. The observed power of the original study may be interpreted as an index of the statistical quality of the original study, with values of .8 or higher signaling good quality (Cohen, 1990). However, we recommend caution in interpreting this alternative observed value, because it is imprecise particularly when both studies' sample sizes is low. To work out an example of this approach we applied it to the example in the introduction and Table 6.1. Following our guidelines in Table 6.3, we use the replication's effect size estimate equal to d=0.164 in combination with the original sample size equal to 80 for our power analysis. Entering these numbers in G*Power 3.1.9.2 (Faul, Erdfelder, Lang, & Buchner, 2007) yields a power equal to .18 of a one-tailed t-test (α=.05), suggesting that the original study had low statistical quality.

We developed R code[30] and a web-based application that enables researchers to apply the hybrid methods, as well as fixed-effect meta-analysis, to their own data (https://rvanaert.shinyapps.io/hybrid). While the hybrid methods can in principle be applied to any effect size measure, the software can currently be applied to three different effect size measures: one-sample mean, two-independent means, and correlation coefficients. For the effect size measures one-sample mean and two-

[30] An R function (called hybrid) for applying the different hybrid methods is included in the "puniform" package and can be installed by running the following code:
devtools::install_github("RobbievanAert/puniform")

independent means, Hedges' *g* effect sizes and their sampling variances are computed by the software before the methods are applied. This is the same procedure as illustrated when we applied the hybrid method to the example in the introduction. If correlation coefficients are used as effect size measure (as was the case in the application to the RPP data), the software first transforms the correlation coefficients to Fisher transformed correlation coefficients and computes the corresponding sampling variances. The Fisher transformed correlation coefficients and their sampling variances are then used for applying the methods, where the output provides the back-transformed correlation coefficients. Figure 6.6 shows a screenshot of the application after it was applied to the example presented in the introduction. Data for one-sample mean and two-independent means can be entered via either group means, sample sizes, and standard deviations or *t*-values and sample sizes. Users should also specify the alpha level and the direction of the hypothesis test which was used in the primary studies. The right-hand side of the web application presents the results (estimate, test statistic [*t*-value, *z*-value, or *x*], two-tailed *p*-value, and confidence interval) of the hybrid, hybrid^{0}, hybrid^{R}, fixed-effect meta-analysis, and the replication. The application includes a link to a short manual on how to use the application.

The hybrid methods assume researchers select statistically significant original findings to replicate. The expected value of a statistically significant finding exceeds the population effect size, irrespective of publication bias, and the hybrid method corrects for this overestimation. A critical question is how to estimate effect size if a researcher wants to replicate a statistically significant original study, but this study was *not* selected because of its significance. How to proceed in this case does depend on the existence of publication bias. If no publication bias exists in the study's field, fixed-effect meta-analysis is the optimal method to combine an original study and replication assuming that both estimate the same underlying true effect size. However, if strong publication bias exists, as seems to be the case in psychology, the literature rather than the researcher has already mainly selected the statistically significant findings. Thus, even though researchers did not select a study to replicate based on it being statistically significant, we recommend applying the presented guidelines (Table 6.3) because the literature mainly presents significant and overestimated effect size estimates.

Another assumption of the hybrid methods is that a common effect (i.e., fixed effect) is underlying the original study and replication. This assumption can be violated if there are substantial discrepancies between the original study and replication. These discrepancies may be caused by differences in methodology that are used in both studies (Gilbert, King, Pettigrew, & Wilson, 2016) discrepancies may also be caused by findings that can only be replicated under specific conditions and that do

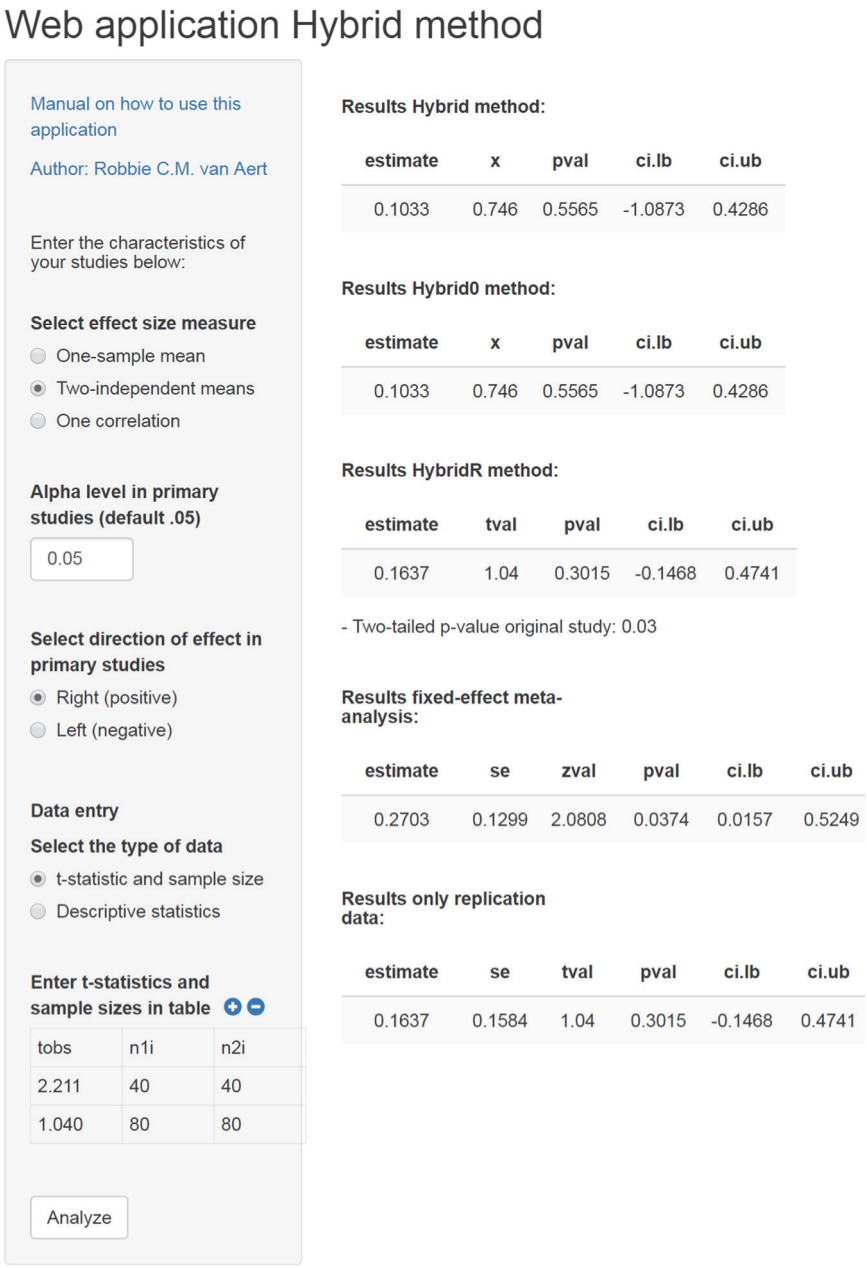

Figure 6.6. Screenshot of the web-based application showing the results of applying the hybrid variants, fixed-effect meta-analysis, and the replication to the exemplary data presented in the introduction.

not generalize to different settings, subjects, or do not persist over time (Amir & Sharon, 1990; Henrich, Heine, & Norenzayan, 2010; Klein et al., 2014; S. Schmidt, 2009). Although the assumption of homogeneity in effect sizes can be tested in a meta-analysis, it is difficult to draw reliable inferences in case of only two studies. The *Q*-test which is used for testing homogeneity lacks statistical power if the number of studies in a meta-analysis is small (e.g., Borenstein et al., 2009; Jackson, 2006).

We will extend the hybrid methods such that they can include more than one original study and one replication. These extended hybrid methods can be applied if, for instance, a researcher replicates a finding on which multiple original studies or a meta-analysis has already been published. These variants would use only the statistically significant findings of the original studies or meta-analysis, as does *p*-uniform (van Aert et al., 2016; van Assen et al., 2015), and combine these with the replication finding(s) to estimate common effect size.

An important implication of our analysis is that it may be optimal to discard information of the original study when estimating effect size. This is the case when being uncertain about population effect size and sample size in the replication is larger than in the original study, a situation that occurs very frequently. For instance, the sample size of 70 out of 100 replications in RPP is larger in the replication than in the original study. This implication may be generalized when multiple original studies and one replication are combined. Fixed-effect meta-analyses overestimate particularly if they incorporate more original studies with a relatively small sample size, and accuracy of estimation is better served by one or few large studies (Button et al., 2013; Gerber et al., 2001; Kraemer et al., 1998; Nuijten, Hartgerink, et al., 2015). We contend that extended hybrid methods, although they can correct for probable overestimation by original studies in the meta-analysis, their accuracy and precision is better served by more replication studies. Discarding all original studies and estimation by only one or a few large replication studies may even be the optimal choice (Nuijten et al., 2015). Omitting biased original studies from a meta-analysis is not a research waste since the effect size estimate will become more accurate.

The present study has several limitations and offer opportunities for future research. First, at present the hybrid method only allows for estimation based on one original and one replication study. We plan to extend the hybrid to incorporate multiple original and replication studies, and to examine its performance as a function of true effect size, publication bias, and the number of studies and their sample sizes. Second, *p*-hacking or questionable research practices distort the distribution of *p*-values and therefore also of conditional probabilities (Bruns & Ioannidis, 2016; Simonsohn et al., 2014a; Ulrich & Miller, 2015; van Aert et al., 2016; van Assen et al., 2015), which will bias effect size estimates of the hybrid methods. However, note that results of traditional meta-analytic methods are also distorted by *p*-hacking. Future research may examine to what extent the results of the hybrid methods become biased due to *p*-hacking. A third limitation is that the performance of hybrid methods

relative to other methods is dependent on the strength of the population effect, which is the object of the research. The guidelines we propose in Table 6.3 acknowledge this fact by advising the researcher what to do if the magnitude of the population effect size is uncertain. We must note, however, that the guidelines are formulated in the context of sample sizes presently used in psychological research. The guidelines lose their practical relevance if the sample size of the original study and replication allow for accurate effect size estimation in both studies. For instance, if original and replication sample sizes are 2,000 and 2,050, respectively, it would be naive to discard the original study and only use the replication for interpretation (Guideline 1a, Table 6.3). In that case, fixed-effect meta-analysis is the recommended method, because overestimation due to publication bias is very small at worst.

The unrealistic high rate of statistically significant results in the published psychological literature suggests that the literature is distorted with false positive results and overestimated effect sizes. Replication research and statistically combining these replications with the published research via meta-analytic techniques can be used to gather insight into the existence of effects. However, traditional meta-analytic techniques generally yield overestimated effect sizes. We developed hybrid meta-analytic methods and demonstrate their good statistical properties. We propose guidelines for conducting meta-analysis by combining the original study and replication, and provide a web-application (https://rvanaert.shinyapps.io/hybrid) that estimates and tests effect size of all methods described in this chapter. Applying the hybrid methods and our guidelines for meta-analyzing an original study and replication will give better insight into psychological phenomena by accurately estimating effect size.

6.5 Appendix A

Table 6.A1. Effect size estimate and standard deviation of this estimate in brackets for the estimators of fixed-effect meta-analysis, replication study and hybrid, hybrid^0, and hybrid^R method as a function of population effect size ρ and sample size of the original study (N_O). The sample size of the replication (N_R) is 783.

		N_R=783		
	ρ	N_O=31	N_O=55	N_O=96
FE	0	0.015 (0.034)	0.02 (0.033)	0.026 (0.032)
	0.1	0.112 (0.034)	0.115 (0.033)	0.116 (0.032)
	0.3	0.306 (0.031)	0.305 (0.031)	0.302 (0.03)
	0.5	0.501 (0.026)	0.5 (0.026)	0.5 (0.025)
Replication	0	0 (0.036)	0 (0.036)	0 (0.036)
	0.1	0.1 (0.035)	0.1 (0.035)	0.1 (0.035)
	0.3	0.3 (0.032)	0.3 (0.032)	0.3 (0.032)
	0.5	0.5 (0.027)	0.5 (0.027)	0.5 (0.027)
Hybrid	0	-0.001 (0.047)	-0.001 (0.046)	-0.001 (0.045)
	0.1	0.099 (0.047)	0.099 (0.045)	0.099 (0.044)
	0.3	0.299 (0.042)	0.299 (0.04)	0.299 (0.036)
	0.5	0.499 (0.033)	0.499 (0.031)	0.499 (0.028)
Hybrid^0	0	0.019 (0.027)	0.018 (0.026)	0.018 (0.025)
	0.1	0.099 (0.046)	0.099 (0.044)	0.099 (0.043)
	0.3	0.299 (0.042)	0.299 (0.04)	0.299 (0.036)
	0.5	0.499 (0.033)	0.499 (0.031)	0.499 (0.028)
Hybrid^R	0	0.013 (0.039)	0.013 (0.039)	0.012 (0.038)
	0.1	0.112 (0.038)	0.112 (0.038)	0.111 (0.036)
	0.3	0.309 (0.035)	0.306 (0.034)	0.303 (0.033)
	0.5	0.503 (0.03)	0.5 (0.03)	0.499 (0.028)

6.6 Appendix B

Table 6.B1. Data of the Reproducibility Project: Psychology and results of applying fixed-effect meta-analysis and the hybrid, hybrid^0, and hybrid^R method to these data. The first column lists the article from which a key effect was replicated. The next two columns show the correlation coefficient (r_o and r_r), sample size (N_O and N_R), and p-value from the original study and replication, respectively. The final three columns present the average effect size estimate, 95% confidence interval (CI), and p-value of fixed-effect meta-analysis (FE MA) and the hybrid and hybrid^R method. ^{0} behind the estimates of the hybrid method indicates that the hybrid^0 method would set the average effect size estimate to zero.

Study	r_o (N_O) [p-value]	r_r (N_R) [p-value]	FE MA (95% CI) [p-value]	Hybrid (95% CI) [p-value]	Hybrid^R (95% CI) [p-value]
Dodson, Darragh, and Williams (2008)	0.561 (39) [< .001]	-0.111 (33) [.543]	0.287 (0.055;0.491) [.016]	0.232 (-0.245;0.641) [.535]	0.232 (-0.245;0.641) [.535]
Ganor-Stern and Tzelgov (2008)	0.699 (30) [< .001]	0.781 (31) [< .001]	0.743 (0.6;0.84) [< .001]	0.743 (0.599;0.838) [< .001]	0.743 (0.599;0.838) [< .001]
Mirman and Magnuson (2008)	0.672 (23) [< .001]	0.466 (31) [.007]	0.561 (0.338;0.725) [< .001]	0.558 (0.318;0.755) [< .001]	0.558 (0.318;0.755) [< .001]
J. R. Schmidt and Besner (2008)	0.195 (96) [.028]	0.247 (243) [< .001]	0.233 (0.129;0.331) [< .001]	0.19 (-0.373;0.304) [.321]	0.247 (0.125;0.361) [< .001]
Oberauer (2008)	0.56 (33) [.001]	0.402 (21) [.071]	0.505 (0.266;0.685) [< .001]	0.482 (0.204;0.666) [.002]	0.482 (0.204;0.666) [.002]
Sahakyan, Delaney, and Waldum (2008)	0.224 (96) [.028]	0.019 (108) [.842]	0.117 (-0.022;0.251) [.099]	0.004 (-0.397;0.198) [.96]	0.019 (-0.17;0.208) [.842]
Bassok, Pedigo, and Oskarsson (2008)	0.364 (154) [< .001]	0.284 (50) [.045]	0.345 (0.217;0.462) [< .001]	0.335 (0.175;0.444) [.001]	0.335 (0.175;0.444) [.001]

Table 6.B1 Continued

Study	r_o (N_O) [*p*-value]	r_r (N_R) [*p*-value]	FE MA (95% CI) [*p*-value]	Hybrid (95% CI) [*p*-value]	Hybrid^R (95% CI) [*p*-value]
Yap, Balota, Tse, and Besner (2008)	0.378 (33) [.029]	0.38 (72) [.001]	0.379 (0.199;0.534) [< .001]	0.294 (-0.689;0.482) [.345]	0.38 (0.162;0.562) [.001]
Turk-Browne, Isola, Scholl, and Treat (2008)	0.738 (9) [.021]	0.704 (16) [.002]	0.715 (0.42;0.873) [< .001]	0.626 (-0.635;0.84) [.169]	0.626 (-0.635;0.84) [.169]
White (2008)	0.623 (38) [< .001]	0.481 (39) [.002]	0.555 (0.374;0.695) [< .001]	0.554 (0.362;0.701) [< .001]	0.554 (0.362;0.701) [< .001]
Farrell (2008)	0.517 (41) [< .001]	0.316 (41) [.044]	0.422 (0.221;0.588) [< .001]	0.408 (0.179;0.603) [.001]	0.408 (0.179;0.603) [.001]
Pacton and Perruchet (2008)	0.714 (22) [< .001]	0.682 (22) [< .001]	0.698 (0.497;0.828) [< .001]	0.696 (0.508;0.816) [< .001]	0.696 (0.508;0.816) [< .001]
Makovski, Sussman, and Jiang (2008)	0.551 (13) [.0499]	0.35 (19) [.144]	0.433 (0.079;0.69) [.018]	-0.312 (-1;0.505) [.865]^0	0.35 (-0.124;0.694) [.144]
Payne, Burkley, and Stokes (2008)	0.352 (69) [.003]	0.15 (178) [.045]	0.208 (0.084;0.325) [.001]	0.202 (0.067;0.419) [.006]	0.202 (0.067;0.419) [.006]
Cox et al. (2008)	0.225 (94) [.029]	-0.052 (194) [.469]	0.039 (-0.078;0.154) [.517]	-0.055 (-0.425; 0.169) [.439]^0	-0.052 (-0.192; 0.089) [.469]
Albarracín et al. (2008)	0.378 (36) [.022]	-0.03 (88) [.779]	0.089 (-0.091;0.263) [.332]	-0.013 (-0.373;0.36) [.894]^0	-0.013 (-0.373;0.36) [.894]

Table 6.B1 Continued

Study	r_o (N_O) [*p*-value]	r_r (N_R) [*p*-value]	FE MA (95% CI) [*p*-value]	Hybrid (95% CI) [*p*-value]	Hybrid^R (95% CI) [*p*-value]
Centerbar, Schnall, Clore, and Garvin (2008)	0.206 (133) [.017]	0.094 (113) [.323]	0.155 (0.03;0.275) [.015]	0.092 (-0.114;0.242) [.258]	0.092 (-0.114;0.242) [.258]
Amodio, Devine, and Harmon-Jones (2008)	0.377 (33) [.03]	0.077 (75) [.514]	0.169 (-0.023;0.35) [.084]	0.04 (-0.707;0.3) [.728]	0.077 (-0.153;0.298) [.514]
van Dijk, van Kleef, Steinel, and van Beest (2008)	0.379 (101) [< .001]	-0.042 (40) [.798]	0.271 (0.109;0.419) [.001]	0.211 (-0.166;0.442) [.363]	0.211 (-0.166;0.442) [.363]
Lemay and Clark (2008)	0.167 (184) [.023]	0.037 (280) [.541]	0.089 (-0.003;0.179) [.057]	0.033 (-0.183;0.163) [.536]	0.033 (-0.183;0.163) [.536]
Ersner-Hershfield, Mikels, Sullivan, and Carstensen (2008)	0.22 (110) [.021]	-0.005 (222) [.944]	0.07 (-0.038;0.177) [.205]	0.008 (-0.188;0.215) [.894]	0.008 (-0.188;0.215) [.894]
Correll (2008)	0.274 (70) [.021]	0.074 (147) [.375]	0.139 (0.005;0.268) [.042]	0.072 (-0.244;0.27) [.378]	0.072 (-0.244;0.27) [.378]
Exline, Baumeister, Zell, Kraft, and Witvliet (2008)	0.432 (43) [.003]	0.012 (133) [.894]	0.117 (-0.033;0.262) [.125]	0.111 (-0.07;0.508) [.266]	0.111 (-0.07;0.508) [.266]
Risen and Gilovich (2008)	0.186 (118) [.044]	0.003 (224) [.964]	0.066 (-0.041;0.172) [.224]	-0.065 (-0.979; 0.077) [.413]^0	0.003 (-0.128;0.134) [.964]
Stanovich and West (2008)	0.222 (375) [< .001]	0.073 (177) [.332]	0.175 (0.093;0.255) [< .001]	0.16 (0.016;0.26) [.028]	0.16 (0.016;0.26) [.028]

Table 6.B1 Continued

Study	r_o (N_O) [*p*-value]	r_r (N_R) [*p*-value]	FE MA (95% CI) [*p*-value]	Hybrid (95% CI) [*p*-value]	Hybrid^R (95% CI) [*p*-value]
Blankenship and Wegener (2008)	0.208 (259) [.001]	0.044 (249) [.485]	0.129 (0.042;0.213) [.004]	0.114 (-0.007;0.25) [.066]	0.114 (-0.007;0.25) [.066]
Shnabel and Nadler (2008)	0.268 (92) [.009]	-0.102 (139) [.234]	0.047 (-0.083;0.176) [.48]	-0.02 (-0.186;0.309) [.861]^0	-0.02 (-0.186;0.309) [.861]
Goff, Steele, and Davies (2008)	0.396 (53) [.003]	0.013 (49) [.929]	0.22 (0.024;0.4) [.028]	0.156 (-0.114;0.468) [.277]	0.156 (-0.114;0.468) [.277]
Murray, Derrick, Leder, and Holmes (2008)	0.317 (85) [.003]	-0.135 (70) [.266]	0.119 (-0.041;0.273) [.144]	0.037 (-0.228;0.379) [.856]	0.037 (-0.228;0.379) [.856]
McCrea (2008)	0.344 (28) [.036]	0.29 (61) [.012]	0.306 (0.101;0.487) [.004]	0.179 (-0.926;0.41) [.545]	0.29 (0.041;0.505) [.023]
Purdie-Vaughns, Steele, Davies, Ditlmann, and Crosby (2008)	0.378 (75) [.001]	-0.037 (1488) [.154]	-0.017 (-0.066; 0.033) [.506]	0.018 (-0.057;0.448) [.879]	0.018 (-0.057;0.448) [.879]
Dessalegn and Landau (2008)	0.382 (36) [.021]	-0.223 (47) [.133]	0.043 (-0.179;0.26) [.707]	-0.153 (-0.44;0.374) [.42]^0	-0.153 (-0.44;0.374) [.42]
Eitam, Hassin, and Schul (2008)	0.222 (86) [.039]	-0.105 (158) [.19]	0.010 (-0.116;0.136) [.874]	-0.146 (-0.889; 0.039) [.094]^0	-0.105 (-0.257; 0.052) [.19]
Farris, Treat, Viken, and McFall (2008)	0.554 (280) [< .001]	0.091 (144) [.278]	0.418 (0.335;0.494) [< .001]	0.385 (0.027;0.585) [.019]	0.385 (0.027;0.585) [.019]

Table 6.B1 Continued

Study	r_o (N_O) [*p*-value]	r_r (N_R) [*p*-value]	FE MA (95% CI) [*p*-value]	Hybrid (95% CI) [*p*-value]	Hybrid^R (95% CI) [*p*-value]
Janiszewski and Uy (2008)	0.333 (57) [.011]	0.226 (118) [.014]	0.261 (0.116;0.395) [.001]	0.226 (0;0.392) [.0501]	0.226 (0;0.392) [.0501]
McKinstry, Dale, and Spivey (2008)	0.701 (11) [.014]	0.75 (11) [.006]	0.727 (0.407;0.888) [< .001]	0.666 (-0.171;0.868) [.079]	0.666 (-0.171;0.868) [.079]
Armor, Massey, and Sackett (2008)	0.681 (126) [< .001]	0.764 (177) [< .001]	0.732 (0.675;0.78) [< .001]	0.728 (0.643;0.787) [< .001]	0.728 (0.643;0.787) [< .001]
Addis, Wong, and Schacter (2008)	0.571 (32) [< .001]	0.653 (32) [< .001]	0.613 (0.428;0.749) [< .001]	0.61 (0.409;0.742) [< .001]	0.61 (0.409;0.742) [< .001]
Nurmsoo and Bloom (2008)	0.502 (33) [.003]	-0.45 (10) [.199]	0.341 (0.033;0.59) [.031]	0.068 (-0.649;0.586) [.903]	0.068 (-0.649;0.586) [.903]
Vul and Pashler (2008)	0.288 (174) [< .001]	0.323 (141) [< .001]	0.303 (0.199;0.401) [< .001]	0.303 (0.204;0.394) [< .001]	0.303 (0.204;0.394) [< .001]
Masicampo and Baumeister (2008)	0.214 (113) [.023]	-0.049 (160) [.54]	0.061 (-0.059;0.179) [.322]	-0.032 (-0.237;0.2) [.661]^0	-0.032 (-0.237;0.2) [.661]
Hajcak and Foti (2008)	0.38 (31) [.017]	0.25 (43) [.053]	0.305 (0.077;0.503) [.009]	0.23 (-0.191;0.464) [.157]	0.23 (-0.191;0.464) [.157]
Alvarez and Oliva (2008)	0.722 (9) [.026]	0.923 (18) [< .001]	0.887 (0.754;0.951) [< .001]	0.847 (-0.865;0.948) [.261]	0.923 (0.801;0.971) [< .001]

Table 6.B1 Continued

Study	r_o (N_O) [*p*-value]	r_r (N_R) [*p*-value]	FE MA (95% CI) [*p*-value]	Hybrid (95% CI) [*p*-value]	Hybrid^R (95% CI) [*p*-value]
Lau, Kay, and Spencer (2008)	0.384 (36) [.02]	-0.034 (70) [.779]	0.11 (-0.085;0.297) [.268]	-0.003 (-0.309; 0.384) [.98][0]	-0.003 (-0.309; 0.384) [.98]
Winawer, Huk, and Boroditsky (2008)	0.685 (30) [< .001]	0.527 (27) [.004]	0.617 (0.418;0.759) [< .001]	0.613 (0.392;0.761) [< .001]	0.613 (0.392;0.761) [< .001]
Nairne, Pandeirada, and Thompson (2008)	0.446 (25) [.025]	0.423 (39) [.007]	0.432 (0.202;0.617) [< .001]	0.338 (-0.552;0.563) [.245]	0.338 (-0.552;0.563) [.245]
Larsen and McKibban (2008)	0.21 (117) [.023]	0.5 (236) [< .001]	0.413 (0.322;0.496) [< .001]	0.382 (-0.223;0.537) [.209]	0.382 (-0.223;0.537) [.209]
Vohs and Schooler (2008)	0.498 (30) [.004]	0.102 (58) [.446]	0.244 (0.032;0.434) [.024]	0.209 (-0.039;0.578) [.098]	0.209 (-0.039;0.578) [.098]
Halevy, Bornstein, and Sagiv (2008)	0.769 (78) [< .001]	0.653 (38) [< .001]	0.736 (0.638;0.811) [< .001]	0.726 (0.573;0.806) [< .001]	0.726 (0.573;0.806) [< .001]
Janssen, Alario, and Caramazza (2008)	0.65 (16) [.005]	0.497 (13) [.085]	0.588 (0.26;0.795) [.001]	0.529 (0.109;0.768) [.021]	0.529 (0.109;0.768) [.021]
Bressan and Stranieri (2008)	0.189 (196) [.008]	-0.03 (261) [.628]	0.064 (-0.028;0.155) [.171]	0.023 (-0.093;0.221) [.715]	0.023 (-0.093;0.221) [.715]
Bressan and Stranieri (2008)	0.189 (196) [.008]	0.018 (316) [.746]	0.084 (-0.003;0.17) [.058]	0.055 (-0.048;0.221) [.284]	0.055 (-0.048;0.221) [.284]

Table 6.B1 Continued

Study	r_o (N_O) [*p*-value]	r_r (N_R) [*p*-value]	FE MA (95% CI) [*p*-value]	Hybrid (95% CI) [*p*-value]	Hybrid^R (95% CI) [*p*-value]
Forti and Humphreys (2008)	0.723 (15) [.002]	0.208 (20) [.385]	0.463 (0.136;0.699) [.007]	0.424 (0;0.804) [.0501]	0.424 (0;0.804) [.0501]
Schnall, Benton, and Harvey (2008)	0.4 (43) [.007]	0.003 (126) [.975]	0.106 (-0.047;0.254) [.176]	0.078 (-0.1;0.463) [.403]	0.078 (-0.1;0.463) [.403]
Palmer and Ghose (2008)	0.86 (9) [.002]	0.12 (9) [.768]	0.608 (0.139;0.854) [.014]	0.516 (-0.211;0.917) [.172]	0.516 (-0.211;0.917) [.172]
Heine, Buchtel, and Norenzayan (2008)	0.43 (70) [< .001]	0.11 (16) [.69]	0.383 (0.182;0.553) [< .001]	0.327 (-0.101;0.517) [.122]	0.327 (-0.101;0.517) [.122]
Moeller, Robinson, and Zabelina (2008)	0.31 (53) [.023]	-0.034 (72) [.778]	0.114 (-0.065;0.286) [.21]	-0.019 (-0.354; 0.287) [.847][0]	-0.019 (-0.354; 0.287) [.847]
Goschke and Dreisbach (2008)	0.375 (40) [.017]	0.411 (95) [< .001]	0.401 (0.247;0.535) [< .001]	0.358 (-0.16;0.504) [.11]	0.358 (-0.16;0.504) [.11]
Lobue and DeLoache (2008)	0.483 (46) [.001]	0.178 (46) [.239]	0.34 (0.141;0.512) [.001]	0.317 (0.055;0.564) [.017]	0.317 (0.055;0.564) [.017]
Estes, Verges, and Barsalou (2008)	0.595 (19) [.006]	0.254 (23) [.245]	0.421 (0.122;0.65) [.007]	0.348 (-0.017;0.678) [.06]	0.348 (-0.017;0.678) [.06]

Note. P-values for the original study (second column) and replication (third column) were two-tailed except for the studies by Beaman et al. (2008), Schmidt and Besner (2008), McCrea (2008), and Hajcak and Foti (2008). These studies reported one-tailed *p*-values. *P*-values for fixed-effect meta-analysis (FE MA), the hybrid and hybrid^R method were two-tailed.

CHAPTER 7

Bayesian evaluation of effect size after replicating an original study

Abstract

The vast majority of published results in the literature is statistically significant, which raises concerns about their reliability. The Reproducibility Project Psychology (RPP) and Experimental Economics Replication Project (EE-RP) both replicated a large number of published studies in psychology and economics. The original study and replication were statistically significant in 36.1% in RPP and 68.8% in EE-RP suggesting many null effects among the replicated studies. However, evidence in favor of the null hypothesis cannot be examined with null hypothesis significance testing. We developed a Bayesian meta-analysis method called *snapshot hybrid* that is easy to use and understand and quantifies the amount of evidence in favor of a zero, small, medium and large effect. The method computes posterior model probabilities for a zero, small, medium, and large effect and adjusts for publication bias by taking into account that the original study is statistically significant. We first analytically approximate the methods performance, and demonstrate the necessity to control for the original study's significance to enable the accumulation of evidence for a true zero effect. Then we applied the method to the data of RPP and EE-RP, showing that the underlying effect sizes of the included studies in EE-RP are generally larger than in RPP, but that the sample sizes of especially the included studies in RPP are often too small to draw definite conclusions about the true effect size. We also illustrate how *snapshot hybrid* can be used to determine the required sample size of the replication akin to power analysis in null hypothesis significance testing and present an easy to use web application (https://rvanaert.shinyapps.io/snapshot/) and R code for applying the method.

This chapter is published as van Aert, R. C. M., & van Assen, M. A. L. M. (2017). Bayesian evaluation of effect size after replicating an original study. PLoS ONE, 12(4), e0175302. doi:10.1371/journal.pone.0175302

Most findings published in the literature are statistically significant (Fanelli, 2010a, 2012; Sterling et al., 1995) and are subsequently interpreted as nonzero findings. However, when replicating these original published studies in conditions as similar as possible to the original studies (so-called direct replications), replications generally provide lower estimates of the effect size that often are not statistically significant and are interpreted as suggesting a null effect. For instance, in medicine findings of only 6 out of 53 (11.3%) landmark studies on the field of hematology and oncology were confirmed in replication studies (Begley & Ellis, 2012). In psychology, the Reproducibility Project Psychology (RPP; Open Science Collaboration, 2015) replicated 100 studies published in major journals in 2008. Of the 97 original findings reported as statistically significant, only 35 (36.1%) had a statistically significant effect in the replication, and 81 of 97 (83.5%) findings were stronger in the original study. In economics, the Experimental Economics Replication Project (EE-RP; Camerer et al., 2016) replicated 18 studies published in high-impact journals. Of 16 findings that were statistically significant in the original study, 11 (68.8%) were statistically significant in the replication, and 13 of 16 (81.3%) had a stronger effect in the original study.

Interpreting the results of the replicability projects as providing evidence of many true null effects among the originally published studies has received criticism (e.g., Gilbert et al., 2016; Maxwell et al., 2015). For instance, Maxwell et al. (2015) argue that, although the replication in RPP generally had higher statistical power than the original study, the power of the replication was still too low to consider as evidence in favor of the null hypothesis. Consequently, the statistically nonsignificant findings of many replications are also consistent with a true nonzero, albeit small effect.

Many researchers adhere to null hypothesis significance testing (NHST) when evaluating the results of replications, and conclude based on a nonsignificant replication that the original study does not replicate (S. F. Anderson & Maxwell, 2016). However, such a vote counting procedure has been largely criticized in the context of a meta-analysis (e.g., Borenstein et al., 2009) and comes along with three fundamental problems. First, one cannot obtain evidence in favor of the null hypothesis of a true zero effect with NHST (e.g., Wagenmakers, 2007). Second, NHST does not tell us the size of the effect. Third, not all available information about the underlying effect is used in NHST because evidence obtained in the original study is ignored. What we need are methods providing evidence on the common true effect underlying both the original study and replications.

The method that immediately comes to mind when the goal is to estimate effect size based on several studies is meta-analysis. Two different traditional meta-analytic models can generally be distinguished: fixed-effect and random-effects model. Fixed-effect meta-analysis assumes that one common true effect underlies all observed effect sizes, whereas random-effects meta-analysis assumes observed effect

sizes arise from a (normal) distribution of true effect sizes (Borenstein et al., 2010). When one study is a direct replication of another, fixed-effect rather than random-effects meta-analysis seems to be the most appropriate method because the two studies are very similar. A small amount of heterogeneity in true effect size, however, may be possible since there could be minor discrepancies in for instance the studied population or experimental design as was sometimes the case in RPP. Publication bias is universally recognized as a major threat to the validity of meta-analyses, leading to overestimation of effect size (e.g., Ioannidis, 2008b; Lane & Dunlap, 1978; van Assen et al., 2015). Publication bias is the suppression of statistically nonsignificant results from being published (Rothstein et al., 2005a). Evidence of publication bias and as a consequence overestimation of effect size is omnipresent (e.g., Fanelli, 2010a; Fanelli, 2012; Lane & Dunlap, 1978; Sterling et al., 1995), and is also obvious from the aforementioned results of the replicability projects; almost all original findings were statistically significant whereas the replication findings were not, and the large majority of original effect size estimates was larger than those in the replication. Hence, traditional meta-analysis will be biased as well and will not suffice. A meta-analysis method is needed that takes into account the statistical significance of the original study, thereby adjusting for publication bias.

This chapter develops and applies a Bayesian meta-analytic method, called *snapshot hybrid*, to evaluate the effect size underlying an original study and replication. A requirement for applying the method is that the effect size of the original study is statistically significant. This requirement hardly restricts the applicability of the proposed method since the vast majority of published studies contain statistically significant results (Fanelli, 2010a, 2012; Sterling et al., 1995) and replications are often conducted when statistical significance is observed.

The snapshot hybrid has many desirable properties. First, the method has few assumptions. It assumes both studies estimate the same true effect size and the effect size in the original study and replication is normally distributed. The second desirable property is that, as opposed to fixed-effect meta-analysis, our method adjusts for publication bias when evaluating the underlying true effect size by taking into account statistical significance of the original study. Third, it provides a very simple interpretation of the magnitude of the true effect size. Its main output is the posterior model probability of a zero, small, medium, and large effect (i.e., probability of a model after updating the prior model probability with the likelihood of the data). Consequently, as opposed to NHST, it also quantifies the evidence in favor of the null hypothesis, relative to a small, medium, and large hypothesized effect. One high posterior model probability suggests certainty about the magnitude of the true effect, whereas several substantial nonzero posterior model probabilities indicate that the magnitude of the effect is rather uncertain. Fourth, the method has great flexibility in dealing with different prior information. Although the method's default prior model probabilities are equal (i.e., zero, small, medium and large effect are equally likely),

using a simple formula one can recalculate the posterior model probabilities for other prior model probabilities, without having to run the analysis again.

The goal of this chapter is fourfold. First, we explain snapshot hybrid and examine its statistical properties. Second, we apply the method to the data of RPP and EE-RP to examine evidence in favor of zero, small, medium, and large true effects. Particularly, we verify if interpreting the statistically nonsignificant findings of replication studies in psychology as evidence for null effects is appropriate. Third, we describe, analogous to conducting a power analysis for determining the sample size in a frequentist framework, how the proposed method can be used to compute the required sample size for the replication in order to get a predefined posterior model probability of the true effect size being zero, small, medium, or large. This goal acknowledges that our method is not only relevant for evaluating and interpreting replicability of effects. Replicating other's research is often the starting point for new research, where the replication is the first study of a multi-study paper (Neuliep & Crandall, 1993). Fourth, we present a web application and R code allowing users to evaluate the common effect size of an original study and replication using snapshot hybrid.

The next section provides a hypothetical example of an original study and replication by Maxwell et al. (2015) and illustrates the problem of evaluating the studies' underlying true effect size. The subsequent section explains snapshot hybrid, and is illustrated by applying it to the example of Maxwell et al. (2015). Then, the statistical properties of our method are examined analytically. Subsequently, the method was applied to the results of RPP and EE-RP. How the required sample size of the replication can be determined with the proposed method in order to achieve a predefined posterior model probability for a hypothesized effect size (zero, small, medium, or large) is discussed next. Then, the computer program is described to determine this required sample size, followed by a conclusion and discussion section.

7.1 Methods related to snapshot hybrid

The proposed snapshot hybrid method is related to several other methods. The meta-analysis methods *p*-uniform (van Aert, Wicherts, et al., 2016; van Assen et al., 2015) and *p*-curve (Simonsohn et al., 2014a) also take statistical significance of studies' effect sizes into account in order to correct the meta-analytic effect size estimate for publication bias. The effect size estimate of *p*-uniform and *p*-curve is equal to the effect size where the statistically significant *p*-values conditional on being statistically significant are uniformly distributed. Both methods have been shown to provide accurate estimates of the underlying true effect size in case of publication bias, but only if the amount of heterogeneity in studies' true effect size is modest (McShane et al., 2016; Simonsohn et al., 2014a; van Aert, Wicherts, et al., 2016; van Assen et al., 2015). We also wrote a paper where we use frequentist statistics to evaluate the common effect size underlying an original study and a replication, taking

into account the statistical significance of the original study (van Aert & van Assen, 2016). This method estimates effect size, provides a confidence interval, and enables testing of the common effect. Advantages of the Bayesian method presented here are that interpretation of its results is more straightforward, and evidence in favor of the null hypothesis is quantified.

Another related paper is a Bayesian re-analysis of the results of RPP (Etz & Vandekerckhove, 2016). In this chapter, Bayes factors were computed for each original study and replication separately, comparing the null hypothesis of no effect with an alternative hypothesis suggesting that the effect is nonzero. For the original studies, publication bias was taken into account when computing Bayes factors by using Bayesian model averaging over four different publication bias models. The most important differences of our Bayesian method and their re-analyses, which we interpret as advantages of our methodology, are: (i) they do not evaluate the underlying effect size, but test hypotheses for original study and replication separately, (ii) they make strong(er) assumptions on publication bias, using Bayesian model averaging over four different models of publication bias, (iii) their methodology lacks flexibility with dealing with different prior information (i.e. another prior requires rerunning the analysis), and (iv) they did not provide software to run the analysis. Both Etz and Vandekerckhove (2016) and van Aert and van Assen (2016) conclude that for many RPP findings no strong conclusions can be drawn on the magnitude of the underlying true effect size.

7.2 Example by Maxwell et al. (2015)

We will illustrate snapshot hybrid using a hypothetical example provided by Maxwell et al. (2015) with a statistically significant original study and nonsignificant replication. They use their example to illustrate that so-called failures to replicate in psychology may be the result of low statistical power in single replication studies. This example was selected because it reflects a common situation in practice. For instance, 62% of the 100 replicated studies in RPP (Open Science Collaboration, 2015) and 39% of the 18 replicated studies in EE-RP (Camerer et al., 2016) did not have a statistically significant effect, as opposed to the effect in the original study. Hence, researchers often face the question what to conclude with respect to the magnitude of the true effect size based on a statistically significant original effect and a nonsignificant replication effect. Does an effect exist? And if an effect exists, how large is it?

The example employs a balanced two-independent groups design. The original study, with 40 participants per group, resulted in Cohen's d=0.5 and $t(78)$=2.24 (two-tailed p-value=.028), which is a statistically significant effect if tested with α=.05. A power analysis was used by Maxwell et al. (2015) to determine the required sample size in the replication to achieve a statistical power of .9 using a two-tailed test, with an expected effect size equal to the effect size observed in the original study. The power analysis revealed that 86 participants per group were required. The

observed effect size in the replication was Cohen's *d*=0.23 with *t*(170)=1.50 (two-tailed *p*-value=.135), which is not statistically significant if tested with α=.05.

In our analyses, like in RPP and EE-RP, we transform effect sizes to correlation coefficients. Correlation coefficients are bounded between -1 and 1, and easy to interpret. Transforming original and replication effect sizes to correlations using $r_o = \frac{d}{\sqrt{d+4}}$ (e.g., Borenstein et al., 2009, p. 48) yields r_o = 0.243 and r_r = 0.114 for original and replication effect size, respectively. Testing individual correlations as well as combining correlations in a meta-analysis is often done using Fisher-transformed correlation coefficients (Borenstein et al., 2009) since these follow a normal distribution with variance $1/(N-3)$ with *N* being the total sample size (Fisher, 1921). The Fisher-transformed correlations (θ) are $\hat{\theta}_o$ =0.247 and $\hat{\theta}_r$ = 0.115 with standard errors .114 and .0769, respectively. Statistically combining the two effects by means of fixed-effect meta-analysis yields $\hat{\theta}$ =0.156 with standard error 0.0638, which is statistically significant (two-tailed *p*=.0142), suggesting a positive effect. Transforming the results of the meta-analysis to correlation coefficients yields the effect size estimate of 0.155 (95% confidence interval; 0.031 to 0.274). Although fixed-effect meta-analysis suggests a positive effect size, it should be interpreted with caution because of the generally overestimated effect size in the original study due to publication bias.

7.3 Snapshot Bayesian hybrid meta-analysis

The snapshot Bayesian hybrid meta-analysis method, *snapshot hybrid* for short, is a *meta-analysis* method because it combines both the original and replication effect size to evaluate the common true effect size. It is a *hybrid* method because it only takes the statistical significance of the original study into account, whereas it considers evidence of the replication study as unbiased. The method is *Bayesian* because it yields posterior model probabilities of the common true effect size. Finally, it is called *snapshot* because only four snapshots or slices of the posterior distribution of effect size are considered, i.e. snapshots/slices at hypothesized effect sizes equal to zero (ρ=0), and small (ρ=0.1), medium (ρ=0.3), and large (ρ=0.5) correlations (Cohen, 1988). We selected these four hypothesized effect sizes, because applied researchers are used to this categorization of effect size. Moreover, point hypotheses enable recalculating the posterior model probabilities for other than uniform encompassing prior distributions (i.e., prior model probabilities derived from other prior distributions than a uniform distribution that results in equal probabilities for the hypothesized effect sizes) as we will show later.

Two assumptions are underlying snapshot hybrid. First, the same effect (i.e., fixed effect) has to be underlying the original study and replication. This assumption seems to be reasonable if the replication is exact although small amounts of

heterogeneity may arise if there are minor discrepancies in studied population or experimental design. Exact replications are often conducted as the first study of a multi-study paper (Neuliep & Crandall, 1993). Second, effect size in the original study and replication are assumed to be normally distributed, which is a common assumption in meta-analysis (Raudenbush, 2009). Furthermore, the original study is required to be statistically significant. This requirement hardly restricts the range of application of the method because most studies in the social sciences contain statistically significant results, particularly in psychology with percentages of about 95% (e.g., Fanelli, 2012; Sterling et al., 1995) or even 97%, as in the RPP (Open Science Collaboration, 2015) and also 89% in the EE-RP. Note that, even if publication bias was absent in science, snapshot hybrid *should* be used if a researcher chooses to replicate an original study because of its statistical significance. It is precisely this selection that biases methods that do not correct for statistical significance, similar to how selecting only ill people for treatment or high scoring individuals on an aptitude test results in regression to the mean when re-tested.

The snapshot hybrid consists of three steps. First, the likelihood of the effect sizes of the original study and replication is calculated conditional on four hypothesized effect sizes (zero, small, medium, and large). Second, the posterior model probabilities of these four effect sizes are calculated using the likelihoods of step 1 and assuming equal prior model probabilities. Equal prior model probabilities are selected by default, because this refers to an uninformative prior distribution for the encompassing model. Third, when desired, the posterior model probabilities can be recalculated for other than equal prior model probabilities. We will explain and illustrate each step by applying the method to the example of Maxwell et al. (2015).

In the first step, the combined likelihood of the effect size of the original study ($\hat{\theta}_o$) and replication ($\hat{\theta}_r$) for each hypothesized effect size (θ) is obtained by multiplying the densities of the observed effect sizes:

$$L(\theta) = f(\hat{\theta}_o, \hat{\theta}_r \mid \theta) = f_o(\hat{\theta}_o \mid \theta) \times f_r(\hat{\theta}_r \mid \theta) \quad (1)$$

Note that densities and likelihood are based on Fisher-transformed correlations and hypothesized effect sizes. Figure 7.1a shows the probability density functions and densities of the observed effect size of the replication ($\hat{\theta}_r$ = .115). The four density functions follow a normal distribution with means θ_0 = 0 (red distribution), θ_S =0.1 (blue distribution), θ_M =0.31 (yellow distribution), θ_L =0.549 (green distribution), and standard deviation $\hat{\sigma}_r$ = .0769. The densities or heights at $\hat{\theta}_r$ = .115 (see vertical dashed line) are 1.705, 5.096, 0.210, 0, for a zero, small, medium, large true effect, respectively.

Figure 7.1b shows the four density functions and densities of the observed effect size in the original study ($\hat{\theta}_o$ =.247). Colors red, blue, yellow, and green refer

again to distributions of a zero (θ_0), small (θ_S), medium (θ_M), and large (θ_L) hypothesized effect, respectively. The density functions take statistical significance of the original finding into account by computing the density of $\hat{\theta}_o$ conditional on the study being statistically significant, i.e. by *truncating* the densities at the critical value of the Fisher-transformed correlation (θ_{cv}). A two-tailed test with α=.05 is assumed reflecting common practice in social science research where two-tailed tests are conducted and only the results in one direction get published. The truncated densities are calculated as

$$f_o(\hat{\theta}_o|\theta) = \frac{\phi\left(\frac{\hat{\theta}_o - \theta}{\hat{\sigma}_o}\right)}{1 - \Phi\left(\frac{\theta_{cv} - \theta}{\hat{\sigma}_o}\right)} \quad (2)$$

with ϕ and Φ being the standard normal density and cumulative distribution function, respectively. The denominator of (2) also represents the power of the test of no effect if the hypothesized effect size is equal to θ. Note how the conditional density f_o in Figure 7.1b of a just significant correlation increases when the hypothesized correlation decreases; the conditional density f_o is virtually identical to the unconditional density for large hypothesized effect size (because statistical power is close to 1), whereas the conditional density is 40 times larger (i.e., $1/(\alpha/2)$) than the unconditional density function for θ =0. The densities at $\hat{\theta}_o$ =.247 are 13.252, 10.852, 3.894, 0.105 for a zero, small, medium, large hypothesized effect, respectively. Note that after taking the statistical significance of the original finding into account, the density is highest for θ_0 and θ_S and substantially lower for θ_M, and θ_L. Hence, it is less likely that $\hat{\theta}_o$ stems from a population with a medium or large effect size than from a population with no effect or a small effect size. The first row of Table 7.1 presents the likelihoods of the observed effect sizes as a function of hypothesized effect size, after multiplying the studies' densities with Equation (1). The likelihood is largest for a small hypothesized effect in comparison with no, medium, and large hypothesized effect, suggesting that there is most probably a small true effect underlying the original study and replication.

Posterior model probabilities of one snapshot (ρ_S) relative to the others are calculated with the snapshot hybrid (second row and last three rows) and without correcting for statistical significance (snapshot naïve, third row). For snapshot hybrid, posterior model probabilities are calculated for four different sets of prior model probabilities; equal prior model probabilities (i.e., uniform encompassing model), prior model probabilities where the hypothesized zero effect gets a weight (p_0) 2 or 6 times higher than the other hypothesized effects, and prior model probabilities when a normal distribution with mean and variance equal to 0 and 1 is the encompassing model, respectively.

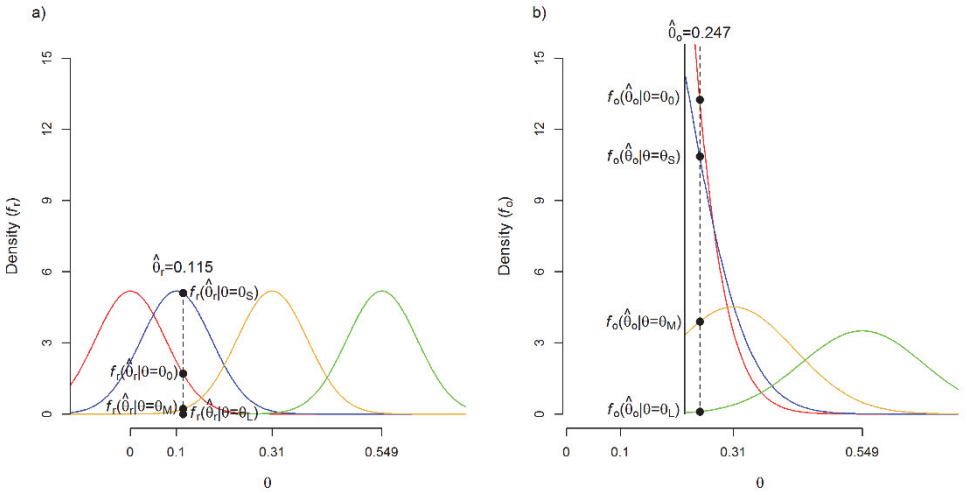

Figure 7.1. Probability density functions of the replication (panel a) and transformed original effect size when statistical significance is taken into account (panel b). The four hypothesized effect sizes (zero, small, medium, and large) are denoted by θ=0 (0, red distribution), θ=0.1 (S, blue distribution), θ=0.31 (M, yellow distribution), and θ=0.549 (L, green distribution). The dashed vertical line refers to the observed effect sizes in the hypothetical example of Maxwell et al. (2015) for the replication (panel a) and original study (panel b). The dots on the vertical dashed line refer to densities for no, small, medium, and large effect in the population.

In the second step, the posterior model probability π_x of each model with hypothesized effect size *x* is calculated using

$$\pi_x = \frac{L(\theta = x)}{L(\theta = \theta_0) + L(\theta = \theta_S) + L(\theta = \theta_M) + L(\theta = \theta_L)} \tag{3}$$

where *x* refers to either a zero (θ_0), small (θ_S), medium (θ_M), or large (θ_L) hypothesized effect size. This posterior model probability is a relative probability because it quantifies the amount of evidence for a model with a particular hypothesized effect size relative to the other included models. Since all likelihoods are weighed equally, implicitly equal prior model probabilities are assumed in Equation (3). The second row of Table 7.1 (method 'snapshot hybrid', uniform prior) presents the four posterior model probabilities of snapshot hybrid for the example. The posterior model probabilities indicate that after observing correlations r_o = 0.243 and r_r = 0.114, the evidence in favor of the null hypothesis slightly increased from .25 to .287, increased a lot (from .25 to .703) in favor of a small hypothesized effect size, and decreased a lot for a medium and large hypothesized effect size.

Table 7.1. Likelihoods and posterior model probabilities for zero, small, medium, and large hypothesized correlations for the example of Maxwell et al. (2015).

	Prior	Method	Zero (ρ_S=0)	Small (ρ_S=0.1)	Medium (ρ_S=0.3)	Large (ρ_S=0.5)
Likelihood			22.594	55.304	0.819	0
Posterior model probabilities	Uniform	Snapshot hybrid	.287	.703	.010	0
		Snapshot naïve	.063	.866	.071	0
	p_0 = 2	Snapshot hybrid	.446	.546	.008	0
	p_0 = 6	Snapshot hybrid	.707	.289	.004	0
	$N(0,1)$	Snapshot hybrid	.288	.702	.010	0

For the sake of comparison, we also calculated the posterior model probabilities using a method we call *snapshot naïve* because it incorrectly does not take the statistical significance of the original finding into account (i.e., without truncating the density at θ_{cv}). Its results are presented in the third row of Table 7.1 (method 'snapshot naïve', uniform encompassing prior distribution). These uncorrected posterior model probabilities provide stronger evidence in favor of a small effect relative to a zero, medium, and large effect, although posterior model probabilities for both a zero and medium hypothesized effect are still larger than zero (.06). Comparing the results of applying snapshot hybrid to those of snapshot naïve to the example shows that snapshot hybrid assigns larger posterior model probabilities to zero hypothesized effect size than snapshot naïve. This always holds. More generally, the snapshot naïve first-order stochastically dominates snapshot hybrid, i.e., snapshot naïve's cumulative posterior model probabilities exceed those of snapshot hybrid. The evaluation of snapshot hybrid may even suggest that the true effect size is smaller than the estimates of *both* the original study and replication. The latter typically occurs when the original effect size is just statistically significant (i.e., has a *p*-value just below .05) and the replication effect size has the same sign as the original effect.

Finally, in the third step the posterior model probabilities of a hypothesized effect size relative to the other hypothesized effect sizes may be recalculated using other than equal model probabilities. The posterior model probability π_x^* for hypothesized effect size *x* can be recalculated using

$$\pi_x^* = \frac{p_x \pi_x}{p_0 \pi_0 + p_S \pi_S + p_M \pi_M + p_L \pi_L}, \quad (4)$$

with prior model probabilities or weights p, and posterior model probabilities π calculated with Equation (3) assuming equal uniform prior probabilities. Note that $\pi_x^* = \pi_x$ for equal prior model probabilities. The values of p_x, with x referring to no (0), small (S), medium (M), or large hypothesized effect (L), can simply be derived from the prior density function of the researcher.

Simple and conservative prior model probabilities are to assign, for instance, a two or even six times higher prior model probability to a zero hypothesized effect than to any of the other hypothesized effects. Note that other prior model probabilities can also be used, and that these probabilities can also be specified for other hypothesized effect sizes than zero. Substituting $p_0 = 2$ and $p_0 = 6$ (and p_S, p_M, and p_L all equal to 1) and the posterior model probabilities presented in row "uniform snapshot hybrid" of Table 7.1, yields the recalculated posterior model probabilities presented in the two subsequent rows of Table 7.1. Naturally, more conservative prior model probabilities yield stronger evidence in favor of a hypothesized zero effect, with posterior model probabilities increasing from .287 (uniform prior) to .707 ($p_0 = 6$).

The posterior model probabilities can also be recalculated when a continuous prior is specified for the encompassing model, for instance a normal distribution with mean and variance equal to 0 and 1, respectively, denoted by $N(0,1)$ in Table 7.1. This normal prior yields prior model probabilities at θ=0, θ=0.1, θ=0.31, θ=0.549 of p_0=0.263, p_S=0.261, p_M=0.250, p_L=0.226, which are close to the equal prior model probabilities. This yields the recalculated posterior model probabilities in the last row of Table 7.1, again showing that assigning higher prior model probability to a hypothesized zero effect results in stronger evidence in favor of the null hypothesis. To sum up, the posterior model probabilities can be recalculated without doing the Bayesian analysis again, by applying Equation (4) using other prior model probabilities. Second, the example demonstrates that the prior model probabilities can have substantial effects on the posterior model probabilities, particularly if there is no (very) strong evidence for a hypothesized effect size.

7.4 Analytical evaluation of statistical properties

We evaluated the statistical properties of snapshot hybrid by comparing it to snapshot naïve. This comparison demonstrates that effect size evaluation often suggests larger effect sizes than the true effect size if statistical significance of the original study is not taken into account. Statistical properties of the methods were evaluated with the correlation coefficient as the effect size measure of interest. However, both methods can also be applied to other effect size measures (e.g., standardized mean differences).

7.4.1 Method

We analytically approximated the statistical properties of both snapshot hybrid and snapshot naïve using numerical integration of the joint probability density function (pdf) of the statistical significant original effect size and effect size in the replication. This joint pdf is a function of the true effect size and both effect sizes' standard error. The joint pdf of the statistically significant observed original effect size and the effect size of the replication was approximated by creating an equally spaced grid of 5,000 x 5,000 values. The pdf of the statistically significant observed original effect sizes was approximated by first selecting 5,000 equally spaced cumulative probabilities given that the effects sizes that accompanied these probabilities were statistically significant. A one-tailed Fisher-*z* test with α=.025 was used to determine the critical value for observing a statistically significant effect size in the original study because this corresponds to a common practice in social science research where two-tailed hypothesis tests are conducted and only the results in the predicted direction are reported. For instance, under the null hypothesis this means that the cumulative probabilities range from $1-0.025+\frac{(1\times.025)}{5{,}001}=.975005$ to $1-0.025+\frac{(5{,}000\times.025)}{5{,}001}=.999995$. All these cumulative probabilities were then transformed to Fisher-transformed correlation coefficients given a true effect size and standard error to approximate the pdf of the original effect size. The pdf of the replication's observed effect size given a true effect size and standard error was created in a similar way as the pdf of the original study's observed effect size, but there was no requirement for the effect size in the replication to be statistically significant. Hence, 5,000 equally spaced cumulative probabilities ranging from $\frac{1}{5{,}001}=.00019996$ to $\frac{5{,}000}{5{,}001}=.9998$ were selected and the pdf of the observed effect size in the replication was obtained by transforming these probabilities to Fisher-transformed correlation coefficients. Combining the marginal pdfs of the observed effect size in the original study and replication resulted in an approximation of the joint pdf consisting of 25,000,000 different combinations of effect sizes that was used for evaluating the statistical properties of snapshot hybrid and snapshot naïve. Both methods were applied to each combination of effect size in the original study and replication.

In order to examine the performance of the methods under different conditions, joint pdfs were created by varying two factors: total sample size for the original study and replication (*N*) and true effect size (ρ). Six different sample sizes were selected (*N* = 31; 55; 96; 300; 1,000; 10,000) and were imposed to be equal in the original study and replication. Sample sizes of 31, 55, and 96 refer to the first quartile, medium, and third quartile of the observed sample sizes of the original study

in RPP (Open Science Collaboration, 2015). Larger sample sizes were also included for two reasons. First, large sample sizes enable us to examine large sample properties of our method, such as convergence of the methods to the correct hypothesized effect size. Second, bias of snapshot naïve can be examined with large sample sizes, because bias is expected to disappear for very large sample size whenever true effect size exceeds zero. For the true effect size, we selected ρ=0 (no effect), 0.1 (small effect), 0.3 (medium effect), 0.5 (large effect), which correspond to the methods' snapshots or hypothesized effect sizes. Our analysis used equal prior model probabilities, assigning probabilities of .25 to each hypothesized effect size.

Posterior model probabilities of snapshot hybrid and snapshot naïve at the four hypothesized effect sizes were computed using Equation (3) for each of the 25,000,000 different combinations of effect sizes. Performances of both methods was then evaluated with respect to three outcomes. The first outcome was the expected value of the posterior model probability for each hypothesized effect size. Second, we calculated the proportion that the posterior model probability of a particular hypothesized effect size relative to the other hypothesized effect sizes was larger than .25, which amounts to the probability that evidence in favor of the true hypothesis increases after observing the data. The third outcome was the proportion that the posterior model probability of a particular hypothesized effect size relative to the other hypothesized effect sizes was larger than .75. This proportion corresponds to a Bayes Factor of 3 when comparing a particular hypothesized effect size to the other hypothesized effect sizes. Since a Bayes Factor exceeding 3 is interpreted as *positive* evidence (e.g., Kass & Raftery, 1995), we interpret posterior model probabilities of .75 or more as positive evidence in favor of that hypothesized effect size. Note that selecting a posterior model probability of 0.75 (and Bayes Factor of 3) is a subjective choice, and that selecting other posterior model probabilities (and Bayes Factors) for the analyses was also possible. In our analyses, we expect all three outcomes to increase in sample size for snapshot hybrid, but not always for the biased snapshot naïve method.

Computations were conducted in the statistical software R and the parallel package was used for parallelizing the computations (R Core Team, 2017). Computer code for the computations is available at https://osf.io/xrn8k/.

7.4.2 Results on statistical properties

Expected value of the posterior model probability Table 7.2 presents the expected values of the posterior model probabilities of snapshot hybrid and snapshot naïve for four different snapshots (ρ_S). The posterior model probabilities are presented for different sample sizes (N) and true effect sizes (ρ). Results for sample sizes per group equal to 10,000 are not shown since expected posterior model probabilities of both methods are always equal to 1 for the correct snapshot and to 0 for the incorrect snapshot. The bold values in the columns for snapshot hybrid and

snapshot naïve indicate the posterior model probability for that particular snapshot that matches the true effect size. Hence, the bold values in these columns should be higher than the posterior model probabilities for the other snapshots. The final column shows the expected value of the estimate of traditional fixed-effect meta-analysis.

Expected values of the posterior model probabilities of snapshot hybrid at the correct snapshot (e.g., ρ_S=0 if ρ=0 and ρ_S=0.1 if ρ=0.1) increase as the sample size increases, as they should (bold values in first four columns in Table 7.2). Expected values of the posterior model probabilities are close to .75 for ρ=0, ρ=0.1, and ρ=0.3 at n_i=300, and at N=55 for ρ=0.5. Snapshot hybrid has difficulties distinguishing whether an effect is absent (ρ=0) or small (ρ=0.1) for N < 1,000 because the expected values of the posterior model probabilities at ρ_S=0 and ρ_S=0.1 are close to each other for both effect sizes. Even if N=1,000, the expected value of the posterior model probability of ρ=0 at snapshot ρ_S=0.1 is .052 and the same holds for the expected value of the posterior model probability of ρ=0.1 at snapshot ρ_S=0.

The expected values of the posterior model probability of snapshot naïve also increase as the sample size increases (bold values in columns seven to ten in Table 7.2). However, the performance of snapshot naïve is worse than of snapshot hybrid for ρ=0 for all N, and for ρ=0.1 at N=31 and 55. Most important is that if ρ=0 the expected posterior model probability of snapshot naïve suggests a small effect (ρ=0.1) up to $N \leq 300$ (i.e., 600 observations in original study and replication combined). Evidence in favor of a small true effect size is even *increasing* in sample size until N=300, where the expected value of the posterior model probability of incorrect snapshot ρ=0.1 is .662 and larger than the .338 for the correct snapshot of zero true effect size. Even when N = 1,000, evidence in favor of a small effect hardly diminished; the expected posterior model probability (.242) is only little lower than .25. The performance of snapshot naïve is better than snapshot hybrid's performance for ρ=0.1 at $N \geq 96$, and for medium and large true effect size, i.e., expected posterior model probabilities at the correct snapshot are highest for snapshot naïve. Note, however, that for a medium true effect size evidence in favor of a strong effect (ρ_S = 0.5) also increases for small sample size (N = 31; expected posterior model probability increases from .25 to .417). All these results can be explained by two related consequences of correcting for the statistical significance of the original study.

The first consequence is that not correcting for statistical significance of the original study leads to overestimation of effect size. The last column of Table 7.2 presents the expected value of fixed-effect meta-analysis, and consequently, its bias. The bias decreases both in true effect size and sample size, and is most severe for ρ=0 and N=31 (0.215). Bias results in a higher expected value of the posterior model probability of 'incorrect' snapshots ($\rho_S \neq \rho$) for snapshot naïve. The fact that snapshot naïve performs relatively worse for a true small effect than for a medium and strong true effect is thus because overestimation is worse for lower true effect size,

particularly for small *N*.

The second consequence is that snapshot hybrid assigns a relatively higher 'weight' to the likelihood of the original effect under a zero true effect, compared to snapshot naïve. This is because, in contrast to snapshot naïve, the replication's likelihood is multiplied by the reciprocal of statistical power (which is the 'weight') under snapshot hybrid (see Equation (2) and Figure 7.1), and statistical power increases in true effect size. In the extreme case, for very large sample size (e.g., $N >$ 10,000), the *only* difference between snapshot hybrid and snapshot naïve is that the likelihood of the original effect under a zero hypothesized effect is multiplied by 40 under snapshot hybrid, because the likelihoods at other snapshots are multiplied by 1 under snapshot hybrid (statistical power then equals 1 at these snapshots). The relatively higher weight assigned to the likelihood of the original study's effect under the hypothesized zero effect explains why snapshot hybrid performs better than snapshot naïve if $\rho=0$, for all values of *N*. This relatively higher weight translates into higher posterior model probabilities for $\rho_S=0$ under snapshot hybrid than snapshot naïve. These higher posterior model probabilities for $\rho_S=0$ under snapshot hybrid also explain why snapshot naïve outperforms snapshot hybrid for nonzero true effect size in combination with large sample size. For nonzero true effect size and small sample size, however, snapshot hybrid outperforms snapshot naïve because then the adverse effect of overestimation in snapshot naïve is stronger than the higher weight of (incorrect) snapshot $\rho_S=0$ in snapshot hybrid.

To sum up, sample sizes of 300 for the original study and replication are needed to obtain expected posterior model probabilities with snapshot hybrid close to .75 or higher for a true effect size of $\rho=0$, $\rho=0.1$, and $\rho=0.3$, whereas a sample size of 55 for the original study and replication is sufficient for $\rho=0.5$. Hence, small sample sizes (sample size of about 50 per study) are sufficient to make correct decisions if true effect size is large, whereas for zero or small true effect size large sample sizes are required (sample size of at least 300 up to 1,000 per study). Not taking the statistical significance of the original study into account results in worse performance of snapshot naïve when true effect size is zero or small, or when sample sizes are small. Snapshot naïve outperforms snapshot hybrid, i.e. gives higher expected posterior model probabilities for the correct snapshot as well as lower ones for all incorrect snapshots whenever $\rho=0.1$ and $N\geq1{,}000$, $\rho=0.3$ and $N\geq300$, and $\rho=0.5$ and $N\geq31$. However, snapshot naïve is biased as a result of not taking the statistical significance of the original study into account. Its better performance is a consequence of its bias, just as the high statistical power of the fixed-effect meta-analysis for small true effect size is a consequence of its overestimation of effect size. Hence, we advise to use snapshot hybrid rather than snapshot naïve. However, if a researcher is certain that the true effect size is, for instance, large, snapshot naïve may be used since this method outperforms snapshot hybrid in most conditions.

Table 7.2. Expected values of the posterior model probabilities of the snapshot hybrid, snapshot naïve, and traditional fixed-effect meta-analysis (FE).

		Snapshot Hybrid				Snapshot Naïve				FE
	N	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5	
ρ=0	31	**0.466**	0.36	0.151	0.023	**0.177**	0.336	0.411	0.076	0.215
	55	**0.535**	0.375	0.089	0.002	**0.212**	0.479	0.304	0.005	0.16
	96	**0.601**	0.368	0.03	0	**0.241**	0.648	0.112	0	0.12
	300	**0.757**	0.243	0	0	**0.338**	0.662	0	0	0.068
	1,000	**0.948**	0.052	0	0	**0.758**	0.242	0	0	0.037
ρ=0.1	31	0.36	**0.351**	0.231	0.057	0.11	**0.258**	0.485	0.147	0.268
	55	0.375	**0.403**	0.211	0.011	0.111	**0.367**	0.501	0.021	0.215
	96	0.368	**0.481**	0.15	0	0.101	**0.552**	0.347	0	0.177
	300	0.243	**0.745**	0.012	0	0.04	**0.94**	0.02	0	0.127
	1,000	0.052	**0.948**	0	0	0.007	**0.993**	0	0	0.103

Table 7.2. Continued

		Snapshot Hybrid				Snapshot Naïve				FE
	N	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5	
ρ=0.3	31	0.151	0.231	**0.367**	0.25	0.027	0.099	**0.456**	0.417	0.381
	55	0.089	0.211	**0.523**	0.178	0.012	0.089	**0.663**	0.236	0.337
	96	0.03	0.15	**0.738**	0.082	0.003	0.066	**0.842**	0.089	0.312
	300	0	0.012	**0.985**	0.003	0	0.008	**0.989**	0.003	0.3
	1,000	0	0	**1**	0	0	0	**1**	0	0.3
ρ=0.5	31	0.023	0.058	0.25	**0.669**	0.002	0.014	0.188	**0.796**	0.516
	55	0.002	0.011	0.178	**0.808**	0	0.002	0.146	**0.851**	0.499
	96	0	0	0.082	**0.918**	0	0	0.075	**0.925**	0.498
	300	0	0	0.003	**0.997**	0	0	0.003	**0.997**	0.499
	1,000	0	0	0	**1**	0	0	0	**1**	0.5

ρ denotes the effect size in the population, *N* is the sample size in the original study and replication, and ρ_S refers to the snapshots of effect size.

Probability of posterior model probability larger than .25 (π>.25) and .75 (π>.75) Table 7.3 shows the probability of how often the posterior model probability is larger than .25 (π>.25), i.e. how often the posterior model probability is larger than the prior model probability. The probability of π>.25 of snapshot hybrid at the correct snapshot is at least .776 and approaches one if the sample sizes increases (bold values in the third to sixth columns of Table 7.3). The same pattern is observed for snapshot naïve, but the probabilities π>.25 at the correct snapshot are smaller for snapshot naïve than snapshot hybrid if ρ=0, and ρ=0.1 and N<300, and higher for ρ=0.3 and ρ=0.5. The lowest probability of π>.25 at the correct snapshot of snapshot naïve is 0.274 for ρ=0 and N=31. However, both methods' probabilities of π>.25 at the incorrect snapshot are also substantial and sometimes even larger for the incorrect than for the correct snapshot. For ρ=0 and N=31, the probability of π>.25 of snapshot hybrid is higher for the incorrect snapshot at ρ_S=0.1 than the correct snapshot (ρ_S=0). The same holds for snapshot naïve at N ≤300 if ρ=0, and at N<96 if ρ=0.1. If the probability of π>.25 is largest for the correct snapshot, the probability of π>.25 at one of the incorrect snapshots can still be substantial. For instance, if ρ=0 or ρ=0.1 and N=300, using snapshot hybrid the probability π>.25 is 0.36 for an incorrect snapshot (ρ_S=0.1 or ρ_S=0, respectively). The probability of π>.25 of snapshot naïve is 0.321 for ρ=0 and N=1,000 at the incorrect snapshot ρ_S=0.1, and 0.495 for ρ=0.1 and N= 96 at ρ_S=0.3. Probabilities of π>.25 at incorrect snapshots also occur for large true effect sizes in combination with small sample sizes. To conclude, a posterior model probability larger than .25 should not be interpreted as evidence in favor of that effect size, but should be interpreted in combination with posterior model probabilities for the other hypothesized effect sizes.

Table 7.4 illustrates the probability of how often the posterior model probability is larger than .75 (π>.75), when evidence can be interpreted as evidence in favor of that true effect. Hence, the probabilities of π>.75 can be interpreted as how often the methods yield the correct conclusion with respect to the magnitude of the true effect size akin to statistical power in null hypothesis significance testing. Table 7.4 also shows how often inconclusive results (columns named "Inconcl.") were obtained, indicating that none of the posterior model probabilities were larger than .75. Focusing first on the results of snapshot hybrid (columns three to seven), the probability of making the wrong decision never exceeds 0.065. However, the probability of obtaining inconclusive results is large for N≤96 when ρ=0 (≥ .706) or ρ=0.1 (≥ .9). The probability of making the correct decision is at least 0.8 (akin to a power of 0.8) for a sample size in between 300 and 1,000 when ρ=0 or ρ=0.1, between 96 and 300 when ρ=0.3, and between 55 and 96 when ρ=0.5.

The probability of making a false decision using snapshot naïve (last five columns) can be substantial for true effect sizes zero to medium. When ρ=0, the probability of making the false decision that the true effect size is of small magnitude is larger than the probability of drawing the correct conclusion for N≤300. The

probability of making false decisions are 0.263 for ρ=0.1 at *N*=55, and 0.157 for ρ=0.3 at *N*=31. The probability of observing inconclusive results with snapshot naïve was large for *N*≤96 when ρ=0 (≥ .628) or ρ=0.1 (≥ .547).The probability of making a correct decision does not exceed 0.8 when ρ=0 for *N*≤1,000, and exceeds 0.8 when ρ=0.1 and sample size between 96 and 300, and ρ=0.3 and ρ=0.5 in combination with sample size between 55 and 96.

To conclude, when true effect size is zero or small, very large sample sizes are required to make correct decisions and snapshot hybrid should be used to take the statistical significance of the original study into account; using snapshot naïve likely results in wrong conclusions when sample size is smaller than 300. When true effect size is medium or large, smaller sample sizes are sufficient to make correct decisions. Snapshot naïve yields both higher probabilities of making correct decisions and lower probabilities of making incorrect decisions when ρ=0.1 and *N*=1,000, ρ=0.3 and *N*≥300, ρ=0.5 and *N*≥96.

7.4.3 Conclusions

The probability of making the correct decision with snapshot hybrid, based on posterior model probabilities larger than .75, is at least 0.8 (akin to a power of 0.8) for a sample size in between 300 and 1,000 when ρ=0 or ρ=0.1, between 96 and 300 when ρ=0.3, and between 55 and 96 when ρ=0.5. The probability of making a false decision using snapshot naïve can be substantial for true effect sizes zero (even for samples sizes of 300 per group in both studies) to medium. Whereas snapshot hybrid outperforms snapshot naïve if there is no or a small true effect, snapshot naïve generally outperformed snapshot hybrid for medium and large true effect sizes. Importantly, the results of both methods also illustrate that it is hard to obtain conclusive results about the magnitude of the true effect size in situations with sample sizes that are illustrative for current research practice. In the penultimate section of this chapter, we use snapshot hybrid to derive the sample size of the replication to obtain evidence of a true effect, akin to power analysis.

Table 7.3. Probability of posterior model probability larger than .25 of snapshot hybrid and snapshot naïve.

		Snapshot Hybrid				Snapshot Naïve			
	N	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5
ρ=0	31	**0.855**	0.914	0.22	0.017	**0.274**	0.725	0.757	0.082
	55	**0.897**	0.84	0.107	0.001	**0.364**	0.891	0.473	0.001
	96	**0.926**	0.705	0.03	0	**0.406**	0.97	0.146	0
	300	**0.939**	0.359	0	0	**0.513**	0.882	0	0
	1,000	**0.984**	0.065	0	0	**0.868**	0.321	0	0
ρ=0.1	31	0.687	**0.87**	0.417	0.063	0.123	**0.512**	0.885	0.206
	55	0.686	**0.891**	0.322	0.007	0.132	**0.677**	0.758	0.015
	96	0.654	**0.895**	0.206	0	0.103	**0.829**	0.495	0
	300	0.36	**0.931**	0.015	0	0.033	**0.992**	0.024	0
	1,000	0.067	**0.982**	0	0	0.007	**0.999**	0	0

Table 7.3 Continued

		Snapshot Hybrid				Snapshot Naïve			
	N	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5
ρ=0.3	31	0.233	0.483	**0.776**	0.385	0.009	0.119	**0.83**	0.656
	55	0.114	0.391	**0.84**	0.252	0.003	0.109	**0.914**	0.338
	96	0.027	0.234	**0.911**	0.104	0	0.081	**0.961**	0.114
	300	0	0.015	**0.995**	0.003	0	0.009	**0.997**	0.003
	1,000	0	0	**1**	0	0	0	**1**	0
	31	0.014	0.062	0.471	**0.887**	0	0.004	0.292	**0.97**
	55	0	0.007	0.265	**0.939**	0	0	0.204	**0.963**
	96	0	0	0.105	**0.973**	0	0	0.096	**0.976**
	300	0	0	0.003	**0.999**	0	0	0.003	**0.999**
	1,000	0	0	0	**1**	0	0	0	**1**

ρ denotes the effect size in the population, *N* is the sample size in the original study and replication, and ρ_S refers to the snapshots of effect size. Note that the sum of probabilities across four snapshots is sometimes larger than 1, because posterior model probabilities can be larger than .25 for more than snapshot.

Table 7.4. Probability of posterior model probability larger than .75 of snapshot hybrid and snapshot naïve.

		Snapshot Hybrid					Snapshot Naïve				
	N	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5	Inconcl.	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5	Inconcl.
ρ=0	31	**0.04**	0	0	0	0.96	**0**	0	0	0.002	0.998
	55	**0.142**	0	0.003	0	0.855	**0.001**	0	0.085	0	0.914
	96	**0.291**	0	0.003	0	0.706	**0.008**	0.343	0.021	0	0.628
	300	**0.641**	0.061	0	0	0.298	**0.118**	0.487	0	0	0.395
	1,000	**0.935**	0.016	0	0	0.049	**0.679**	0.132	0	0	0.189
ρ=0.1	31	0.01	**0**	0	0.002	0.988	0	**0**	0	0.012	0.988
	55	0.033	**0**	0.017	0	0.95	0	**0**	0.263	0.001	0.736
	96	0.057	**0**	0.043	0	0.9	0	**0.288**	0.165	0	0.547
	300	0.065	**0.625**	0.004	0	0.306	0.001	**0.943**	0.007	0	0.049
	1,000	0.018	**0.933**	0	0	0.049	0.001	**0.993**	0	0	0.006

Table 7.4 Continued

		Snapshot Hybrid					Snapshot Naïve				
	N	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5	Inconcl.	ρ_S=0	ρ_S=0.1	ρ_S=0.3	ρ_S=0.5	Inconcl.
ρ=0.3	31	0	0	**0**	0.06	0.94	0	0	**0**	0.157	0.843
	55	0	0	**0.115**	0.05	0.835	0	0	**0.513**	0.077	0.41
	96	0	0	**0.645**	0.025	0.33	0	0.006	**0.802**	0.029	0.163
	300	0	0.005	**0.982**	0.001	0.012	0	0.003	**0.988**	0.001	0.008
	1,000	0	0	**1**	0	0	0	0	**1**	0	0
ρ=0.5	31	0	0	0	**0.498**	0.502	0	0	0	**0.699**	0.301
	55	0	0	0.018	**0.732**	0.25	0	0	0.034	**0.796**	0.17
	96	0	0	0.027	**0.895**	0.078	0	0	0.024	**0.904**	0.072
	300	0	0	0.001	**0.997**	0.002	0	0	0.001	**0.997**	0.002
	1,000	0	0	0	**1**	0	0	0	0	**1**	0

ρ denotes the effect size in the population, *N* is the sample size in the original study and replication, and ρ_S refers to snapshots of effect size. The columns "Inconcl." indicate the probability of observing inconclusive results (i.e., none of the posterior model probabilities for the hypothesized effect sizes was larger than .75)

7.5 Replicability projects

The Reproducibility Project Psychology (RPP; Open Science Collaboration, 2015) and the Experimental Economics Replication Project (EE-RP; Camerer et al., 2016) are two projects that studied the replicability of psychological and economic experimental research by replicating published research. Articles for inclusion in RPP were selected from three high impact psychological journals (Journal of Experimental Psychology: Learning, Memory, and Cognition, Journal of Personality and Social Psychology, and Psychological Science) published in 2008. EE-RP included all articles with a between-subject experimental design published in the American Economic Review and the Quarterly Journal of Economics between 2011 and 2014. The most important finding from these articles was selected for both projects to be replicated and the replication was conducted according to a predefined analysis plan in order to ensure that the replication was as close as possible to the original study.

RPP contained 100 studies that were replicated. A requirement for applying snapshot hybrid is that the original study has to be statistically significant. Three observed effect sizes were reported as not being statistically significant in the original studies of RPP. However, of the remaining 97 original effect sizes reported as statistically significant, recalculation of their *p*-values revealed that four were actually not statistically significant either, but slightly larger than .05 (Open Science Collaboration, 2015); these were excluded as well. The remaining 93 study-pairs included 26 study-pairs that had to be excluded because the correlation coefficient and standard error could not be computed for these study-pairs. This was, for instance, the case for $F(df_1 >1, df_2)$ or χ^2. Hence, the snapshot hybrid and snapshot naïve were applied to 67 study-pairs. EE-RP included 18 study-pairs. The effect size measure of the study-pairs included in EE-RP was also the correlation coefficient. Only two studies had to be excluded because the original study was not statistically significant. Hence, the snapshot hybrid and snapshot naïve were applied to 16 study-pairs of the EE-RP. Information on effect sizes and sample sizes of the study-pairs and the results of the snapshot hybrid and snapshot naïve are reported in Table S1 (https://osf.io/u5gzh/) for EE-RP and Table S2 (https://osf.io/6zpu4/) for RPP.

Table 7.5 lists the posterior model probabilities averaged over all the study-pairs of the snapshot hybrid and snapshot naïve. The difference between the average posterior model probabilities of snapshot hybrid and snapshot naïve was largest for the study-pairs in RPP at $\rho_S=0$ (0.293 vs. 0.126). Average posterior model probabilities of both the snapshot hybrid and snapshot naïve based on the study-pairs in EE-RP were larger at $\rho_S=0.3$ and $\rho_S=0.5$ than the study-pairs in RPP, whereas this was the other way around at $\rho_S=0$ and $\rho_S=0.1$.

Table 7.5. Average posterior model probabilities for the study-pairs in EE-RP and RPP for snapshot hybrid and snapshot naïve at four different snapshots (ρ_S=0; 0.1; 0.3; 0.5).

		ρ_S			
		0	0.1	0.3	0.5
Snapshot Hybrid	EE-RP	0.084	0.137	0.34	0.44
	RPP	0.293	0.234	0.217	0.256
Snapshot Naïve	EE-RP	0.03	0.165	0.361	0.444
	RPP	0.126	0.285	0.267	0.321

Table 7.6 shows the proportions of how often the posterior model probability was larger than .25 for the snapshot hybrid and snapshot naïve. Seven different categories were used because the posterior model probability could be larger than .25 for two snapshots. A study-pair was assigned to one of the categories belonging to snapshots ρ_S=0, 0.1, 0.3, and 0.5 if the posterior model probability of only one of these snapshots was larger than .25. If the posterior model probability of a study-pair was, for instance, 0.4 of snapshot ρ_S=0 and 0.5 of snapshot ρ_S=0.1, the study-pair was assigned to category 0-0.1. The proportion of study-pairs in the categories ρ_S=0 and 0-0.1 was larger for snapshot hybrid than snapshot naïve. On the contrary, snapshot naïve resulted in a higher proportion of study-pairs in the categories ρ_S=0.3, 0.3-0.5 and 0.5 than snapshot hybrid. The large proportions for the categories including two snapshots (e.g., 0-0.1) indicated that drawing definite conclusions about the magnitude of the effect size was often impossible. Comparing the results between EE-RP and RPP shows that the proportion of study-pairs in the categories ρ_S=0 and 0-0.1 was larger for RPP than EE-RP. The proportion of study-pairs in the categories ρ_S=0.3, 0.3-0.5, and 0.5 was larger for EE-RP than RPP.

Table 7.7 presents the proportions of how often the posterior model probability was larger than .75 for the snapshot hybrid and snapshot naïve for snapshots ρ_S=0, 0.1, 0.3, and 0.5. None of the posterior model probabilities at the different snapshots was larger than .75 for snapshot hybrid in 18.8% and 62.7% of the study-pairs in EE-RP and RPP (last column of Table 7.7, respectively. For snapshot naïve, no posterior model probability was larger than .75 in 6.3% of the study-pairs in EE-RP and 52.2% of the study-pairs in RPP. Hence for most of the effects studied in the RPP, no decisions can be made on the magnitude of the true effect size, whereas for EE-RP decisions can be made in the majority of effects studied.

Table 7.6. Proportions of how often the posterior model probability is larger than .25 for the study-pairs in EE-RP and RPP for snapshot hybrid and snapshot naïve at seven different snapshots (ρ_S=0; 0-0.1; 0.1; 0.1-0.3; 0.3; 0.3-0.5; 0.5).

		ρ_S						
		0	0-0.1	0.1	0.1-0.3	0.3	0.3-0.5	0.5
Snapshot Hybrid	EE-RP	0	0.125	0.062	0.062	0.312	0	0.438
	RPP	0.134	0.284	0.045	0.119	0.06	0.164	0.194
Snapshot Naïve	EE-RP	0	0.062	0.125	0	0.375	0	0.438
	RPP	0.015	0.194	0.149	0.06	0.164	0.179	0.239

For the EE-RP, no evidence in favor of a zero true effect was obtained, whereas the majority of effects examined showed evidence in favor of a medium (31.2% for snapshot hybrid and 37.5% for snapshot naïve) or large true effect (43.8% for both snapshot hybrid and snapshot naïve). In RPP, evidence in favor of the null hypothesis was obtained for 13.4% of the effects examined according to snapshot hybrid. This number is much lower than the percentage of statistically nonsignificant replications in RPP (73.1%). A small percentage of study-pairs obtained evidence in favor of a large true effect (16.4%-23.9%).

Table 7.7. Proportions of how often the posterior model probability is larger than .75 for the study-pairs in EE-RP and RPP for snapshot hybrid and snapshot naïve at four different snapshots (ρ_S=0; 0.1; 0.3; 0.5) and how often none of the posterior model probabilities is larger than .75 (Inconclusive results, final column).

		ρ_S				
		0	0.1	0.3	0.5	Inconcl.
Snapshot Hybrid	EE-RP	0	0.062	0.312	0.438	0.188
	RPP	0.134	0.030	0.045	0.164	0.627
Snapshot Naïve	EE-RP	0	0.125	0.375	0.438	0.062
	RPP	0.015	0.119	0.104	0.239	0.522

The studies in RPP can be divided into social and cognitive psychology studies. The proportions of how often the posterior model probability was larger than .75 for social psychology and cognitive psychology is presented in Table 7.8. According to snapshot hybrid, the true effect size was more often zero in studies in social psychology than in cognitive psychology (23.5% vs. 3.0%), whereas it was more often large in cognitive psychology than in social psychology (21.2% vs 11.8%). For approximately half of the study-pairs in both fields, none of the posterior model probabilities of snapshot hybrid and snapshot naïve was larger than .75 (final column of Table 7.8).

Table 7.8. Proportions of how often the posterior model probability is larger than .75 for the study-pairs in RPP grouped by social and cognitive psychological studies for snapshot hybrid and snapshot naïve at four different snapshots (ρ_S=0; 0.1; 0.3; 0.5) and how often none of the posterior model probabilities is larger than .75 (Inconclusive results, final column).

		ρ_S				
		0	0.1	0.3	0.5	Inconcl.
Snapshot Hybrid	Social	0.235	0.059	0	0.118	0.588
	Cognitive	0.030	0	0.091	0.212	0.667
Snapshot Naïve	Social	0.029	0.176	0.059	0.118	0.618
	Cognitive	0	0.061	0.152	0.364	0.424

7.6 Determining sample size of replication with snapshot hybrid

Snapshot hybrid can also be used for computing the required sample size where $P(\pi_x \geq a)=b$ with a being the desired posterior model probability and b the desired probability for a correct decision (i.e., desired probability of observing a posterior model probability larger than a). Computing the required sample size with snapshot hybrid is akin to computing the required sample size with a power analysis in null hypothesis significance testing. A value for a is 0.75 that corresponds to a Bayes Factor of 3 (Kass & Raftery, 1995) and b equal to 0.8 reflecting 80% statistical power. Note that any other desired values for a and b can be chosen. We do not compute the required sample size with snapshot naïve because it falsely does not take the significance of the original study into account and is unsuitable for ρ=0.

For computing the required sample size of the replication, we need information on the effect size or test statistic and sample size(s) of the original study and the expected true effect size in the population. The four different hypothesized effect sizes or snapshots (zero, small, medium, large) are used as before. $P(\pi_x \geq a)$ for

the hypothesized effect size is calculated using numerical integration. The required sample size of the replication can be obtained by optimizing the sample size until *b* is obtained. The required sample size of the replication is also computed when the original study is ignored. A researcher may opt to ignore information of the original study if he or she believes that the original study does not estimate the same true effect or has other reasons to discard this information.

The procedure for determining the sample size of the replication is programmed in R and requires as input the observed effect size and sample size of the original study, α-level, desired posterior model probability (*a*), and desired probability (*b*). Users can also specify (besides specifying the α-level, *a*, and *b*) the two group means, standard deviations, and sample sizes or a *t*-value and sample sizes in order to compute the required sample size in case of a two-independent groups design. The output is a 4×2 table with for each hypothesized effect size the required total replication sample size when the original effect size is included or excluded. An easy to use web application is available to compute the required sample size for researchers who are not familiar with R (https://rvanaert.shinyapps.io/snapshot/).

Determining the sample size of the replication with snapshot hybrid for the example by Maxwell et al. (2015) with r_o = .243 and *N*=80 resulted in the sample sizes presented in Table 7.9 for each hypothesized effect size, using *a*=0.75 and *b*=0.8. Higher sample sizes are needed for zero and small hypothesized effect size than for medium and strong hypothesized effect size. When ignoring the original study, *less* observations are needed for nonzero hypothesized effect sizes than after incorporating the original study. The reason is that, after taking the statistical significance of the original effect into account, the original effect provides evidence in favor of a zero true effect. This is also the reason that *more* observations are needed for a zero hypothesized effect size when the original study is ignored (*N*=645) relative to incorporating it (*N*=587). We note that Maxwell et al. (2015) conducted a power analysis based on the results of the original study to compute the required sample size in the replication and ended up with a sample size of 172. The explanation for their low required sample size is that they likely overestimate effect size with the original study by not taking its statistical significance into account. Sample size of the replication obtained with snapshot hybrid may also be larger than the sample size obtained with a power analysis, because four different hypothesized effect sizes are examined with snapshot hybrid instead of one in a usual power analysis. However, this will only have a minor influence since the posterior model probability for at least two out of four effect sizes will be small if the sample size is large.

Finally, we emphasize that snapshot hybrid is sensitive to the observed effect size in the original study. An observed effect size in the original study close to a hypothesized effect size results in a smaller required sample size for the replication than if the observed effect size substantially deviates from the hypothesized effect size. Our web application can be used to examine the sensitivity of the required

sample size to the results of the original study. This provides information on how much evidence there is for a particular hypothesized effect size in the original study after taking into account statistical significance in this study. The first column is affected by the statistics of the original study, whereas the last is not because it ignores the original study. For instance, if Maxwell et al. (2015) had postulated r_o = .243 and N=800, the required sample size for the replication by taking into account the information of the original study is 3,691, 561, a sample size of less than 4, and 1,521 for ρ=0; 0.1; 0.3; 0.5, respectively. The reason that only very few observations are required for medium hypothesized effect size and very large sample size for zero and large hypothesized effect size is that the original study provides strong evidence of a close to medium true effect size.

Table 7.9. Required sample size computed with snapshot hybrid based on characteristics of the original study as described in Maxwell et al. (2015); r_o = .243 and N=80.

	With original study	Without original study
ρ_S=0	587	645
ρ_S=0.1	709	664
ρ_S =0.3	223	215
ρ_S =0.5	284	116

Sample size was computed with snapshot hybrid for a desired posterior model probability of a=0.75 and the desired probability of observing a posterior model probability larger than a was b=0.8. The hypothesized effect size was equal to ρ_S=0 (no effect), 0.1 (small), 0.3 (medium), and 0.5 (large). The penultimate column refers to the required sample size where information of the original study is included and the last column where this information is excluded.

7.7 Conclusion and discussion

The high number of statistical significant findings in the literature (e.g., Fanelli, 2010a; Fanelli, 2012; Sterling et al., 1995) does not match the average low statistical power (Bakker et al., 2012; Button et al., 2013; Cohen, 1990), and raises concerns about the reliability of published findings. Several projects recently systematically replicated published studies in medicine (Begley & Ellis, 2012), psychology (RPP; Open Science Collaboration, 2015), and economics (EE-RP; Camerer et al., 2016) to examine their replicability. Characteristic of all these projects is that most effects were originally statistically significant, but not significant in the

replication. Problems with traditional methods to analyze these results are that (1) NHST is not informative for the magnitude of the true effect size, (2) no evidence can be obtained for a true zero effect, and (3) they do not take into account the statistical significance of the original study. To solve these problems, we developed a method (snapshot Bayesian hybrid meta-analysis method, snapshot hybrid for short) that computes the posterior model probability for a set of effect sizes (no, small, medium, and large effect) by statistically combining the original study and replication, while at the same time taking the statistical significance of the original study into account. Desirable properties of snapshot hybrid are its few assumptions, its straightforward interpretation as the probability that the true effect size is zero, small, medium or large, and the ease with which posterior model probabilities can be recalculated using different sets of prior model probabilities.

Researchers can apply snapshot hybrid with the R function "snapshot" in the "puniform" package (package can be installed with the following R code devtools::install_github("RobbievanAert/puniform"). The "req_ni_r" function which is also in the "puniform" package can be used for computing the required sample size of the replication to achieve a certain posterior model probability for hypothesized effect sizes equal to zero, small, medium, and large, akin to power analysis. Researchers not familiar with R can use the web application (https://rvanaert.shinyapps.io/snapshot/) for applying snapshot hybrid and computing the required sample size of the replication.

We examined the performances of snapshot hybrid and a method that does not take into account that the original study is statistically significant (snapshot naïve). Our analysis shows that snapshot naïve hardly ever can provide evidence in favor of a true zero effect; even if both original and replication effect have a sample size of 1,000, the expected posterior model probability in favor of a small effect is close to the prior model probability of .25. Hence, we recommend not using any method that does not take into account statistical significance of the original study (including fixed-effect meta-analysis) when the goal is examining if a nonzero true effect size exists. Snapshot naïve outperformed snapshot hybrid for medium true effect size and sample sizes of 300 per study, and for large true effect size and sample sizes of 31. Thus, we only recommend using methods that do not correct for statistical significance in the original study when true effect size is strongly suspected to be large, or medium in combination with large sample sizes (> 300) of both the original study and the replication.

By taking the statistical significance of the original effect into account snapshot hybrid yields accurate evaluations of not only zero true effect size, but larger true effect size as well. The probability of making the correct decision with snapshot hybrid, based on posterior probabilities larger than .75, is at least 0.8 (akin to a power of 0.8) for sample sizes in between 300 and 1,000 when $\rho=0$ or $\rho=0.1$, between 96 and 300 when $\rho=0.3$, and between 55 and 96 when $\rho=0.5$. Due to its accurate evaluations,

particularly if true effect size is zero, we recommend using snapshot hybrid when evaluating effect size based on a statistically significant original study and a replication. Importantly, our results also confirm previous research (e.g., Etz & Vandekerckhove, 2016; Maxwell et al., 2015) that it is hard to obtain conclusive results about the magnitude of the true effect size in situations with sample sizes that are illustrative for current research practice.

Several conclusions can be drawn from the application of snapshot hybrid to the data of RPP and EE-RP. First, in the majority of study-pairs in RPP no evidence was found for any of the true effects, as opposed to in EE-RP where evidence was found for one true effect size considering a zero, small, medium, or large true effect size in about 80% of the study-pairs. This shows that sample sizes of the original study and replication in RPP were generally often not large enough to draw definite conclusions about the magnitude of the true effect size (e.g, Etz & Vandekerckhove, 2016; Maxwell et al., 2015). Second, true effect size was generally higher in EE-RP than in RPP. However, evidence in favor of the null hypothesis was found for only 13.4% of the study-pairs in RPP, as opposed to the much higher percentage of statistically nonsignificant replications in RPP (73.1%). This is in line with the argumentation of Maxwell et al. (2015), who argue that sample sizes of the replications in RPP are generally too small to draw conclusions on the absence of a true effect. Finally, within RPP true effect size was generally lower for study-pairs in social than cognitive psychology.

Our study and snapshot hybrid have several limitations. First, we analytically evaluated the statistical properties of the snapshot hybrid and snapshot naïve by assuming equal sample sizes of the original study and replication. Most often their sample sizes are somewhat different, with the replication generally having larger sample size than the original study. Hence, our results on statistical properties should be considered as illustrations of the effect of sample size on the performance of both snapshot naïve and snapshot hybrid. Note that our web application can be used to examine what the effect is of different sample sizes of the original study when calculating the required replication sample size.

A limitation of snapshot hybrid seems to be the requirement that the original study is statistically significant. However, most studies in the social sciences contain statistically significant results; about 95% of the studies in psychology contain significant results (e.g., Fanelli, 2012; Sterling et al., 1995) and 97% and 89% of the original findings in RPP (Open Science Collaboration, 2015) and EE-RP (Camerer et al., 2016) were statistically significant. Another apparent limitation of snapshot hybrid is that it assumes that the same true effect is underlying the original study and replication. However, an exact replication is highly similar to an original study and no or a small amount of heterogeneity in true effect size may be expected. Furthermore, two studies are not sufficient to estimate the amount of heterogeneity (e.g., Borenstein et al., 2010; IntHout et al., 2014).

Our current implementation of snapshot hybrid assumes discrete values of hypothesized effect size, rather than distributions of hypothesized effect size as in continuous Bayesian analyses. A disadvantage of using discrete values is that if the true effect size is in between these values, the results of our analysis on statistical properties do no longer apply. That is, higher samples sizes are needed to obtain evidence in favor of the discrete value closest to the actual true effect size. Other hypothesized effect sizes can be used as a sort of sensitivity analysis to examine whether the true effect size is in between the originally proposed hypothesized effect sizes. For example, if the true effect size is $\rho=0.2$ and thus between $\rho=0.1$ and $\rho=0.3$, the highest posterior model probability will be observed for $\rho=0.2$ when hypothesized effect sizes $\rho=0$, $\rho=0.1$, $\rho=0.2$, and $\rho=0.3$ are chosen. Snapshot hybrid could also be implemented using intervals of hypothesized effect size, say 0, 0-0.1, 0.1-0.3, 0.3-0.5, > 0.5, while keeping all of its desirable properties except for one: The posterior model probabilities can no longer be easily updated using Equation (4) when assuming other than equal prior model probabilities. However, we chose for discrete hypothesized effect size values in the current implementation of snapshot hybrid because we believe most researchers think in terms of zero, small, medium, and large effect size, and wish to carry out power analyses assuming these effect sizes as in our web application.

Another limitation of snapshot hybrid in its current implementation is that it can only deal with one (statistically significant) original study and one replication. Including more than one original study or replication will usually yield more divergence in the posterior model probabilities of the set of effect sizes and enable researchers to draw more reliable conclusions. We will extend the current snapshot hybrid method such that it can deal with multiple original studies and replications in the future. A final limitation is that the results of snapshot hybrid will be biased in case of questionable research practices or *p*-hacking in the original study. Questionable research practices bias the *p*-values (e.g., Bruns & Ioannidis, 2016; Simonsohn et al., 2014a; Ulrich & Miller, 2015; van Aert, Wicherts, et al., 2016; van Assen et al., 2015) and therefore also the truncated density of the original study. The extent to which the results of snapshot hybrid becomes biased due to questionable research practices may be subject for further study. We note, however, that *no* existing method can deal with questionable research practices.

To conclude, the unrealistic high rate of statistically significant findings in the published literature and the results of RPP and EE-RP suggest that the literature is distorted with false positive findings and too high effect size estimates. We propose and recommend snapshot hybrid for evaluating the magnitude of the true effect size underlying an original study and replication that computes the posterior model probability for a zero, small, medium, and large hypothesized effect. The method has the advantage over other existing methods, because it is the first method that adjusts for publication bias by taking statistical significance of the original study into account.

Moreover, the method can also be used for determining the sample size in the replication akin to power analysis in NHST. The snapshot hybrid method is easy to understand and to apply and provides useful insights in evaluating an original study and replication.

CHAPTER 8

Multi-step estimators of the between-study variance: The relationship with the Paule-Mandel estimator

Abstract

A wide variety of estimators of the between-study variance are available in random-effects meta-analysis. Many, but not all, of these estimators are based on the method of moments. The DerSimonian-Laird estimator is widely used in applications, but the Paule-Mandel estimator is an alternative that is now recommended. Recently, DerSimonian and Kacker have developed two-step moment based estimators of the between-study variance. We extend these two-step estimators so that multiple (more than two) steps are used. We establish the surprising result that the multi-step estimator tends towards the Paule-Mandel estimator as the number of steps becomes large. Hence, the iterative scheme underlying our new multi-step estimator provides a hitherto unknown relationship between two-step estimators and Paule-Mandel estimator. Our analysis suggests that two-step estimators are not necessarily distinct estimators in their own right, instead they are quantities that are closely related to the usual iterative scheme that is used to calculate the Paule-Mandel estimate. The relationship that we establish between the multi-step and Paule-Mandel estimator is another justification for the use of the latter estimator. Two-step and multi-step estimators are perhaps best conceptualized as approximate Paule-Mandel estimators.

This chapter is published as van Aert, R. C. M., & Jackson, D. (2018). Multi-step estimators of the between-study variance: The relationship with the Paule-Mandel estimator. *Statistics in Medicine.* doi: doi:10.1002/sim.7665

8.1 Introduction

Meta-analysis statistically combines effect size estimates from different studies in order to calculate a quantitative summary of the evidence base. Two important outcomes from a meta-analysis are the estimates of the overall effect size and the between-study variance (the variance of the studies' true effect sizes). Between-study heterogeneity refers to the possibility that there is more variation in the studies' observed effect sizes than what would be expected by sampling variability alone (DerSimonian & Laird, 1986; Higgins & Thompson, 2002), and is often present in meta-analyses (Higgins, 2008; Higgins et al., 2009; Kontopantelis et al., 2013). Characteristics of the included studies (e.g., differences between populations from which participants were sampled or treatments across studies) can be incorporated as moderators in meta-regressions to explore and explain the between-study heterogeneity (Borenstein et al., 2009; Thompson & Sharp, 1999; van Houwelingen, Arends, & Stijnen, 2002). However, random-effects meta-analyses are often used to account for, but not explain, between-study heterogeneity.

A wide variety of estimators are available for the between-study variance. Two recent papers (Langan et al., 2016; Veroniki et al., 2016) review existing research on these estimators and recommended either the Paule-Mandel (PM) estimator (Paule & Mandel, 1982) or the restricted maximum likelihood (REML) estimator (Raudenbush, 2009). However, the DerSimonian-Laird (DL) estimator is most often used in practice (Jackson, Bowden, & Baker, 2010; Kontopantelis et al., 2013; Wiksten et al., 2016). The popularity of the DL estimator is due to its simplicity, because it is calculated from an easily computed non-iterative method, and also because it is already familiar to applied meta-analysts. In this chapter, we focus on estimators that are motivated by the method of moments, which includes the DL and PM estimators, but not REML.

In particular, we use the general method of moments estimator (i.e., with an arbitrary set of weights for the effect sizes) proposed by DerSimonian and Kacker (2007) to develop a new multi-step DL estimator. This idea extends the two-step DL (DL_2) estimator which was also proposed by DerSimonian and Kacker (2007). The usual (one-step) DL estimator uses the inverse of the studies' within-study sampling variances as weights to estimate the between-study variance. In the two step-estimation procedure, the estimate of the usual DL estimator is calculated in the first step and this estimate is then included in the weights of the second step. Full details of the DL_2 estimator are provided in section 8.3. The statistical properties of the DL_2 estimator are largely unknown, because the method has rarely been topic of further study. Bhaumik et al. (2012) studied the statistical properties of the DL_2 estimator and concluded that for rare events both the DL_2 and PM estimators are negatively biased. It was our initial intuition that allowing the number of steps to tend to infinity in our new multi-step estimator would define a new type of estimator. However, working empirically to begin with and then mathematically, we will demonstrate that the PM

estimator is obtained if the number of steps tends towards infinity. Hence, we will instead establish the relationship between the two-step estimators and PM estimator which is another justification for the use of the Paule-Mandel estimator.

The rest of this chapter is set out as follows. We continue with describing the random-effects model for meta-analysis in section 8.2. In section 8.3, we describe three existing moments based estimators, DL, DL_2, and PM. Our new multi-step estimator is introduced in section 8.4. Subsequently, we apply these estimators to three contrasting examples in section 8.5 where we empirically show that the multi-step estimator tends towards the PM estimator as the number of steps becomes large, where this convergence occurs quickly in practice. Section 6 contains mathematics that formally establishes the relationship between the multi-step estimators and PM estimator. We explore the use of meta-regression models in section 8.7, and we conclude with a short discussion in section 8.8.

8.2 The random-effects model

The random-effects model assumes that the effect size estimates y_i, i=1, ..., n, are extracted from separate studies. This model can be written as

$$y_i = \mu + \mu_i + \epsilon_i \tag{1}$$

where μ is the average true effect size, μ_i is a random effect indicating the difference between the ith study's true effect size and μ, and ϵ_i is the ith study's sampling error. It is commonly assumed that $\mu_i \sim N(0, \tau^2)$ where τ^2 is the between-study variance and $\epsilon_i \sim N(0, \sigma_i^2)$, where σ_i^2 is the within-study sampling variance of the ith study. Furthermore, all μ_i and ϵ_i are assumed to be mutually independent. The within-study sampling variances σ_i^2 are usually estimated in practice, and then assumed to be known in the analysis. We will emphasize that the σ_i^2 are estimated by writing $\widehat{\sigma}_i^2$ as their estimates.

The parameter μ is usually of primary interest. The usual method for making inferences about μ initially estimates τ^2 and then treats the resulting estimate as fixed and known (Biggerstaff & Tweedie, 1997; Veroniki et al., 2016). Hence, the conventional weights in the random-effects model, $1/(\widehat{\sigma}_i^2 + \hat{\tau}^2)$, are treated as fixed and known and the usual inferential procedure for μ is straightforward (Borenstein et al., 2009). However, the estimate of the between-study variance, $\hat{\tau}^2$, is our primary interest here with moment based estimators as our focus.

8.3 Moment based methods for estimating the between-study variance

Most of the moment based estimators for τ^2 are a special case of a general method of moments estimator (DerSimonian & Kacker, 2007). To derive this general

estimation method, DerSimonian and Kacker (2007) propose methodology for estimating τ^2 using an arbitrary set of weights a_i, i=1, ..., n, where all a_i are fixed positive constants. To estimate τ^2, DerSimonian and Kacker (2007) propose equating $\sum_{i=1}^{n} a_i\,(y_i - \widehat{\mu})^2$, where $\widehat{\mu} = \sum_{i=1}^{n} a_i\,y_i / \sum_{i=1}^{n} a_i$, to its expected value. As explained by DerSimonian and Kacker (2007), this results in the estimating equation

$$\hat{\tau}^2_{MM} = \frac{[\sum_{i=1}^{n} a_i\,(y_i - \widehat{\mu})^2] - [\sum_{i=1}^{n} a_i\,\widehat{\sigma}_i^2 - \sum_{i=1}^{n} a_i^2\,\widehat{\sigma}_i^2 \,/\, \sum_{i=1}^{n} a_i]}{\sum_{i=1}^{n} a_i - \sum_{i=1}^{n} a_i^2 \,/\, \sum_{i=1}^{n} a_i} \qquad (2)$$

where negative estimates $\hat{\tau}^2_{MM}$ from equation (2) are truncated to zero (because $\tau^2 \geq 0$). An often overlooked point is that the calculation of the expectation of $\sum_{i=1}^{n} a_i\,(y_i - \widehat{\mu})^2$, that gives rise to the estimating equation (2), ignores the uncertainty in the $\widehat{\sigma}_i^2$ and has taken $\sigma_i^2 = \widehat{\sigma}_i^2$. Although when presenting equation (2), we have emphasized that the estimates $\widehat{\sigma}_i^2$ are used in the calculation, this does not clearly convey the fact that the estimation does not take their uncertainty into account. Kulinskaya and Dollinger (2015) and Hoaglin (2016a) criticize moment based methods for this type of reason, because ignoring uncertainty in $\widehat{\sigma}_i^2$ may cause bias in the estimate of τ^2 especially if the sample size of the studies is small. By ignoring the uncertainty in the within-study variances we have that $\hat{\tau}^2_{MM}$ is unbiased before truncation to zero, but a positive bias in the estimator is introduced by the truncation (Rukhin, 2013; Viechtbauer, 2005).

8.3.1 The DerSimonian-Laird estimator

The DL estimator (DerSimonian & Laird, 1986), $\hat{\tau}^2_{DL}$, is obtained by taking $a_i = 1/\widehat{\sigma}_i^2$ in equation (2). We then have $\sum_{i=1}^{n} a_i\,\widehat{\sigma}_i^2 - \sum_{i=1}^{n} a_i^2\,\widehat{\sigma}_i^2 \,/\, \sum_{i=1}^{n} a_i = n - 1$, so that equation (2) simplifies when using this standard set of weights. Negative estimates are again truncated to zero. Uncertainty in $\widehat{\sigma}_i^2$ is, as in equation (2), neglected by treating the weights $a_i = 1/\widehat{\sigma}_i^2$ as fixed constants. This may result in bias when estimating τ^2 using the DL estimator especially if sample sizes of the studies is small (Hoaglin, 2016a; Kulinskaya & Dollinger, 2015).

8.3.2 The two-step DerSimonian-Laird estimator

DerSimonian and Kacker (2007) propose an alternative estimator that is an extension of the DL estimator. The usual DL estimate $\hat{\tau}^2_{DL}$, described in the previous section, is calculated in the first step. The two-step DL (DL_2) estimator adds a second step by incorporating $\hat{\tau}^2_{DL}$ into the weights, and computes $\hat{\tau}^2_{DL_2}$ using estimating equation (2) with $a_i = 1/(\widehat{\sigma}_i^2 + \hat{\tau}^2_{DL})$.

To describe the two-step DerSimonian-Laird estimator more explicitly, and also to define the PM and multi-step DL estimators below, it is convenient to define the quantity

$$Q_{GEN}(\tau^2) = \sum_{i=1}^{n} \frac{(y_i - \widehat{\mu}\,(\tau^2))^2}{\widehat{\sigma}_i^2 + \tau^2} \tag{3}$$

where $\widehat{\mu}\,(\tau^2) = \sum_{i=1}^{n} y_i\,/(\widehat{\sigma}_i^2 + \tau^2) / \sum_{i=1}^{n} 1\,/(\widehat{\sigma}_i^2 + \tau^2)$. Then $Q_{GEN}(0)$ is the usual Q statistic used in meta-analysis (Cochran, 1954; Hoaglin, 2016b). From equation (2) with $a_i = 1/(\widehat{\sigma}_i^2 + \hat{\tau}_{DL}^2)$, we have

$$\hat{\tau}_{DL_2}^2 = \frac{Q_{GEN}(\hat{\tau}_{DL}^2) - [\sum_{i=1}^{n} \widehat{\sigma}_i^2\,/(\widehat{\sigma}_i^2 + \hat{\tau}_{DL}^2) - \sum_{i=1}^{n} \widehat{\sigma}_i^2\,/(\widehat{\sigma}_i^2 + \hat{\tau}_{DL}^2)^2 / \sum_{i=1}^{n} 1\,/(\widehat{\sigma}_i^2 + \hat{\tau}_{DL}^2)]}{\sum_{i=1}^{n} 1\,/(\widehat{\sigma}_i^2 + \hat{\tau}_{DL}^2) - \sum_{i=1}^{n} 1\,/(\widehat{\sigma}_i^2 + \hat{\tau}_{DL}^2)^2 / \sum_{i=1}^{n} 1\,/(\widehat{\sigma}_i^2 + \hat{\tau}_{DL}^2)} \tag{4}$$

where we again truncate negative estimates to zero. The weights $a_i = 1/(\widehat{\sigma}_i^2 + \hat{\tau}_{DL}^2)$ are intuitively appealing, because we then weight by estimates of the studies' total precisions which are also the standard weights when making inferences about μ in the random-effects model (Borenstein et al., 2009; Shadish & Haddock, 2009). Using these weights raises further statistical issues, because they are now functions of both the $\widehat{\sigma}_i^2$ and the estimated between-study variance $\hat{\tau}_{DL}^2$. There is statistical error in both of these estimated variance components, and so treating the weights $a_i = 1/(\widehat{\sigma}_i^2 + \hat{\tau}_{DL}^2)$ as fixed constants continues to have the potential to have unfortunate implications for the estimation.

It is possible to use other estimators in the first step, and DerSimonian and Kacker (DerSimonian & Kacker, 2007) also propose using the Cochran ANOVA estimator (Cochran, 1954; Hedges, 1983) that is based on an unweighted sum of squares for this purpose. However, the DL estimator is so common in application that we only explore the use of two-step and multi-step estimators that use this particular estimator. Nonetheless, our main results will apply regardless of the type of estimator used in the first step as we will explain below. Hence, generalisability of our results is not restricted by using the DL estimator in the first step, but the results also apply if, for instance, the Cochran ANOVA estimator is used in the first step.

8.3.3 The Paule-Mandel estimator

Another moment based estimator for τ^2 is the Paule-Mandel (PM) estimator (Paule & Mandel, 1982). This estimation method exploits the fact that $Q_{GEN}(\tau^2) \sim \chi_{n-1}^2$, so that $\hat{\tau}_{PM}^2$ is obtained by matching $Q_{GEN}(\tau^2)$ to its expected value. Hence, $\hat{\tau}_{PM}^2$ is the solution to

$$Q_{GEN}(\hat{\tau}^2_{PM}) = n - 1. \tag{5}$$

For any given dataset, $Q_{GEN}(\tau^2)$ is a monotonically decreasing continuous function of τ^2. As a consequence, equation (5) always provides a unique estimate if $Q_{GEN}(0) \geq (n-1)$ (DerSimonian & Kacker, 2007; Knapp, Biggerstaff, & Hartung, 2006; Viechtbauer, 2007b; Viechtbauer, López-López, Sánchez-Meca, & Marín-Martínez, 2015). If $Q_{GEN}(0) < (n-1)$ then no positive solution to the estimating equation (5) exists, and we take $\hat{\tau}^2_{PM} = 0$. The estimating equation (5) is non-linear and so must be solved numerically, but this is straightforward in practice. An empirical Bayes estimator for estimating τ^2 (Berkey, Hoaglin, Mosteller, & Colditz, 1995; Morris, 1983) was developed independently, but this has subsequently been shown to be equivalent to the PM estimator (Veroniki et al., 2016; Viechtbauer et al., 2015).

Unlike the DL and DL_2 estimator and other moment based estimators, the PM estimator does not directly use estimating equation (2). This is because the general method of moments treats the weights a_i as fixed (and therefore known) constants, but the PM estimator uses weights $1/(\hat{\sigma}^2_i + \tau^2)$ that are explicitly unknown (because τ^2 is unknown). The PM estimator is motivated using the method of moments, but otherwise there is no direct connection between the PM estimator and other moment based estimators. We introduce our new multi-step estimator in the next section, and we will illustrate the relationship between the PM and the two-step estimator.

8.4 The multi-step DerSimonian-Laird estimator

In this section, we develop the multi-step DL estimator as a natural extension of the DL_2 estimator. From equation (4), we have that the DL_2 estimator is simply the estimate from the more general estimating equation (2) where the weights are $a_i = 1/(\hat{\sigma}^2_i + \hat{\tau}^2_{DL})$. The key observation is that the two-step estimator uses weights that are the reciprocal of the estimated total study variances, where the between-study variance is estimated using the usual DL estimator. A natural way to extend this estimator to define a three-step estimator is to use weights that are reciprocal of the estimated total study variances, where the between-study variance is estimated using the DL_2 estimator. Hence, we define $\hat{\tau}^2_{DL_3}$ to be

$$\hat{\tau}^2_{DL_3} = \frac{Q_{GEN}(\hat{\tau}^2_{DL_2}) - [\sum_{i=1}^n \hat{\sigma}^2_i / (\hat{\sigma}^2_i + \hat{\tau}^2_{DL_2}) - \sum_{i=1}^n \hat{\sigma}^2_i / (\hat{\sigma}^2_i + \hat{\tau}^2_{DL_2})^2 / \sum_{i=1}^n 1 / (\hat{\sigma}^2_i + \hat{\tau}^2_{DL_2})]}{\sum_{i=1}^n 1 / (\hat{\sigma}^2_i + \hat{\tau}^2_{DL_2}) - \sum_{i=1}^n 1 / (\hat{\sigma}^2_i + \hat{\tau}^2_{DL_2})^2 / \sum_{i=1}^n 1 / (\hat{\sigma}^2_i + \hat{\tau}^2_{DL_2})}$$

where as before we truncate negative estimates to zero. We can then define a four-step estimator in a similar way, using equation (2) with weights $a_i = 1/(\hat{\sigma}^2_i + \hat{\tau}^2_{DL_3})$, and then a five-step estimator using weights $a_i = 1/(\hat{\sigma}^2_i + \hat{\tau}^2_{DL_4})$, and so on. In general, we define the $(k+1)$th step DL estimator as

$$\hat{\tau}^2_{DL_{k+1}} = \frac{Q_{GEN}(\hat{\tau}^2_{DL_k}) - [\sum_{i=1}^n \hat{\sigma}_i^2/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k}) - \sum_{i=1}^n \hat{\sigma}_i^2/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})^2 / \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})]}{\sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k}) - \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})^2 / \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})} \quad (6)$$

for $k \geq 1$, where $\hat{\tau}^2_{DL_1}$ is defined to be the usual DL estimator $\hat{\tau}^2_{DL}$. As usual, we truncate the resulting estimate from equation (6) to zero if the solution is negative. Written explicitly in terms of this truncation, the $(k+1)$th step DL estimator is

$$\hat{\tau}^2_{DL_{k+1}} = \max\left(0, \frac{Q_{GEN}(\hat{\tau}^2_{DL_k})}{\sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k}) - \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})^2 / \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})} - \frac{[\sum_{i=1}^n \hat{\sigma}_i^2/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k}) - \sum_{i=1}^n \hat{\sigma}_i^2/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})^2 / \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})]}{\sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k}) - \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})^2 / \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})}\right). \quad (7)$$

In practice, we compute $\hat{\tau}^2_{DL_k}$ recursively by first computing $\hat{\tau}^2_{DL}$, then $\hat{\tau}^2_{DL_2}$, then $\hat{\tau}^2_{DL_3}$ and so on until we reach the required value of k. However, all of these estimators are available in closed form and so it is in principle also possible to write $\hat{\tau}^2_{DL_k}$ in this way. Assuming that the limit exists, we define $\lim_{k\to\infty} \hat{\tau}^2_{DL_k} = \hat{\tau}^2_{DL_\infty}$. We will see below that, whenever convergence occurs, $\hat{\tau}^2_{DL_\infty} = \hat{\tau}^2_{PM}$, so that instead of defining a new estimator we establish the relationship between existing estimates by taking this limit.

8.5 Examples

In this section, we apply the DL, PM, DL_2, and multi-step DL estimator to three contrasting examples. Having illustrated our main findings empirically using these examples, we will demonstrate them mathematically in section 8.6.

8.5.1 Characteristics of the three examples

Our first example is a meta-analysis by Bangert-Drowns et al. (2004) studying the effect of school-based writing-to-learn interventions on academic achievement. This meta-analysis consists of 48 estimated standardized mean differences (i.e., Hedges' g). The second example is obtained from Sterne et al. (2001), and is a meta-analysis on the effectiveness of intravenous magnesium in acute myocardial infarction. This meta-analysis consists of sixteen estimated log odds ratios. The third example is a meta-analysis on the efficacy of two treatments for post-traumatic stress disorder (Ho & Lee, 2012). This meta-analysis consists of ten standardized mean differences. The metafor package (Viechtbauer, 2010) was used to calculate the DL and PM estimators, and we used our own bespoke code to recursively calculate the

multi-step DL_k estimator. R code for applying these estimators to the examples is available via https://osf.io/paqzm/.

8.5.2 Results

Table 8.1 shows the DL, DL_2, DL_k and PM estimates of τ^2 for all three examples. For each example, we calculated the multi-step DL estimator until the $(k+1)$th step DL estimator was the same as the kth step estimator up to 4 decimal places. Convergence was taken to have been reached at this point, so that any further steps would result in the same estimate to this level of numerical accuracy. From Table 8.1, we can see that this convergence was reached in 6, 10 and 4 steps, for examples one, two and three, respectively. Furthermore, we can see that in each case the DL_2 estimate is closer to the PM estimate than the DL estimate, and that the DL_k estimate converges to the PM estimate. The way in which this convergence occurred was different for each example. For the first example obtained from Bangert-Drowns et al. (2004), the DL estimate was notably less than the PM estimate. Then the DL_2 estimate took a large step towards the PM estimator, and after this convergence was quickly reached. For the second example obtained from Sterne et al. (2001), the DL estimate was notably greater than the PM estimate and once again the DL_2 estimate took a large step towards the PM estimator (and in fact 'overshot' this). Convergence of the multi-step DL estimator was reasonably fast although the sequence produced by the DL_k estimates was not monotone until $k \geq 7$. For the third example obtained from Ho and Lee (2012), the DL and PM estimators are similar and convergence was very quickly reached.

8.5.3 Conclusions

Although the way in which the multi-step DL estimator converged to the PM estimator was different in each case, all three examples illustrated the surprising finding that $\lim_{k\to\infty} \hat{\tau}^2_{DL_k} = \hat{\tau}^2_{DL_\infty} = \hat{\tau}^2_{PM}$. A large number of simulations (see https://osf.io/dpuzs/ for R code) using $a_i = 1/\hat{\sigma}_i^2$ and $a_i = 1/(\hat{\sigma}_i^2 + \hat{\tau}^2)$, where $\hat{\tau}^2$ is either the DL estimate or the Cochran ANOVA estimate, as study weights in the first step confirmed that multi-step estimators converge to the PM estimator. Hence, this indicates that convergence was not only a property of the selected data sets, and that convergence also occurred if the DL estimator was not used in the first step. Our findings are in agreement with the observation by DerSimonian and Kacker (2007) that two-step estimators better approximate the method of Paule and Mandel, and the conclusion by Bhaumik et al. (2012) that performance of the DL_2 and PM estimator are similar. This is because we have observed that DL_2 is the second step in an iterative scheme that takes us from $\hat{\tau}^2_{DL}$ to $\hat{\tau}^2_{PM}$.

Table 8.1. The DerSimonian-Laird (DL), two-step DerSimonian-Laird (DL_2), multi-step DerSimonian-Laird DL_k (where k refers to the kth step) and Paule-Mandel (PM) estimates for the three example data sets.

Estimate	Bangert-Drowns et al. (2004)	Sterne et al. (2001)	Ho and Lee (2012)
DL	0.0455	0.2239	0.0076
DL_2	0.0652	0.1587	0.0078
DL_3	0.0684	0.1841	0.0079
DL_4	0.0688	0.1736	0.0079
DL_5	0.0689	0.1778	
DL_6	0.0689	0.1761	
DL_7		0.1768	
DL_8		0.1765	
DL_9		0.1766	
DL_{10}		0.1766	
PM	0.0689	0.1766	0.0079

8.6 Proving (when convergence occurs) that the multi-step estimator converges to the Paule-Mandel estimator

As explained above, in addition to our three examples, many simulated datasets have shown that multi-step estimators converge to the PM estimator. In this section, we provide mathematical proofs to formally establish this limit. We will explain why it is not necessary that the DL estimator is used in the first step, so that our findings apply to multi-step estimators regardless of the nature of the estimation used in the first step.

8.6.1 Lemma: Agreement with respect to truncation to zero of the DerSimonian-Laird and Paule-Mandel estimators

We start by proving the Lemma that the DL and the PM estimators always agree in the sense that, for any given dataset, they are either both zero (if $Q_{GEN}(0) \leq (n-1)$) or both positive (if $Q_{GEN}(0) > (n-1)$). It is conceptually appealing that these two estimators agree in this way, and this is easily proved, but we do not think that this result has been stated previously.

Proof: If $Q_{GEN}(0) < (n-1)$, where $Q_{GEN}(\tau^2)$ is defined in equation (3), then the PM estimator is truncated to zero as explained in section 8.3.3. Furthermore, the first term in the numerator of equation (2) is also $Q_{GEN}(0)$ when the DL weights of $a_i = 1/\widehat{\sigma}_i^2$ are used. As noted in section 8.3.1, we then also have $\sum_{i=1}^{n} a_i \widehat{\sigma}_i^2 - \sum_{i=1}^{n} a_i^2 \widehat{\sigma}_i^2 / \sum_{i=1}^{n} a_i = (n-1)$ in the numerator of equation (2). Hence, the DL

estimator is also truncated to zero if $Q_{GEN}(0) < (n-1)$. If $Q_{GEN}(0) = (n-1)$ then, immediately from their estimating equations, both the DL and PM estimators are zero. Finally, if $Q_{GEN}(0) > (n-1)$ then no truncation for either estimator is required, so that the DL and PM estimators are both positive.

8.6.2 Proving that if convergence of the multi-step estimator occurs then it is to the Paule-Mandel estimate

Having established our Lemma, we will prove that the estimate of the multi-step estimator equals the PM estimate if convergence occurs. We will prove this first for cases where the convergence is to a positive estimate and then to an estimate of zero.

The case where the estimate converged to is positive Assume that convergence occurs and the resulting estimate is positive, so that $\hat{\tau}^2_{DL_{k+1}} = \hat{\tau}^2_{DL_k} = \hat{\tau}^2 > 0$. We substitute $\hat{\tau}^2_{DL_{k+1}} = \hat{\tau}^2_{DL_k} = \hat{\tau}^2$ into equation (6), where this equation correctly describes the iteration from DL_k to DL_{k+1} (because the estimate is positive and no truncation is necessary). Then solving the resulting equation for $Q_{GEN}(\hat{\tau}^2)$ results in

$$Q_{GEN}(\hat{\tau}^2) = \sum_{i=1}^{n} \frac{\hat{\sigma}_i^2 + \hat{\tau}^2}{\hat{\sigma}_i^2 + \hat{\tau}^2} - \frac{\sum_{i=1}^{n} (\hat{\sigma}_i^2 + \hat{\tau}^2)/(\hat{\sigma}_i^2 + \hat{\tau}^2)^2}{\sum_{i=1}^{n} 1/(\hat{\sigma}_i^2 + \hat{\tau}^2)} = (n-1)$$

which from equation (5) means that $\hat{\tau}^2 = \hat{\tau}^2_{PM}$.

The case where the estimate converged to is zero Assume that convergence occurs and the resulting estimate is either zero or truncated to zero, so that $\hat{\tau}^2_{DL_{k+1}} = \hat{\tau}^2_{DL_k} = \hat{\tau}^2 = 0$. If we substitute $\hat{\tau}^2_{DL_{k+1}} = \hat{\tau}^2_{DL_k} = \hat{\tau}^2 = 0$ into equation (7), the term in square brackets of (7) simplifies to $(n-1)$ and this equation becomes

$$0 = \max\left(0, \frac{Q_{GEN}(0) - (n-1)}{c}\right) \tag{8}$$

where $c = \sum_{i=1}^{n} 1/\hat{\sigma}_i^2 - \sum_{i=1}^{n} 1/\hat{\sigma}_i^4 / \sum_{i=1}^{n} 1/\hat{\sigma}_i^2 > 0$. Equation (8) is satisfied only if $Q_{GEN}(0) - (n-1) \leq 0$, from which the Lemma in section 8.6.1 implies that both the DL and PM estimators are zero (which is also the assumed value of $\hat{\tau}^2_{DL_{k+1}} = \hat{\tau}^2_{DL_k} = \hat{\tau}^2$). Hence, if the convergence of the multi-step estimator is to zero then the PM estimate is also zero, so that $\hat{\tau}^2 = \hat{\tau}^2_{PM}$.

Failure of convergence of the multi-step estimator Although we have observed convergence of the multi-step estimators in thousands of simulated datasets (see

https://osf.io/dpuzs/), it is possible to create examples where the multi-step estimator does not converge. As a concrete example of non-convergence, imagine a meta-analysis with four effect sizes $y_1 = -0.2$, $y_2 = 0.1$, $y_3 = -0.05$, and $y_4 = -0.3$, with corresponding $\sigma_1^2 = \sigma_2^2 = 0.01$ and $\sigma_3^2 = \sigma_4^2 = 0.2$. The DL estimate is $\hat{\tau}_{DL}^2 = 0.016$. Using this $\hat{\tau}_{DL}^2$ in estimating equation (4) gives $\hat{\tau}_{DL_2}^2 = 0$. Hence $\hat{\tau}_{DL_3}^2$ is then the usual DL estimator and, instead of achieving convergence, the multi-step estimator oscillates between 0.016 and 0, and does not converge to $\hat{\tau}_{PM}^2 = 0.0066$. The difficulties for achieving convergence in this example would appear to be due to the fact that the DL and PM estimates differ so substantially, and also because the within-study variances are of different magnitudes (so that $Q_{GEN}(\tau^2)$ is sensitive to the value of τ^2 when this is small). This example is a counterexample to the conjecture that the multi-step estimator always converges to the PM estimator.

Conclusions Regardless of whether or not the convergence of the multi-step estimator is to a positive estimate, we have proved that if convergence occurs then this is to the PM estimate. Simulating thousands of meta-analyses (see https://osf.io/dpuzs/) did not reveal the convergence problems suggesting that these problems only occur in rare cases such as the artificial one described above. We conclude that that, in practice, multi-step estimators converge to the PM estimate and also that they cannot converge to anything other than the PM estimate.

Although the finding that multi-step estimators may not converge reduces the utility of our analysis, our analytical results are more general than might be supposed, because it is not limited to using the DL estimator in the first step. All that is necessary for our results is that subsequent steps weight by the reciprocal of the estimated total study variances where the estimated between-study variance is the estimate at the previous step. Hence, our work establishes a link between multi-step estimators *per se* and the PM estimator rather than between just the DL$_k$ and PM estimators.

8.6.3 The relationship with an established Newton-Raphson method for calculating the Paule-Mandel estimate

DerSimonian and Kacker (2007) propose a Newton-Raphson algorithm for calculating the PM estimate (see their Appendix A). This algorithm sets $\hat{\tau}_{PM}^2$ to zero if $Q_{GEN}(0) \leq (n-1)$. If $Q_{GEN}(0) > (n-1)$, then $\hat{\tau}_{PM}^2 > 0$ and an initial value for the algorithm must be chosen. Then the Newton-Raphson algorithm takes $\hat{\tau}_{k+1}^2 = \hat{\tau}_k^2 + \Delta\,\hat{\tau}_{NR}^2$, where

$$\Delta\,\hat{\tau}_{NR}^2 = \frac{Q_{GEN}(\hat{\tau}_k^2) - (n-1)}{\sum_{i=1}^{n} \frac{1}{(\hat{\sigma}_i^2 + \hat{\tau}_k^2)^2} (y_i - \hat{\mu}\,(\hat{\tau}_k^2))^2} \tag{9}$$

where $\hat{\mu}(\hat{\tau}_k^2) = \sum_{i=1}^n y_i/(\hat{\sigma}_i^2 + \hat{\tau}_k^2) / \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}_k^2)$. Negative estimates are truncated to zero and the algorithm keeps iterating until convergence is reached. Jackson, Turner, Rhodes, and Viechtbauer (2014) explain how to generalize this Newton-Raphson procedure so that it can be applied to meta-regression models.

We can also calculate the corresponding $\Delta\hat{\tau}^2$ when using equation (6) in the iterative scheme that produces our multi-step estimators as $\Delta\hat{\tau}^2 = \hat{\tau}^2_{DL_{k+1}} - \hat{\tau}^2_{DL_k}$. From equation (6) this is

$$\Delta\hat{\tau}^2 = \frac{Q_{GEN}(\hat{\tau}^2_{DL_k}) - [\sum_{i=1}^n \hat{\sigma}_i^2/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k}) - \sum_{i=1}^n \hat{\sigma}_i^2/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})^2 / \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})]}{\sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k}) - \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})^2 / \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})} - \hat{\tau}^2_{DL_k}$$

Putting the right-hand side of the numerator over a common denominator results in

$$\Delta\hat{\tau}^2 = \frac{Q_{GEN}(\hat{\tau}^2_{DL_k}) - (n-1)}{\sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k}) - \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})^2 / \sum_{i=1}^n 1/(\hat{\sigma}_i^2 + \hat{\tau}^2_{DL_k})}. \quad (10)$$

Equation (10) also illustrates why the multi-step estimator converges to the PM estimator in practice. This is because the multi-step estimator converges if and only if $\Delta\hat{\tau}^2 = 0$, so that $Q_{GEN}(\hat{\tau}^2_{DL_k}) - (n-1) = 0$ and $\hat{\tau}^2_{DL_k} = \hat{\tau}^2_{PM}$. If instead the PM estimate has not been converged to, equation (10) shows that the estimator takes a step in the direction of the PM estimate in the kth step, because if $Q_{GEN}(\hat{\tau}^2_{DL_k}) < (n-1)$ then $\Delta\hat{\tau}^2 < 0$ and if $Q_{GEN}(\hat{\tau}^2_{DL_k}) > (n-1)$ then $\Delta\hat{\tau}^2 > 0$.

Comparing equations (9) and (10), we can also see that the iterative scheme for the multi-step estimator is closely related to the established Newton-Raphson method for calculating $\hat{\tau}^2_{PM}$. In the Appendix, we show that the expectation of the denominator of equation (9) under the model $y_i \sim N(\mu, \hat{\sigma}_i^2 + \hat{\tau}_k^2)$ and where the y_i are independent (where we suppress the distinction between $\hat{\tau}_k^2$ and $\hat{\tau}^2_{DL_k}$), is equal to the denominator of equation (10). This is reminiscent of the relationship between Fisher's scoring and Newton-Raphson methods in maximum likelihood estimation. This is because Fisher's scoring algorithm solves the likelihood based estimating equation by replacing the observed information in the denominator in a Newton-Raphson procedure by its expectation (the expected information). This observation provides us with intuition into why multi-step estimators tend towards the PM estimator as the number of steps becomes large.

8.7 The random-effects meta-regression model

For ease of exposition, we have presented our main results for random-effects meta-analyses, but these are readily extended to meta-regression models where study level covariate effects are included in the model. In order to establish that our results

generalise in this way, we consider meta-regression models with an arbitrary number of covariates in this section. All of the results in this section simplify to those shown previously.

The random-effects meta-regression model is an extension of model (1), where we assume that

$$y_i = \mathbf{x}_i\boldsymbol{\beta} + \mu_i + \epsilon_i$$

where $\mathbf{x}_i$ is the $1 \times p$ row vector of covariates associated with this study and $\boldsymbol{\beta}$ is the $p \times 1$ column vector of regression parameters of interest. Unless an intercept free regression is required, the first 'covariate' in each study is taken to be one to include the intercept. A matrix formulation of this standard model is

$$\mathbf{Y}|\mathbf{X} \sim N(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\Delta} + \tau^2\mathbf{I}) \tag{11}$$

where $\mathbf{Y}$ is a column vector containing the y_i, $\mathbf{X}$ is the $n \times p$ design matrix (sometimes referred to as the model matrix) whose ith row is $\mathbf{x}_i$, $\boldsymbol{\Delta} = \text{diag}(\widehat{\sigma}_i^2)$ and $\mathbf{I}$ is the $n \times n$ identity matrix. The parameter τ^2 in model (11) is called the residual between-study variance and describes the heterogeneity in the effect size estimates that is not explained by the covariates.

8.7.1 The general method of moments for meta-regression

Jackson et al. (2014) generalise the general method of moments (equation 2) to the meta-regression setting. They define $\mathbf{A} = \text{diag}(a_i)$, a diagonal matrix containing the weights, and $\mathbf{B} = \mathbf{A} - \mathbf{AX}(\mathbf{X}^t\mathbf{AX})^{-1}\mathbf{X}^t\mathbf{A}$. They also define the Q_a statistic

$$Q_a = \mathbf{Y}^t\mathbf{BY}$$

Jackson et al. (2014) use the subscript a to emphasise that the weights a_i are used, and so use the notation Q_a for this quadratic form. This Q_a statistic reduces to the the quadratic form in the numerator of equation (2) in the meta-analysis setting. Jackson et al. (2014) show that the meta-regression version of the generalised method of moments in equation (2) is

$$\hat{\tau}^2_{MM} = \frac{Q_a - \text{tr}(\mathbf{B}\boldsymbol{\Delta})}{\text{tr}(\mathbf{B})}. \tag{12}$$

where $\text{tr}(\cdot)$ denotes the trace of a matrix and $\text{tr}(\mathbf{B}) > 0$. As in the meta-analysis setting, we truncate $\hat{\tau}^2_{MM}$ when the solution to equation (12) is negative.

8.7.2 Paule-Mandel and DerSimonian-Laird estimators for meta-regression

The Paule-Mandel estimator The PM type estimator in the meta-regression setting proposed by Jackson et al. (2014) uses weights $a_i = 1/(\widehat{\sigma}_i^2 + \tau^2)$ when computing the Q_a statistic. We denote the resulting Q_a statistic using the notation $Q_{GEN}(\tau^2)$ in order to emphasise the dependence of the weights on the unknown parameter τ^2. This is a direct generalisation of $Q_{GEN}(\tau^2)$ in equation (3). Since $Q_{GEN}(\tau^2)$ follows a χ^2-distribution with $n-p$ degrees of freedom, the PM estimator is obtained by solving

$$Q_{GEN}(\hat{\tau}^2_{PM}) = n - p \tag{13}$$

If $Q_{GEN}(0) < n-p$ then, because for any given dataset $Q_{GEN}(\tau^2)$ is a monotonically decreasing continuous function in τ^2, there is no solution to this equation and we take $\hat{\tau}^2_{PM} = 0$ (Jackson et al., 2014). Following similar arguments as in the meta-analysis case, if $Q_{GEN}(0) \leq n-p$ then $\hat{\tau}^2_{PM} = 0$ and if $Q_{GEN}(0) > n-p$ then $\hat{\tau}^2_{PM} > 0$.

The DerSimonian and Laird estimator The standard weights of $a_i = 1/\widehat{\sigma}_i^2$ produce a DL type estimator of τ^2 when using equation (12), so that this estimator is just a special case of the general method of moments. We then have $\mathbf{A} = \mathbf{\Delta}^{-1}$ so that $\mathbf{B} = \mathbf{\Delta}^{-1} - \mathbf{\Delta}^{-1}\mathbf{X}(\mathbf{X}^t\mathbf{\Delta}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{\Delta}^{-1}$. Hence with these weights the numerator of equation (12) becomes $Q_{GEN}(0) - \text{tr}(\mathbf{B\Delta}) = Q_{GEN}(0) - \text{tr}(\mathbf{\Delta}^{1/2}\mathbf{B}\mathbf{\Delta}^{1/2})$, where this final equality is because $\text{tr}(\mathbf{CD}) = \text{tr}(\mathbf{DC})$, where $\mathbf{C}$ and $\mathbf{D}$ are square matrices of the same size, and because $\mathbf{\Delta}^{1/2}\mathbf{\Delta}^{1/2} = \mathbf{\Delta}$. We can then further simplify this expression by taking $\text{tr}(\mathbf{\Delta}^{1/2}\mathbf{B}\mathbf{\Delta}^{1/2}) = n - p$. This identity is because $\text{tr}(\mathbf{\Delta}^{1/2}\mathbf{B}\mathbf{\Delta}^{1/2}) = \text{tr}(\mathbf{I}) - \text{tr}(\mathbf{\Delta}^{-1/2}\mathbf{X}(\mathbf{X}^t\mathbf{\Delta}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{\Delta}^{-1/2})$, where $\text{tr}(\mathbf{I}) = n$ and $\text{tr}(\mathbf{\Delta}^{-1/2}\mathbf{X}(\mathbf{X}^t\mathbf{\Delta}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{\Delta}^{-1/2}) = p$. This final equality follows from the observation that the hat matrix corresponding to a design matrix $\mathbf{X}$ is given by $\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$, where $\text{tr}(\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t) = \text{tr}(\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1})$. For an identifiable regression $\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}$ is a $p \times p$ identity matrix, which results in the well known result that the trace of the hat matrix is p. Then we simply observe that $\mathbf{\Delta}^{-1/2}\mathbf{X}(\mathbf{X}^t\mathbf{\Delta}^{-1}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{\Delta}^{-1/2}$ is the hat matrix corresponding to the design matrix $\mathbf{\Delta}^{-1/2}\mathbf{X}$, so that its trace is also p. The numerator of equation (12) therefore simplifies to $Q_{GEN}(0) - (n-p)$ for the DL estimator.

8.7.3 Multi-step estimators for meta-regression

We can motivate multi-step estimators of τ^2 for meta-regression in exactly the same way as in meta-analysis. For example, using the DL estimator we first calculate $\hat{\tau}^2_{DL}$ using equation (12) and weights of $a_i = 1/\widehat{\sigma}_i^2$, truncating the estimate to zero if the solution is negative. We can then calculate $\hat{\tau}^2_{DL_2}$ using equation (12) and weights of $a_i = 1/(\widehat{\sigma}_i^2 + \hat{\tau}^2_{DL})$, from which we can then calculate $\hat{\tau}^2_{DL_3}$ and so on. In

general, we calculate $\hat{\tau}^2_{DL_{k+1}}$ using equation (12) with weights of $a_i = 1/(\widehat{\sigma}^2_i + \hat{\tau}^2_{DL_k})$. Any negative solutions are truncated to zero. This process generalises the multi-step estimators for meta-analysis described in section 8.4.

Let $\mathbf{A}_{\hat{\tau}^2_{DL_k}} = (\boldsymbol{\Delta} + \hat{\tau}^2_{DL_k}\,\mathbf{I})^{-1}$ denote the diagonal matrix containing the weights when computing the $(k+1)$th step DL estimator, for $k \geq 1$. Let $\mathbf{B}_{\hat{\tau}^2_{DL_k}}$ denote the corresponding matrix $\mathbf{B}$ computed using $\mathbf{A}_{\hat{\tau}^2_{DL_k}}$. From equation (12) we can then write

$$\hat{\tau}^2_{DL_{k+1}} = \frac{Q_{GEN}(\hat{\tau}^2_{DL_k}) - \mathrm{tr}\left(\mathbf{B}_{\hat{\tau}^2_{DL_k}}\boldsymbol{\Delta}\right)}{\mathrm{tr}\left(\mathbf{B}_{\hat{\tau}^2_{DL_k}}\right)} \tag{14}$$

for $k \geq 1$, where we truncate the resulting estimate to zero if the solution is negative. Equation (14) is a direct generalisation of equation (6) for meta-regression. Written explicitly in terms of the truncation, the $(k+1)$th step estimator is

$$\hat{\tau}^2_{DL_{k+1}} = \max\left(0, \frac{Q_{GEN}(\hat{\tau}^2_{DL_k}) - \mathrm{tr}\left(\mathbf{B}_{\hat{\tau}^2_{DL_k}}\boldsymbol{\Delta}\right)}{\mathrm{tr}\left(\mathbf{B}_{\hat{\tau}^2_{DL_k}}\right)}\right), \tag{15}$$

and equation (15) is a direct generalisation of equation (7).

8.7.4 Lemma: Agreement with respect to truncation to zero of the DerSimonian-Laird and Paule-Mandel estimators

In this section, we generalise the Lemma for the univariate meta-analysis to the meta-regression model. As explained above, the PM estimator is positive if and only if $Q_{GEN}(0) > n - p$. As also explained above, the numerator of equation (12) simplifies to $Q_{GEN}(0) - (n-p)$ when using the DL estimator ($a_i = 1/\,\widehat{\sigma}^2_i$). Hence, the DL estimator is also positive if and only if $Q_{GEN}(0) > n - p$. If instead $Q_{GEN}(0) \leq n - p$ then both the DL and PM estimators are zero. We therefore have established that the type of weak agreement described in section 8.6.1 also applies in the meta-regression setting.

8.7.5 Proving that if convergence occurs then it is to the Paule-Mandel estimate

The case where the estimate converged to is positive Assume that convergence occurs and the resulting estimate is positive, so that $\hat{\tau}^2_{DL_{k+1}} = \hat{\tau}^2_{DL_k} = \hat{\tau}^2 > 0$. We substitute $\hat{\tau}^2_{DL_{k+1}} = \hat{\tau}^2_{DL_k} = \hat{\tau}^2$ into equation (14), where this equation correctly

describes the iteration from DL_k to DL_{k+1} (because the estimate is positive and no truncation is necessary). Then solving the resulting equation for $Q_{GEN}(\hat{\tau}^2)$ results in

$$Q_{GEN}(\hat{\tau}^2) = \text{tr}\left(\mathbf{B}_{\hat{\tau}^2}(\boldsymbol{\Delta} + \hat{\tau}^2\,\mathbf{I})\right) = (n-p)$$

where the final equality follows from an argument involving a hat matrix that is very similar to the one made in section 8.7.2. Equation (13) implies that $\hat{\tau}^2 = \hat{\tau}^2_{PM}$.

The case where the estimate converged to is zero Assume that convergence occurs and the resulting estimate is either zero or truncated to zero, so that $\hat{\tau}^2_{DL_{k+1}} = \hat{\tau}^2_{DL_k} = \hat{\tau}^2 = 0$. If we substitute $\hat{\tau}^2_{DL_{k+1}} = \hat{\tau}^2_{DL_k} = \hat{\tau}^2 = 0$ into equation (15) then this equation becomes

$$0 = \max\left(0, \frac{Q_{GEN}(0) - (n-p)}{c}\right)$$

where $c = \text{tr}(\mathbf{B}_0) > 0$. Equation (8) is satisfied only if $Q_{GEN}(0) - (n-p) \leq 0$, from which the Lemma in section 8.7.4 implies that both the DL and PM estimators are zero (which is also the assumed value of $\hat{\tau}^2_{DL_{k+1}} = \hat{\tau}^2_{DL_k} = \hat{\tau}^2$). Hence, if the convergence of the multi-step estimator is to zero then the PM estimate is also zero, so that $\hat{\tau}^2 = \hat{\tau}^2_{PM}$. We have therefore established that multi-step estimates also converge to the PM estimator in meta-regression models.

8.8 Discussion

Two-step estimators have recently been presented as estimators of the between-study variance. We have extended these two-step estimators to a multi-step estimator and show by means of empirical examples, simulations, and also analytically that the multi-step estimator converges to the PM estimator if the number of steps is sufficiently large. This convergence occurs quickly in practice. Although examples can be produced where the multi-step estimator does not converge, we have shown that the PM estimator is obtained in the limit when convergence is obtained, and that convergence problems seldom occur in practice. Hence, our analysis suggests that the two-step estimators are better conceptualized as part of the usual iterative scheme that is used to calculate estimates using the PM estimator. Our findings also clarify why previous work (Bhaumik et al., 2012; DerSimonian & Kacker, 2007) observed that the DL_2 estimator was closer to the PM estimator than the DL estimator. We therefore suggest that the two-step estimators, as well as the proposed multi-step estimator, are not seen as truly distinct estimators but as steps in an iterative procedure that results in the PM estimator.

Now that REML and the PM estimator are computationally feasible and established in standard software, we align ourselves with those who argue that these

estimators should be preferred over the DL estimator (Langan et al., 2016; Veroniki et al., 2016). The case for REML becoming the default estimation method is now strong. However, the PM estimator is a viable alternative that is currently the best estimator that uses the method of moments. An advantage of the PM estimator compared to REML is that, in a small proportion of meta-analyses, REML suffers from convergence problems (Kontopantelis et al., 2013). A byproduct of our work is the development of a new iterative scheme that can be used to calculate the PM estimator.

Our work is a good example of scientists exploring an issue of interest with the expectation of discovering something new and then making new, but unanticipated, discoveries. However, discovering the link between the multi-step and PM estimator is in some respects even more satisfying than inventing a new class of estimators of the between-study variance. We have already explained that the PM estimator has been found to be equivalent to the empirical Bayes estimator, and our results provide another justification for the use of the PM estimator. This estimator would therefore seem to have a very wide variety of justifications and connections with other approaches which suggests that it has a useful role in both methodological and applied work.

We have considered the random-effects models for meta-analysis and meta-regression. Both of these models assume that the outcome data are independent. More sophisticated models that allow for correlated data include multivariate meta-analysis (Jackson, Riley, & White, 2011) and network meta-analysis (Salanti, 2012). Jackson, Veroniki, Law, Tricco, and Baker (2017) have already developed PM estimators for network meta-analysis, but our connection between multi-step and PM estimators provides an alternative possibility for motivating them. There is currently no PM estimator for the between-study covariance matrix in multivariate meta-analysis, but two extensions of the DL estimator have been proposed (Jackson, White, & Riley, 2013; Jackson, White, & Thompson, 2010). Generalising one or both of these estimators to allow an arbitrary set of weights, and so develop a general method of moments estimator, could then motivate the development of multi-step estimators in the context of multivariate meta-analysis. When convergence is reached as the number of steps becomes large, PM estimators of the between-study covariance matrix could then be defined in this limit. However, considerable methodological development is needed to extend our work to the network and multivariate meta-analysis settings, because this would first require the development of a generalised method of moments for correlated outcome data. We therefore leave this as a tantalising possibility for further work. However, enthusiasm for this idea is likely to be mitigated by the finding that the multi-step estimator does not always converge. Matters will become more complicated in the multivariate setting and some convention for defining a PM estimator in this way when convergence is not obtained would be needed.

To summarize, we have extended the two-step estimator so that multiple

steps can be used, and reproduced the PM estimator in the limit when the number of steps are sufficiently large. The PM estimator therefore has another justification as a result of its relationship with the proposed multi-step estimator. We suggest that the meta-analysis community should no longer consider the two-step and multi-step estimators to be truly distinct estimators, but should instead regard these type of estimators as approximate PM estimators.

8.9 Appendix

In this Appendix, we prove that

$$\mathbb{E}[\sum_{i=1}^{n} \frac{1}{(\widehat{\sigma}_i^2 + \hat{\tau}_k^2)^2} (y_i - \widehat{\mu}\,(\hat{\tau}_k^2))^2] = \sum_{i=1}^{n} \frac{1}{\widehat{\sigma}_i^2 + \hat{\tau}_k^2} - \frac{\sum_{i=1}^{n} \frac{1}{(\widehat{\sigma}_i^2 + \hat{\tau}_k^2)^2}}{\sum_{i=1}^{n} \frac{1}{\widehat{\sigma}_i^2 + \hat{\tau}_k^2}}.$$

under the model $y_i \sim N(\mu, \widehat{\sigma}_i{}^2 + \hat{\tau}_k^2)$, where all y_i are independent and $\widehat{\mu}\,(\hat{\tau}_k^2) = \sum_{i=1}^{n} y_i\,/(\widehat{\sigma}_i^2 + \hat{\tau}_k^2) / \sum_{i=1}^{n} 1\,/(\widehat{\sigma}_i^2 + \hat{\tau}_k^2)$.

To simplify the notation let $w_i = 1/(\widehat{\sigma}_i^2 + \hat{\tau}_k^2)$. Then the required expectation is

$$\mathbb{E}[\sum_{i=1}^{n} w_i^2\,(y_i - \widehat{\mu}\,(\hat{\tau}_k^2))^2] = \mathbb{E}[\sum_{i=1}^{n} w_i^2\,(y_i - \mu + \mu - \widehat{\mu}\,(\hat{\tau}_k^2))^2]$$

$$= \mathbb{E}[\sum_{i=1}^{n} w_i^2\,(y_i - \mu)^2] + \mathbb{E}[\sum_{i=1}^{n} w_i^2\,(\widehat{\mu}\,(\hat{\tau}_k^2) - \mu)^2] - 2\mathbb{E}[\sum_{i=1}^{n} w_i^2\,(y_i - \mu)(\widehat{\mu}\,(\hat{\tau}_k^2) - \mu)]. \qquad (17)$$

The first term in equation (17) is

$$\mathbb{E}[\sum_{i=1}^{n} w_i^2\,(y_i - \mu)^2] = \sum_{i=1}^{n} w_i^2\,\mathbb{E}[(y_i - \mu)^2] = \sum_{i=1}^{n} w_i^2\,\mathrm{Var}[y_i] = \sum_{i=1}^{n} w_i. \qquad (18)$$

The second term in equation (17) is

$$\mathbb{E}[\sum_{i=1}^{n} w_i^2\,(\widehat{\mu}\,(\hat{\tau}_k^2) - \mu)^2] = \sum_{i=1}^{n} w_i^2\,\mathbb{E}[(\widehat{\mu}\,(\hat{\tau}_k^2) - \mu)^2] = \mathrm{Var}[\widehat{\mu}\,(\hat{\tau}_k^2)] \sum_{i=1}^{n} w_i^2$$

where $\mathrm{Var}[\widehat{\mu}\,(\hat{\tau}_k^2)] = \mathrm{Var}[\sum_{i=1}^{n} w_i\,y_i / \sum_{i=1}^{n} w_i] = 1/\sum_{i=1}^{n} w_i$. Hence, the second term in (17) is equal to

$$\frac{\sum_{i=1}^{n} w_i^2}{\sum_{i=1}^{n} w_i} \qquad (19)$$

The third term in equation (17) is

$$-2\mathbb{E}[\sum_{i=1}^{n} w_i^2\,(y_i-\mu)(\widehat{\mu}\,(\hat{\tau}_k^2)-\mu)] = -2\sum_{i=1}^{n} w_i^2\,\mathrm{Cov}\left[y_i, \frac{\sum_{j=1}^{n} w_j\,y_j}{w_+}\right]$$

where $w_+ = \sum_{i=1}^{n} w_i$ and we have used the definition of $\widehat{\mu}\,(\hat{\tau}_k^2)$ and use the summation over j to compute it, because $\widehat{\mu}\,(\hat{\tau}_k^2)$ takes the same value for all i in the summation. The y_i are independent so that all covariances in the above summation are zero unless $i = j$. Hence, the third term is

$$-2\sum_{i=1}^{n} w_i^2\,\mathrm{Cov}[y_i, \frac{\sum_{j=1}^{n} w_j\,y_j}{w_+}] = -2\sum_{i=1}^{n} w_i^2\,\mathrm{Cov}[y_i, \frac{w_i y_i}{w_+}] = -2\frac{\sum_{i=1}^{n} w_i^2\,\mathrm{Cov}[y_i, w_i y_i]}{w_+}$$

where $\mathrm{Cov}[y_i, w_i y_i] = w_i \mathrm{Cov}(y_i, y_i) = w_i \mathrm{Var}(y_i) = 1$ so that the third term is

$$-2\frac{\sum_{i=1}^{n} w_i^2}{w_+}. \tag{20}$$

The summation of equations (18), (19) and (20), recalling that $w_i = 1/(\widehat{\sigma}_i^2 + \hat{\tau}_k^2)$, gives the required expectation.

CHAPTER 9

Statistical properties of methods based on the *Q*-statistic for constructing a confidence interval for the between-study variance in meta-analysis

Abstract

The effect sizes of studies included in a meta-analysis do often not share a common true effect size due to differences in for instance the design of the studies. Estimates of this so-called between-study variance are usually imprecise. Hence, reporting a confidence interval together with a point estimate of the amount of between-study variance facilitates interpretation of the meta-analytic results. Two methods that are recommended to be used for creating such a confidence interval are the *Q*-profile and generalized *Q*-statistic method that both make use of the *Q*-statistic. These methods are exact if the assumptions underlying the random-effects model hold, but these assumptions are usually violated in practice such that confidence intervals of the methods are approximate rather than exact confidence intervals. We illustrate by means of two Monte-Carlo simulation studies that coverage probabilities of both methods can be substantially below the nominal coverage rate in situations that are representative for meta-analyses in practice. We also show that these too low coverage probabilities are caused by violations of the assumptions of the random-effects model (i.e., normal sampling distributions of the effect size measure and known sampling variances), and are especially prevalent if the sample sizes in the primary studies are small.

This chapter is submitted as van Aert, R. C. M., van Assen, M. A. L. M., & Viechtbauer, W. (2017). Statistical properties of methods based on the *Q*-statistic for constructing a confidence interval for the between-study variance in meta-analysis.

Meta-analysis refers to a set of statistical techniques for combining the estimates of similar studies providing commensurable evidence about some phenomenon of interest (e.g., the effectiveness of a treatment, the size of a group difference, or the strength of the association between two variables). By combining the evidence, we aim to increase statistical power to find effects or relationships that individual studies may fail to detect. Moreover, by examining the variability in the estimates, we can draw more generalizable conclusions about the consistency of the effect or relationship over multiple studies and/or examine the degree to which effects or relationships vary and under what conditions.

If the included studies in a meta-analysis share the same common true effect size, any differences between the studies' effect size estimates are in theory only caused by sampling variability. However, the true effect sizes can also vary and sampling variability alone can then not explain the differences in effect size estimates. The effect sizes are then said to be heterogeneous. Such between-study variance may be due to systematic differences between the studies (e.g., differences in the sample characteristics or differences in the length or dose of a treatment). If information on how the studies differ is available, it may be possible to account for the between-study variance by incorporating this information in the model with a meta-regression analysis (Borenstein et al., 2009).

The *Q*-test (Cochran, 1954) is commonly used to test the null hypothesis of no between-study variance. A drawback of the *Q*-test is that the test can have low statistical power if a small number of studies are included and can have very high power if a large number of studies are included even if the amount of variability in the true effects is negligible (Hardy & Thompson, 1998; Higgins et al., 2003; Viechtbauer, 2007c). These undesirable statistical properties of the *Q*-test call attention to the importance for estimating the amount of between-study variance. The amount of between-study variance as well as the average effect size of the set of studies can be estimated by means of a random-effects model. Estimating the between-study variance is equally important as estimating the average effect size because it indicates the amount of consistency among the effects (Higgins et al., 2009). However, estimates of the between-study variance are rather imprecise if the number of studies in a meta-analysis is small (Chung et al., 2013; Kontopantelis et al., 2013; Sidik & Jonkman, 2007). Hence, reporting a confidence interval (CI) around the estimate is highly desirable and improves interpretability (Higgins et al., 2009; Ioannidis et al., 2007; Kepes et al., 2013; Langan et al., 2016).

Numerous methods for constructing a CI around the estimate of the between-study variance have been proposed, including the profile likelihood method(Hardy & Thompson, 1996), Wald-type methods (Biggerstaff & Tweedie, 1997), bootstrapping (Switzer, Paese, & Drasgow, 1992; Turner, Omar, Yang, Goldstein, & Thompson, 2000), a method by Sidik and Jonkman based on weighted least squares estimation (Sidik & Jonkman, 2005), the *Q*-profile method (Viechtbauer, 2007b), two different methods

that approximate the distribution of the test statistic of the *Q*-test (Biggerstaff & Jackson, 2008; Biggerstaff & Tweedie, 1997; Jackson, 2013), and also Bayesian methods to estimate a corresponding credible interval (Smith, Spiegelhalter, & Thomas, 1995). Since the method proposed by Biggerstaff and Jackson (2008) is a special case of the method described by Jackson (2013), we will refer to this method as the generalized *Q*-statistic method (GENQ method for short). A recent review of the aforementioned methods (Veroniki et al., 2016) recommended to use the *Q*-profile method if the between-study variance is large and the GENQ method if the between-study variance is small.

The *Q*-profile and GENQ method make use of the distribution of the test statistic of the *Q*-test to compute a CI. If the assumptions underlying the random-effects model hold, the null distribution of the *Q*-statistic is χ^2 with the number of studies minus one as the degrees of freedom (Cochran, 1954). However, violations of these assumptions are likely to occur in practice. For instance, an assumption of the random-effects model is that the sampling distribution of each study's effect size is normally distributed. This assumption is violated in most meta-analyses because the sampling distribution of most effect size measures is only asymptotically normal (i.e., approximates a normal distribution as the primary study's sample size gets large) (Hardy & Thompson, 1998; Hoaglin, 2016a, 2016b). Another assumption is that the sampling variances are known whereas they are usually estimated and then simply assumed to be known (Biggerstaff & Tweedie, 1997; Raudenbush, 2009). These assumptions become more acceptable if the primary studies' sample sizes increase, because the sampling distributions are then better approximated by normal distributions and the primary studies' observed sampling variances are closer to the true sampling variances. Nevertheless, violations of the assumptions of the random-effects model will result in a *Q*-statistic that does not exactly follow a χ^2 distribution under the null hypothesis. Hence, the *Q*-profile and GENQ method may not yield exact CIs (i.e., coverage probability equal to 1-α) if these assumptions do not hold.

The aim of this chapter is to study the performance of the *Q*-profile and GENQ method under conditions that are representative for meta-analyses in practice. We selected the log odds ratio as the effect size measure in our analyses, because it is often used in medical research. Note that the above discussed assumptions of normal sampling distributions and known sampling variances are violated if the log odds ratio is the effect size measure, and that these violations can be substantial particularly if the primary studies' sample sizes are small. The statistical properties of the *Q*-profile method have already been examined under conditions that are representative for meta-analyses in practice where the assumptions of the random-effects model are violated (Viechtbauer, 2007b). However, statistical properties of the GENQ method have only been studied under conditions where all assumptions of the random-effects model hold (Jackson, 2013; Jackson, Bowden, & Baker, 2015). This chapter is therefore the first that compares the statistical properties of the *Q*-profile and GENQ method

when the assumptions of the random-effects model do not hold in combination with conditions that are representative for meta-analysis in practice.

This chapter continues by briefly outlining the random-effects model and the *Q*-test. Subsequently, the *Q*-profile and GENQ method are described. Next, we describe the Monte-Carlo simulation study that we use to examine the statistical properties of the two methods and present their results. This chapter ends with a conclusion and discussion section with recommendations for when to use the *Q*-profile and GENQ method.

9.1 The random-effects model and *Q*-test

Assume that $i = 1, 2, \ldots k$ independent effect sizes have been derived from a set of studies. Each study's observed effect size (Y_i) is assumed to be an unbiased estimate of the study specific true effect size (θ_i). However, Y_i is not equal to θ_i due to sampling error (ε_i). This can be written as

$$Y_i = \theta_i + \varepsilon_i,$$

where $\varepsilon_i \sim N(0, \sigma_i^2)$ with σ_i^2 denoting the true sampling variance in the ith study. All ε_i are assumed to be independent of each other and each σ_i^2 is estimated in practice and then assumed to be known. Hence, we will write $\hat{\sigma}_i^2$ to refer to the estimated sampling variances. Each θ_i consists of an average true effect (μ) and the random effect $u_i \sim N(0, \tau^2)$ that denotes the difference between θ_i and μ (Raudenbush, 2009). Hence, the random-effects model can be written as

$$Y_i = \mu + u_i + \varepsilon_i,$$

where it is assumed that the u_i are independent of each other and u_i is independent of ε_i. The random-effects model reduces to the common- or fixed-effects model if $\tau^2 = 0$.

Several hypothesis tests for testing H_0: $\tau^2 = 0$ have been proposed (Viechtbauer, 2007c), of which the *Q*-test is most often used (Hoaglin, 2016b). The *Q*-statistic is computed with

$$Q = \sum_{i=1}^{k} \frac{(Y_i - \hat{\theta})^2}{\hat{\sigma}_i^2}, \quad (1)$$

where $\hat{\theta}$ is given by

$$\hat{\theta} = \frac{\sum_{i=1}^{k} w_i Y_i}{\sum_{i=1}^{k} w_i}, \tag{2}$$

with $w_i = 1/\hat{\sigma}_i^2$. Under the null hypothesis, Q follows a χ^2 distribution with $k - 1$ degrees of freedom if the primary studies' sample sizes are large (Cochran, 1954).

9.2 *Q*-profile method

The *Q*-profile method generalizes the *Q*-statistic in Equation 1 to a random-effects model by incorporating τ^2, so that

$$Q(\tau^2) = \sum_{i=1}^{k} \frac{(Y_i - \hat{\mu})^2}{\tau^2 + \hat{\sigma}_i^2}, \tag{3}$$

with $\hat{\mu}$ given by Equation 2 with $w_i = 1/(\tau^2 + \hat{\sigma}_i^2)$. This generalized version of the *Q*-statistic also follows a χ^2 distribution with $k - 1$ degrees of freedom (Viechtbauer, 2007b) and is a function of τ^2. Hence, a CI for τ^2 can be obtained by means of test inversion (Casella & Berger, 2002). If $\chi^2_{k-1;0.025}$ and $\chi^2_{k-1;0.975}$ are the 2.5th and 97.5th percentiles of a χ^2 distribution with $k - 1$ degrees of freedom, the 95% CI ($\hat{\tau}^2_{LB}; \hat{\tau}^2_{UB}$) is equal to the two values for τ^2 where

$$(Q(\tau^2 = \hat{\tau}^2_{LB}) = \chi^2_{k-1;0.975}; Q(\tau^2 = \hat{\tau}^2_{UB}) = \chi^2_{k-1;0.025}).$$

The method is called *Q*-profile because different values for τ^2 are entered in Equation 3 (i.e., profiling) until the generalized *Q*-statistic equals the critical values of the χ^2 distribution. If $Q(\tau^2 = 0) < \chi^2_{k-1;0.975}$, the lower bound of the CI is in principle negative but outside of the parameter space and hence truncated to zero (Viechtbauer, 2007b). If $Q(\tau^2 = 0) < \chi^2_{k-1;0.025}$, the estimate of the upper bound is also negative, and the CI is set equal to the null set. Under the assumptions of the random-effects model (i.e., unbiased observed effect size estimates, normal sampling distributions, known sampling variances, and uncorrelated sampling errors and random effects), the *Q*-profile method yields exact CIs. Viechtbauer (2007a) showed by means of a simulation study with log odds ratios as effect size measure (which do not fulfill the model assumptions exactly) that the *Q*-profile method still yields accurate coverage probabilities for the majority of the conditions included in the simulations. One exception was that undercoverage occurred when meta-analyzing a large number of studies with small sample sizes.

9.3 Generalized *Q*-statistic method

The generalized *Q*-statistic (GENQ) method (Biggerstaff & Jackson, 2008; Jackson, 2013) constructs a CI for τ^2 based on the exact distribution of the *Q*-statistic under the assumptions of the random-effects model. This method uses the generalized form of the *Q*-statistic as described by DerSimonian and Kacker (2007) where the weights are no longer $w_i = 1/\hat{\sigma}_i^2$, but could be any set of positive constants denoted by a_i. The exact distribution of the *Q*-statistic (Q_a) was derived by Biggerstaff and Jackson (2008, p. 6095) and Jackson (2013, p. 222). The distribution of Q_a is the weighted sum (weighted by $\lambda_i \geq 0$ where λ_i are the eigenvalues of a matrix that is a function of a_i, $\hat{\sigma}_i^2$, and τ^2) of mutually independent χ^2-distributed random variables with one degree of freedom each, so that

$$Q_a = \sum_{i=1}^{k}{}^{d} \lambda_i \chi_i^2(1). \tag{4}$$

Jackson (2013) proved that the cumulative distribution function of Q_a is a continuous and decreasing function in τ^2. The cumulative distribution function of a positive linear combination of χ^2-distributed random variables can be obtained by Farebrother's algorithm (Farebrother, 1984). The lower and upper bound of the 95% CI ($\hat{\tau}_{LB}^2; \hat{\tau}_{UB}^2$) can then be obtained again by test inversion (Casella & Berger, 2002), that is, given the observed value q_a of Q_a, we find those two values of τ^2 for which

$$(P(Q_a \geq q_a; \tau^2 = \hat{\tau}_{LB}^2) = 0.025; P(Q_a \geq q_a; \tau^2 = \hat{\tau}_{UB}^2) = 0.975).$$

The upper and lower bounds of the CI can also be negative. If the estimate of the lower bound is negative, it is recommended to truncate the estimate to zero. In case the lower and upper bounds are both negative, the CI is set equal to the null set. The GENQ method yields exact CIs if the assumptions underlying the random-effects model (i.e., unbiased observed effect size estimates, normal sampling distributions, known sampling variances, and uncorrelated standard errors and random effects) are fulfilled.

Different values for a_i can be selected for weighting the observed effect sizes. If $a_i = 1/\hat{\sigma}_i^2$, the results of the methods by Biggerstaff and Jackson (2008) and Jackson (2013) are equivalent. Other suggestions for a_i are an unweighted analysis with a_i equal to a constant, $1/(\hat{\tau}^2 + \hat{\sigma}_i^2)$, and $1/(\hat{\tau}^2 + \hat{\sigma}_i^2)^{0.5}$ (Jackson, 2013; Jackson et al., 2014). Note that, even when all model assumptions are fulfilled, the CIs are no longer

exact if the last two weights are used, because the weights are then a function of a random variable (since τ^2 has to be estimated).

9.4 Monte-Carlo simulation study 1

The *Q*-profile and GENQ method both yield exact CIs under the assumptions of the random-effects model. However, these assumptions usually do not hold in practice, but become more acceptable if the primary studies' sample sizes increase. Hence, the generalized *Q*-statistics that are used for constructing the CIs with the *Q*-profile and GENQ method only approximate a χ^2 distribution if the primary studies' sample sizes are large (Cochran, 1954), and therefore the CIs are really just approximations in practice instead of exact CIs. We will study the statistical properties of the CIs obtained with the *Q*-profile and GENQ method by means of two Monte-Carlo simulation studies with the log odds ratio as effect size measure whose sampling distribution is only well-approximated by a normal distribution for large sample sizes in the primary studies.

Data in both simulation studies were generated by first drawing the true log odds ratios, θ_i for *i*=1, ..., *k*, from $N(\mu, \tau^2)$, with μ denoting the mean of the distribution of the studies' true effect sizes and τ^2 the variance of this distribution. Based on the sampled θ_i, *k* 2x2 frequency tables were simulated by first generating the number of cases with the outcome of interest in the control group (x_i^C). A value for x_i^C was sampled from a binomial distribution with n_i^C being the sample size of the control group and probability π_i^C for the outcome of interest in the control group. A study's true log odds ratio (θ_i) and π_i^C were used for computing the probability of the outcome of interest in the experimental group with $\pi_i^E = \pi_i^C \exp(\theta_i) / [1 - \pi_i^C + \pi_i^C \exp(\theta_i)]$. The number of cases with the outcome of interest in the experimental group, x_i^E, was sampled from a binomial (n_i^E, π_i^E) distribution with n_i^E being the total number of cases in the experimental group. Before computing the observed log odds ratio and corresponding sampling variance for each study, 0.5 was added to each cell of the frequency tables to decrease bias in the estimator of the log odds ratio (Walter & Cook, 1991). Furthermore, this adjustment allows calculation of the log odds ratio and its sampling variance in case of zero cells. Therefore, the observed log odds ratio was computed with

$$Y_i = \log\left[\frac{x_i^E + 0.5}{n_i^E - x_i^E + 0.5} \Big/ \frac{x_i^C + 0.5}{n_i^C - x_i^C + 0.5} \right]$$

and its observed sampling variance with

$$\hat{\sigma}_i^2 = \frac{1}{x_i^E + 0.5} + \frac{1}{n_i^E - x_i^E + 0.5} + \frac{1}{x_i^C + 0.5} + \frac{1}{n_i^C - x_i^C + 0.5}.$$

The Y_i and $\hat{\sigma}_i^2$ values were used as input for the *Q*-profile and GENQ method. Two different weights were used for applying the GENQ method, $a_i = 1/\hat{\sigma}_i^2$ and $a_i = 1/\hat{\sigma}_i$ These two weights were selected because the GENQ method yields exact CIs for these two weights if the assumptions underlying the random-effects model hold.

Values for the true effect size (μ) in the first simulation study were 0, 0.25, 0.5, 0.75, and 1. The amount of between-study heterogeneity (τ) was varied between 0 and 0.5 with steps equal to 0.1, and three fixed values for π_i^C were selected: 0.1, 0.3, and 0.5. For the condition with large heterogeneity ($\tau = 0.5$) and $\mu = 0$, the 95% prediction interval for θ_i ranges from -0.980 to 0.980, corresponding to odds ratios of 2.66 in favor of the control group to 2.66 in favor of the experimental group. The total number of observed effect sizes in a meta-analysis (*k*) was 5, 10, 20, 40, and 160. Values for *k* are in line with previous Monte-Carlo simulation studies (Jackson, 2013; Viechtbauer, 2007b) that examined the statistical properties of the *Q*-profile and GENQ method. We also included the condition *k*=160 to examine the statistical properties of the methods for a very large number of studies. The sample size in the control and experimental group in each study was set equal to each other, but sample sizes were allowed to differ across the studies within a meta-analysis. Sample sizes per group (30, 50, 100, 150, and 300) were replicated *k*/5 times in each meta-analysis in order to hold the average sample size of the studies constant across conditions.

The outcome variables in our simulation study were the coverage probability (how often is τ^2 in the CI of the *Q*-profile and GENQ method), the average width of the CI, the standard deviation of the width of the CI over all replications, and the number of times the width of a particular method's CI was larger than the width of the other methods. The simulations were programmed in R (R Core Team, 2017) with 10,000 replications per condition. The "parallel" package (R Core Team, 2017) was used to parallelize the computations and the "metafor" package(Viechtbauer, 2010) was used for applying the *Q*-profile and GENQ method. R code of this simulation study is available via: https://osf.io/3x5rg/.

9.4.1 Results Monte-Carlo simulation study 1

We only present the results for $\mu = 0$, *k* = (5, 10, 40, 160), and $\pi_i^C = (0.1, 0.5)$, because these conditions are illustrative for the performance of the methods. Results were hardly affected by the selected values of μ, whereas results for $\pi_i^C = 0.3$ were in between the two other conditions of π_i^C. Results of all other conditions are

available via https://osf.io/qjv5x/. We will refer to the two different weights used for the GENQ method as 'variance weights' for $a_i = 1/\hat{\sigma}_i^2$ and 'standard error weights' for $a_i = 1/\hat{\sigma}_i$. Figure 9.1 shows the coverage probabilities of these two methods and the *Q*-profile method as a function of true heterogeneity τ. The solid lines refer to coverage probabilities for $\pi_i^C = 0.5$ and the dashed lines to the coverage probabilities for $\pi_i^C = 0.1$. Coverage probabilities of the *Q*-profile method are indicated with triangles, the GENQ method with variance weights with plus signs, and the GENQ method with standard error weights with crosses. Note that we concluded that τ was not included in the CI if a CI was equal to the null set. Hence, coverage probabilities of the methods equal to 0.95 indicate nominal coverage for all conditions.

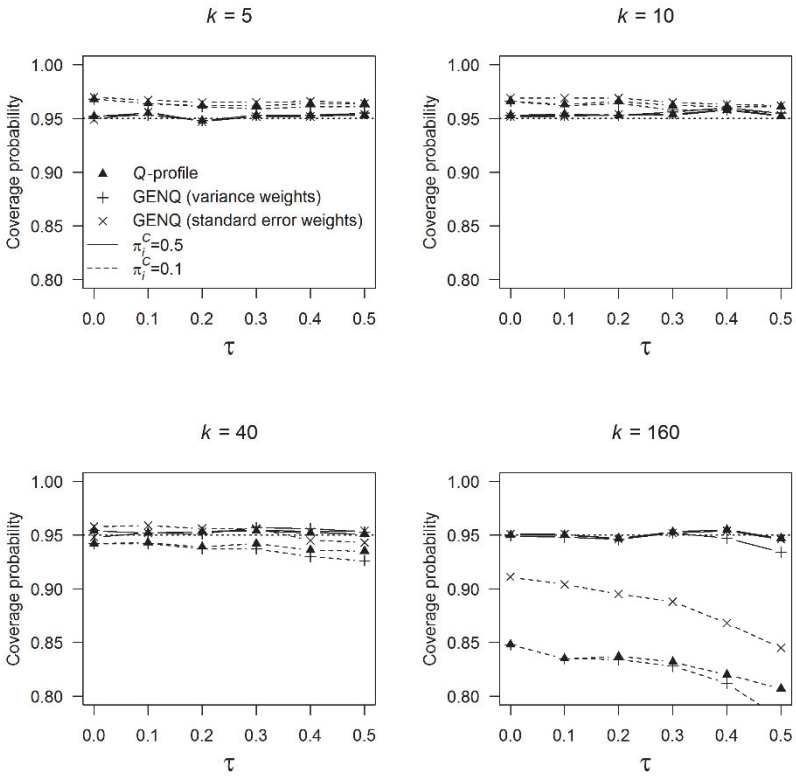

Figure 9.1. Coverage probabilities of the *Q*-profile method, GENQ method with variance weights ($a_i = 1/\hat{\sigma}_i^2$), and GENQ method with standard error weights ($a_i = 1/\hat{\sigma}_i$). The probability of the outcome of interest in the control group is denoted by π_i^C, the number of primary studies in a meta-analysis with *k*, and the amount of between-study heterogeneity with τ.

For all values of *k*, coverage probabilities of the *Q*-profile and GENQ methods for $\pi_i^C = 0.5$ were equal or close to 0.95. However, coverage of the methods for $\pi_i^C = 0.1$ and *k*=5 or 10 was slightly too large especially for τ=0. Since coverage probabilities decreased when *k* was increased, coverage probabilities were reasonably close to the nominal coverage rate for *k*=40 and $\pi_i^C = 0.1$, but severe undercoverage was observed for *k*=160.

The lowest coverage probability for all methods was obtained in the condition *k*=160, $\pi_i^C = 0.1$, and $\tau = 0.5$; for *Q*-profile 0.807, GENQ with variance weights 0.779, and GENQ with standard error weights 0.845. For this condition, the undercoverage was fully explained by the upper bounds of the CIs being smaller than τ suggesting that the generalized *Q*-statistic was too low. This also explains why the undercoverage for $\pi_i^C = 0.1$ and *k*=160 was least severe for the GENQ method with standard error weights. Large (both positive and negative) effect sizes go together with unequally distributed cases in the 2x2 frequency table and thus large sampling variances. Equation 3 shows that effect sizes that deviate substantially from $\hat{\mu}$ have only a minimal contribution to the generalized *Q*-statistic because of their large sampling variance. If standard error weights are used instead of variance weights, more extreme effect sizes contribute more to the generalized *Q*-statistic resulting in larger values for this statistic. Hence, undercoverage was less severe for the GENQ method with standard error weights than with variance weights.

Table 9.1 presents the average and the standard deviation of the width of a method's CI over all replications. Bold values indicate the method with the smallest average width of the CI within a particular condition. As expected, the average width of the CIs decreased as a function of *k*. Coverage probabilities of the methods were in general close to the nominal coverage rate for $\pi_i^C = 0.5$, so the method with the smallest CI is preferred in this condition. The CI of the GENQ method with variance weights was the smallest for the majority of the conditions. With the exception of one condition (i.e., *k* = 20 and $\tau = 0.4$), the average width of the CI for $\pi_i^C = 0.5$ of the *Q*-profile method was larger than of the GENQ methods. However, the difference between the method with the smallest and largest average width of a CI was at most 0.1 for $\tau \leq 0.1$ and at most 0.05 for $\tau > 0.1$.

The standard deviations of the width of the methods' CIs over all replications were similar for $\pi_i^C = 0.5$ and *k* < 160; the method with the highest standard deviation never had a standard deviation that was more than twice as large as the standard deviation of the method with the smallest standard deviation. The width of the CIs obtained with the GENQ method with variance weights was in at most 92.6% and 99.2% of the conditions smaller than that of the *Q*-profile and GENQ method with standard error weights whereas the width of the CIs obtained with the *Q*-profile

Table 9.1. Average and standard deviation (in parentheses) of the confidence interval width of the *Q*-profile method, GENQ method with variance weights ($a_i = 1/\hat{\sigma}_i^2$), and GENQ method with standard error weights (SE; $a_i = 1/\hat{\sigma}_i$). The probability of having the outcome of interest in the control group is denoted by π_i^C, the number of primary studies in a meta-analysis with *k*, and the amount of between-study heterogeneity with τ.

		$\pi_i^C = 0.5$					
		τ=0	τ=0.1	τ=0.2	τ=0.3	τ=0.4	τ=0.5
k = 5	*Q*-profile	0.831 (0.392)	0.869 (0.393)	0.979 (0.413)	1.128 (0.43)	1.285 (0.462)	1.446 (0.489)
	GENQ (variance)	**0.747 (0.309)**	**0.797 (0.321)**	**0.933 (0.361)**	1.112 (0.401)	1.295 (0.455)	1.483 (0.512)
	GENQ (SE)	0.777 (0.33)	0.818 (0.338)	0.94 (0.365)	**1.104 (0.39)**	**1.272 (0.423)**	**1.44 (0.455)**
k = 20	*Q*-profile	0.46 (0.186)	0.487 (0.185)	0.564 (0.175)	0.641 (0.161)	**0.71 (0.154)**	0.783 (0.165)
	GENQ (variance)	**0.403 (0.139)**	**0.439 (0.144)**	**0.531 (0.142)**	**0.622 (0.132)**	0.711 (0.138)	0.798 (0.165)
	GENQ (SE)	0.434 (0.155)	0.462 (0.156)	0.547 (0.153)	0.635 (0.135)	**0.71 (0.119)**	**0.779 (0.127)**
k = 40	*Q*-profile	0.221 (0.075)	0.251 (0.068)	0.284 (0.041)	0.283 (0.03)	0.299 (0.03)	0.327 (0.034)
	GENQ (variance)	**0.199 (0.059)**	**0.23 (0.054)**	**0.266 (0.028)**	**0.269 (0.015)**	**0.294 (0.026)**	0.332 (0.035)
	GENQ (SE)	0.227 (0.068)	0.254 (0.064)	0.299 (0.041)	0.295 (0.023)	0.295 (0.015)	**0.319 (0.024)**
k = 160	*Q*-profile	0.13 (0.04)	0.161 (0.028)	0.137 (0.012)	0.133 (0.007)	0.143 (0.007)	0.157 (0.008)
	GENQ (variance)	**0.122 (0.034)**	**0.152 (0.024)**	**0.126 (0.011)**	**0.125 (0.003)**	0.14 (0.006)	0.159 (0.009)
	GENQ (SE)	0.145 (0.041)	0.174 (0.032)	0.158 (0.024)	0.134 (0.003)	**0.139 (0.004)**	**0.152 (0.006)**

Table 9.1 Continued

		$\pi_i^C = 0.1$					
		τ=0	τ=0.1	τ=0.2	τ=0.3	τ=0.4	τ=0.5
k = 5	Q-profile	1.399 (0.712)	1.412 (0.714)	1.488 (0.728)	1.587 (0.751)	1.732 (0.759)	1.889 (0.793)
	GENQ (variance)	**1.201 (0.5)**	**1.233 (0.518)**	**1.315 (0.537)**	**1.449 (0.589)**	**1.614 (0.627)**	**1.791 (0.682)**
	GENQ (SE)	1.276 (0.559)	1.297 (0.569)	1.372 (0.581)	1.486 (0.618)	1.639 (0.641)	1.807 (0.682)
k = 20	Q-profile	0.734 (0.311)	0.76 (0.314)	0.814 (0.317)	0.899 (0.316)	0.981 (0.306)	1.067 (0.306)
	GENQ (variance)	**0.626 (0.223)**	**0.655 (0.23)**	**0.718 (0.239)**	**0.814 (0.247)**	**0.912 (0.243)**	**1.005 (0.241)**
	GENQ (SE)	0.696 (0.256)	0.722 (0.259)	0.774 (0.264)	0.862 (0.272)	0.955 (0.266)	1.051 (0.262)
k = 40	Q-profile	0.317 (0.113)	0.337 (0.114)	0.395 (0.11)	0.455 (0.087)	0.474 (0.065)	0.477 (0.055)
	GENQ (variance)	**0.289 (0.094)**	**0.31 (0.096)**	**0.37 (0.093)**	**0.429 (0.069)**	**0.447 (0.042)**	**0.448 (0.028)**
	GENQ (SE)	0.348 (0.111)	0.365 (0.111)	0.417 (0.109)	0.485 (0.094)	0.526 (0.071)	0.525 (0.053)
k = 160	Q-profile	0.159 (0.06)	0.184 (0.063)	0.249 (0.05)	0.256 (0.032)	0.224 (0.016)	0.221 (0.012)
	GENQ (variance)	**0.153 (0.056)**	**0.178 (0.058)**	**0.243 (0.046)**	**0.245 (0.034)**	**0.208 (0.013)**	**0.205 (0.005)**
	GENQ (SE)	0.199 (0.067)	0.22 (0.067)	0.279 (0.059)	0.321 (0.042)	0.271 (0.039)	0.237 (0.012)

method was in at most 59.3% and 90% of the conditions smaller than that of the GENQ method with variance and standard error weights, respectively. To summarize the results for $\pi_i^C = 0.5$, the GENQ method with variance weights outperformed the other two methods for $\tau \leq 0.3$ in the majority of the conditions, and the GENQ method with standard error weights had the best statistical properties if $\tau > 0.3$ in the majority of the conditions.

Results for $\pi_i^C = 0.1$ are also presented in Table 9.1, but can hardly be interpreted. Coverage probabilities for these conditions often substantially deviated from the nominal coverage rate. Hence, drawing conclusions based on the width of a CI is not informative. Noteworthy though is that the GENQ method with variance weights always yielded smaller CIs than the *Q*-profile and GENQ method with standard error weights. Based on the results for $\pi_i^C = 0.1$, we conclude that the GENQ method with standard error weights performs best, because its undercoverage is considerably less than that of the other two methods.

We created heat maps to gain further insight into whether there is a specific set of conditions for *k*, τ, π_i^C, n_i^E, and n_i^C for which the coverage probability substantially diverges from the nominal coverage rate. For these conditions, researchers should be reluctant in applying these methods and interpreting their results. The heat maps show the coverage probabilities for different values of *k* (5, 10, 20, 40, 80, and 160) and π_i^C ranging from 0.01 to 0.5 at a fixed sample size of 30 in both groups (i.e., $n_i^E = n_i^C = 30$). We also created heat maps in the same conditions but with n_i^E and n_i^C both being equal to either 15, 30, 80, 160, 320, or 800 while fixing *k* to 20. The heat maps were created for each of the three methods for $\tau = 0$ and $\tau = 0.5$. The procedure for creating the heat maps as well as the heat maps themselves are available via https://osf.io/e35qc/.

The heat maps confirmed the results presented in Figure 9.1 that τ only had a small effect on the coverage probabilities of the methods. Coverage probabilities decreased if π_i^C decreased, and if undercoverage was present for a combination of π_i^C and sample size, then this undercoverage became more severe as *k* increased. Furthermore, coverage probabilities also decreased if the sample size decreased, because the sampling variances were then less accurately estimated. The maximum coverage probability was equal to 0.97, so no severe overcoverage was observed. Specifically, coverage probabilities of all three methods were acceptable (i.e., > 0.9) at a fixed sample size of 30 in both groups when *k* = 5 or 10 and $\pi_i^C \geq 0.05$, *k* = 20 and $\pi_i^C \geq 0.1$, *k* = 40 or 80 and $\pi_i^C \geq 0.2$, and *k* = 160 and $\pi_i^C \geq 0.35$. If *k* was fixed to 20, coverage probabilities were acceptable for $n_i^E = n_i^C = 15$ and $\pi_i^C \geq 0.2$, $n_i^E = n_i^C = 30$ and $\pi_i^C \geq 0.1$, and $n_i^E = n_i^C = 80$ and $\pi_i^C \geq 0.05$. The finding that coverage

probabilities deviate from the nominal coverage rate for low values of π_i^C and not for π_i^C close to 0.5 hints at a systematic bias that is caused by violated assumptions of the random-effects model in case of rare events in the primary studies. This bias will be examined in simulation study 2.

9.5 Monte- Carlo simulation study 2

Simulation study 1 showed that coverage probabilities of both the *Q*-profile and GENQ methods can substantially deviate from the nominal coverage rate. The goal of simulation study 2 was to examine the cause of under- and overcoverage by the methods that was apparent for $\pi_i^C = 0.1$, but not for $\pi_i^C = 0.5$.

The *Q*-profile and GENQ methods with the specified weights are exact if the assumptions underlying the random-effects model hold, so deviations from the nominal coverage rate in simulation study 1 were caused by violations of assumptions of the random-effects model. One of the assumptions that is violated is that the primary studies' sampling variances are not known but estimated, which particularly affects the methods' coverage if the studies' sample sizes are small. Hence, we set out to compare the methods' coverage rates and the distribution of the generalized *Q*-statistic used by the *Q*-profile method when the sampling variances are estimated as in simulation study 1 (denoted by $\hat{\sigma}^2$) and when the true variances are used.

In order to compute the true variances, we first created all possible 2x2 frequency tables based on n_i^C and n_i^E given a particular value for π_i^C and π_i^E. For example, this yields 31x31=961 possible frequency tables if the sample size in both groups was equal to 30. A selection of these 961 frequency tables is presented in Table 9.2 (first four columns). The probability of observing a particular frequency table (fifth column) was computed by multiplying $B(x^E; n^E, \pi^E)$ with $B(x^C; n^C, \pi^C)$ where *B* refers to the probability mass function of the binomial distribution. Log odds ratios (last column) were computed for each frequency table after adding 0.5 to each cell to reduce bias in the estimator of the log odds ratios(Walter & Cook, 1991) and to make computation of the log odds ratio possible in all tables, even those with zero cells. We used the probability of observing a frequency table and the log odds ratio for each frequency table for computing the expected value of the log odds ratio ($E[Y]$) and the true sampling variance ($\sigma_T^2 = E[Y^2] - E[Y]$). We expect that the methods' coverage probabilities computed with σ_T^2 will be closer than the nominal coverage rate than with $\hat{\sigma}^2$, because instead of using estimated sampling variances, the true variances are used. Differences between $\hat{\sigma}^2$ and σ_T^2 are especially prevalent if one of the cells in the observed frequency table is equal to 0.

Table 9.2. Selection of all possible 2x2 frequency tables, probabilities of observing a table $B(x^E; n^E, \pi^E) \times B(x^C; n^C, \pi^C)$ with *B* denoting the probability mass function of the binomial distribution, and log odds ratios (*Y*) if the sample size in the experimental and control group equals 30. Cell frequencies are denoted by x^E, $n^E - x^E$, x^C, and $n^C - x^C$.

x^E	$n^E - x^E$	x^C	$n^C - x^C$	$B(x^E; n^E, \pi^E) \times B(x^C; n^C, \pi^C)$	Y
0	30	0	30	$B(0;30, \pi^E) \times B(0;30, \pi^C)$	0
1	29	0	30	$B(1;30, \pi^E) \times B(0;30, \pi^C)$	1.132
2	28	0	30	$B(2;30, \pi^E) \times B(0;30, \pi^C)$	1.677
3	27	0	30	$B(3;30, \pi^E) \times B(0;30, \pi^C)$	2.049
4	26	0	30	$B(4;30, \pi^E) \times B(0;30, \pi^C)$	2.338
⋮	⋮	⋮	⋮	⋮	⋮
0	30	1	29	$B(0;30, \pi^E) \times B(1;30, \pi^C)$	-1.132
⋮	⋮	⋮	⋮	⋮	⋮
30	0	30	0	$B(30;30, \pi^E) \times B(30;30, \pi^C)$	0

The computation time of σ_T^2 was large, and therefore we could not include the same conditions as in simulation study 1. One value for the true effect size was selected ($\mu = 0$), π_i^C was 0.1 or 0.5, and *k* was set equal to 5, 40, or 160. The sample size of all studies was set equal to 30, because the methods' coverage probabilities were expected to deviate the most from the nominal coverage rate in this condition with the smallest study sample sizes. The amount of between-study heterogeneity (τ) was, as in simulation study 1, varied from 0 to 0.5 in steps of 0.1. This simulation study was also programmed in R (R Core Team, 2017) and the packages "parallel" (R Core Team, 2017) and "metafor" (Viechtbauer, 2010) were used. A total number of 3,000 replications per condition were used. R code of simulation study 2 is available via: https://osf.io/xba4y/.

9.5.2 Results Monte-Carlo simulation study 2

Figure 9.2 shows the coverage probabilities of the *Q*-profile and GENQ methods when using estimator $\hat{\sigma}^2$ and when using the true variances σ_T^2. Similar to Figure 9.1, triangles refer to the *Q*-profile method, plus signs to the GENQ method with variance weights, and crosses to the GENQ method with standard error weights. The estimator of the sampling variance that was also used in simulation study 1 ($\hat{\sigma}^2$) is indicated with solid black lines, while σ_T^2 is indicated with dashed gray lines. Note that the results of both simulation studies cannot directly be compared because a sample size of 30 for each primary study was used in simulation study 2 instead of different sample sizes as in simulation study 1.

In general, coverage probabilities were closer to the nominal coverage rate if $\pi_i^C = 0.1$ and σ_T^2 was used. This can be seen in the top left panel of Figure 9.2 (*k*=5; $\pi_i^C = 0.1$) where coverage probabilities were closer to the nominal coverage rate (although slightly too low) when σ_T^2 was used instead of $\hat{\sigma}^2$. If $\pi_i^C = 0.5$ (second row of panels in Figure 9.2), no severe undercoverage was observed for the three methods when using $\hat{\sigma}^2$ or σ_T^2 since all coverage probabilities were larger than 0.9. Simulation study 1 showed that coverage probabilities most notably diverged from the nominal coverage rate when *k*=160 and $\pi_i^C = 0.1$. This is also apparent here; coverage probabilities based on $\hat{\sigma}^2$ are below 0.8 for each value of τ and therefore not visible in the figure. Coverage probabilities of the *Q*-profile and GENQ method with variance weights were not above 0.268, and coverage probabilities of the GENQ method with standard error weights were not above 0.662. However, although coverage probabilities substantially improved when using σ_T^2 (e.g., for $\tau = 0.5$ *Q*-profile: 0.212 vs. 0.918, GENQ with variance weights: 0.108 vs. 0.917, and GENQ with standard error weights: 0.470 vs. 0.917), coverage probabilities still deviated from the nominal coverage rate. To conclude, using true sampling variances rather than estimated sampling variances considerably improved the coverage probability of the *Q*-profile and GENQ methods, but did not always provide nominal CIs. It follows that these deficiencies must be caused by two other assumptions of the random-effects model that were violated in our simulation study; normal sampling distributions of the effect sizes and uncorrelated random effects and sampling errors.

To increase our understanding of how violating the assumption of known sampling variances as well as violations of other assumptions underlying the random-effects model affect the generalized *Q*-statistic, we computed the generalized *Q*-statistic as described in Equation 3 based on $\hat{\sigma}^2$ and σ_T^2 and examined how well its probability density function (pdf) was approximated by a χ^2 distribution with *k*-1

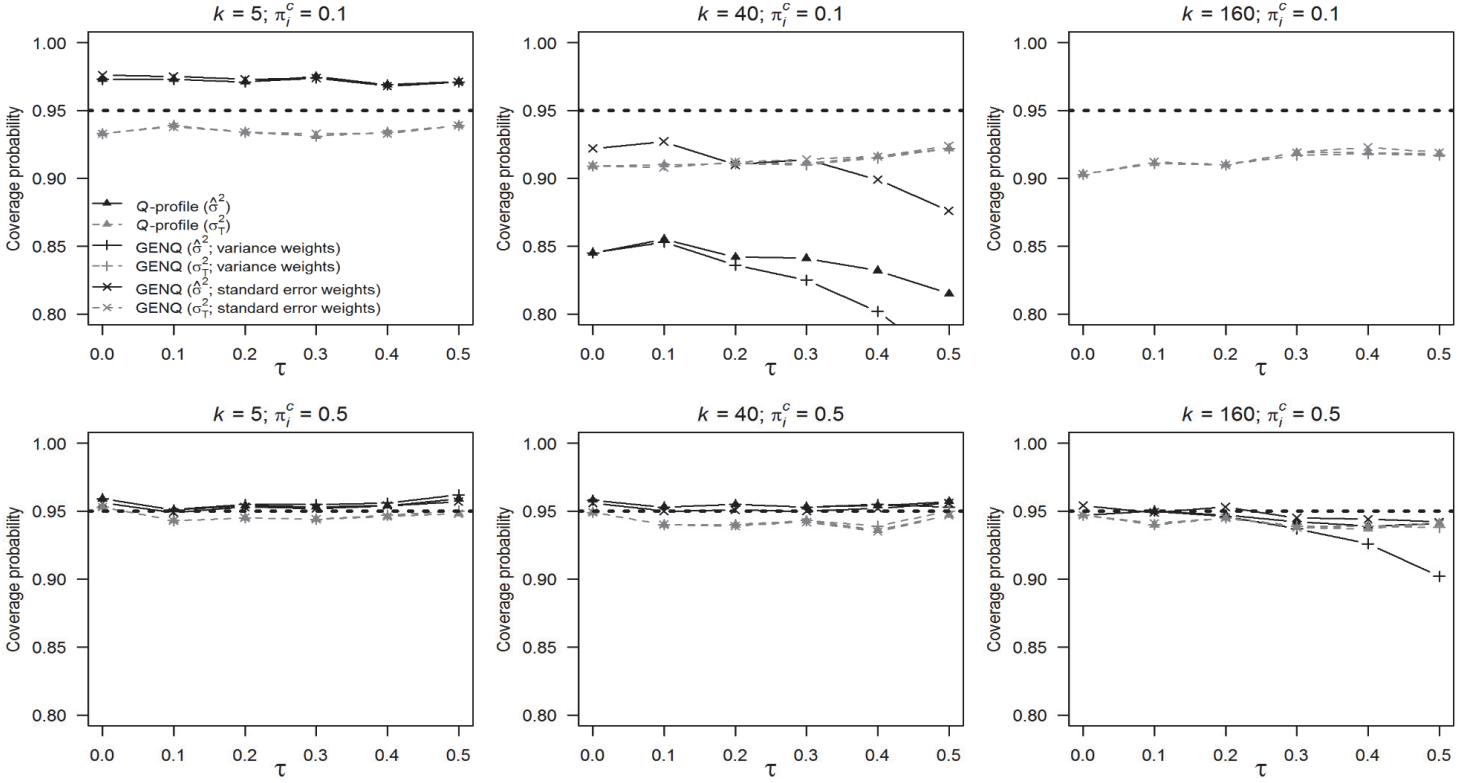

Figure 9.2. Coverage probabilities of the *Q*-profile method, GENQ method with variance weights, and GENQ method with standard error weights. The probability of the outcome of interest in the control group is denoted by π_i^C, the number of primary studies in a meta-analysis with *k*, and the amount of between-study heterogeneity with τ. The estimator of the sampling variance ($\hat{\sigma}^2$) is indicated with solid black lines and the true sampling variance (σ_T^2) with dashed gray lines. Note that the methods' coverage probabilities in the condition with estimated sampling variances are not shown in the upper right panel for *k* =160 and $\pi_i^C = 0.1$, since all these coverage probabilities were too low (< .8).

degrees of freedom. Since the generalized *Q*-statistic follows a χ^2 distribution if the assumptions of the random-effects model hold and these assumptions become less objectionable for larger sample sizes in the primary studies, we also computed the pdf with $\hat{\sigma}^2$ if there were 10 times more cases in the control and experimental group (300 instead of 30). Three different conditions were selected, representing coverage probabilities of the methods equal to the nominal coverage rate (*k*=5, $\pi_i^C = 0.5$, $\tau = 0$), overcoverage (*k*=5, p_i^C =0.1, τ=0), and undercoverage (*k*=160, $\pi_i^C = 0.1$, $\tau = 0.5$). Pdfs were created based on 5,000 generated generalized *Q*-statistics and R code for creating these pdfs is available via https://osf.io/bdhn8/. We focus in this section on the approximation of the generalized *Q*-statistic by the χ^2 distribution in the *Q*-profile method, because Q_a as used by the GENQ methods depends on the weights λ_i (see Equation 4), and therefore does not follow a single reference distribution. However, because coverage probabilities of the GENQ method with variance weights and the *Q*-profile method were comparable, we expect similar deviations of the weighted χ^2 distribution for the GENQ method as the deviations we find for *Q*-profile method.

Figure 9.3 shows the pdfs of the generalized *Q*-statistic when the coverage probability was close to the nominal coverage rate (left panel; *k*=5, $\pi_i^C = 0.5$, $\tau = 0$), when coverage was too large (middle panel; *k*=5, $\pi_i^C = 0.1$, $\tau = 0$), and when coverage was too low (right panel; *k*=160, $\pi_i^C = 0.1$, $\tau = 0.5$). The pdf of the generalized *Q*-statistic when the sampling variance is computed with $\hat{\sigma}^2$ is illustrated with a solid black line and σ_T^2 with a dashed gray line. The bold gray line corresponds to a χ^2 distribution with *k*-1 degrees of freedom, which in theory should be the distribution that is approximated by the other pdfs. Starting with the left panel (close to accurate coverage; *k*=5, $\pi_i^C = 0.5$, $\tau = 0$), the mean of the generalized *Q*-statistics was indeed close to the mean (4) of the χ^2 distribution (3.86 for $\hat{\sigma}^2$, 3.96 for σ_T^2). However, the variance (4x2=8) was somewhat different for $\hat{\sigma}^2$ (6.73), but not for σ_T^2 (7.69). As expected, the pdf was closely approximated by the χ^2 distribution if the primary studies' sample size was equal to 300 and $\hat{\sigma}^2$ was used to estimate the sampling variance (mean 4.01 and variance 8.20). These results suggest that the sampling variance was accurately estimated with $\hat{\sigma}^2$ for *k*=5, $\pi_i^C = 0.5$, and $\tau = 0$, and that the sample size of 30 was sufficiently large for this condition to approximate the pdf of the generalized *Q*-statistic with a χ^2 distribution.

The pdfs of the generalized *Q*-statistic for the condition with overcoverage

(k=5, $\pi_i^C = 0.1$, $\tau = 0$) are presented in the middle panel of Figure 9.3. The pdf of the generalized Q-statistic based on σ_T^2 was closer to the χ^2 distribution than based on $\hat{\sigma}^2$. Especially the variance of the generalized Q-statistic based on $\hat{\sigma}^2$ was too low (mean 3.07 < 4, and variance 3.27 < 8) whereas the mean and variance of the generalized Q-statistics were 3.97 and 9.87 for σ_T^2. The approximation was again best for sample sizes equal to 300 (dotted black line; mean of generalized Q-statistics 3.90 and variance 7.41). Here, the coverage probability of the Q-profile method (0.952) also approached the nominal coverage rate. These results indicate that a sample size of 30 was not sufficiently large to accurately approximate the χ^2 distribution with $\hat{\sigma}^2$ when k=5, $\pi_i^C = 0.1$, and $\tau = 0$. However, this approximation improved if σ_T^2 was used for computing the sampling variance or the sample size was equal to 300. Overcoverage of the methods for k=5, $\pi_i^C = 0.1$, $\tau = 0$ and sample sizes equal to 30 can be explained by the distribution of the generalized Q-statistic. Since the distribution of the generalized Q-statistic is to the left of the χ^2 distribution and its variance is smaller than that of the χ^2 distribution, the CIs will too often include $\tau = 0$.

For the condition with too low coverage probability (right panel; k=160, $\pi_i^C = 0.1$, $\tau = 0.5$), the pdf of the generalized Q-statistic based on estimator $\hat{\sigma}^2$ with a sample size of 30 per group (solid black line) deviated from the pdf of the χ^2 statistic. The mean (117.10) and variance (124.12) of the generalized Q-statistics were both substantially lower than those of a χ^2 distribution with 159 degrees of freedom (mean 159 and variance 318). Using $\hat{\sigma}_T^2$ resulted in a pdf markedly closer to the pdf of the χ^2 statistic. However, the generalized Q-statistics computed with σ_T^2 (dashed gray line; mean 156.40 and variance 384.30) still deviated from those of the χ^2 distribution. Again, increasing the primary studies' sample size to 300 yielded a pdf of the generalized Q-statistic that was better approximated by the χ^2 distribution (dotted black line; mean 153.90, variance 288.81). For this condition, the coverage probability of the Q-profile method (0.945) was also close to the nominal coverage rate. These results suggest that for k=160, $\pi_i^C = 0.1$, and $\tau = 0.5$ a sample size of 30 was too small to accurately approximate the χ^2 distribution even if the true sampling variances were used (σ_T^2).

Using the pdf of the generalized Q-statistic, we can now explain the undercoverage of the Q-profile method. Because the distribution of the Q-statistic is to the left of the χ^2 distribution, the lower and upper bounds of the CI around τ have to be obtained by decreasing τ^2 in Equation 3 till the 2.5th and 97.5th percentiles of this

χ^2 distribution are reached. Consequently, the CIs of the *Q*-profile method have too low lower and upper bounds, with τ often being larger than the upper bound. This was also apparent in the results of simulation study 1 in the condition k=160, $\pi_i^C = 0.1$, and $\tau = 0.5$, because the lower bound was never lower than τ and the undercoverage was fully explained by the upper bound being often smaller than τ.

9.6 Conclusion and discussion

Between-study variance is often present in a meta-analysis (Higgins, 2008; Higgins et al., 2009; Kontopantelis et al., 2013; van Erp et al., 2017). The amount of between-study variance can be estimated, but estimates are usually rather imprecise (Chung et al., 2013; Sidik & Jonkman, 2007). An estimate of the amount of between-study variance can be surrounded by a CI to illustrate its imprecision. Two recommended methods (Veroniki et al., 2016) to compute such a CI are the *Q*-profile (Viechtbauer, 2007b) and GENQ method (Biggerstaff & Jackson, 2008; Jackson, 2013). Both methods yield exact CIs under the assumptions of the random-effects model (i.e., unbiased observed effect size estimates, normal sampling distributions of the effect sizes, known sampling variances, and uncorrelated sampling errors and random effects). However, these assumptions are most likely violated in practice (Biggerstaff & Tweedie, 1997; Hoaglin, 2016a, 2016b) such that CIs of the *Q*-profile and GENQ method are approximations rather than exact CIs. The goal of this chapter was to study the performance of both methods under situations that are representative for research in practice where the assumptions underlying the random-effects model are violated.

Results of two Monte-Carlo simulation studies revealed that coverage probabilities of both methods can be substantially below the nominal coverage rate if model assumptions are violated. Coverage probabilities of both methods were especially too low if both the sample sizes of the primary studies and the probability of the outcome of interest were low in combination with a large number of studies in a meta-analysis. This result is in line with Viechtbauer (Viechtbauer, 2007b) who also showed that the coverage probability of the *Q*-profile method was too low if the number of studies was large in a meta-analysis in combination with large between-study heterogeneity. Coverage probabilities of the *Q*-profile method and the GENQ method with variance weights were comparable in our simulation studies. If coverage of the *Q*-profile and GENQ method with variance weights was close to the nominal rate, coverage probability of the GENQ method with standard error weights deviated more from the nominal rate than the other two methods. However, the GENQ method with standard error weights yielded better coverage probabilities than the *Q*-profile and GENQ method with variance weights if the coverage probability of the *Q*-profile and GENQ method with variance weights was substantially too low. This was caused

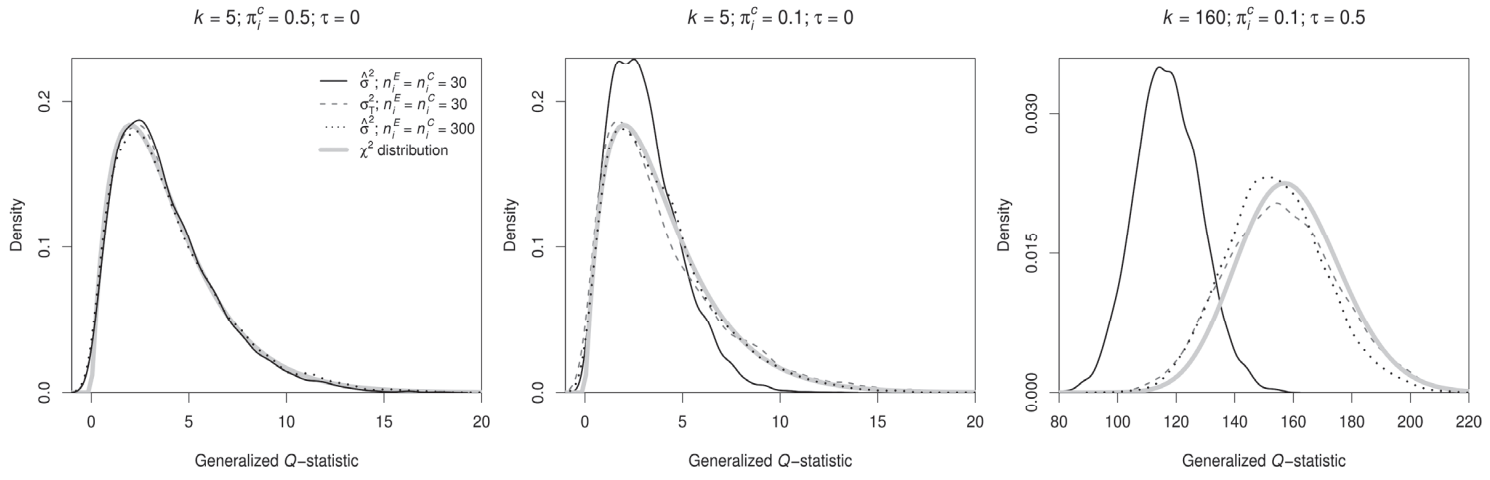

Figure 9.3. Probability density functions (pdfs) of the generalized *Q*-statistic (Equation 3) for *k*=5, $\pi_i^C = 0.5$, and $\tau = 0$ (coverage probability equal to nominal coverage rate), *k*=5, $\pi_i^C = 0.1$, and $\tau = 0$ (overcoverage), and *k*=160, $\pi_i^C = 0.1$, and $\tau = 0.5$ (undercoverage). Pdfs based on a sample size of 30 in the experimental and control group were obtained with three different estimators for the sampling variance: $\hat{\sigma}^2$ (solid black line) and σ_T^2 (dashed gray line). The pdf based on a sample size of 300 in both groups with estimator $\hat{\sigma}^2$ is presented with the dotted black line. The pdf of the χ^2 statistic is denoted by the bold gray line.

by the difference in weights, because more extreme observed log odds ratios (with larger sampling variances/standard errors) have a larger influence on the exact distribution of the *Q*-statistic if standard error weights are used instead of variance weights. However, coverage probabilities of the methods substantially deviated from the nominal coverage rate if the probability of the outcome of interest was low.

Our second simulation study showed that the mean and variance of the sampling distribution of the generalized *Q*-statistic may be too small in comparison to a χ^2 distribution with *k*-1 degrees of freedom if the probability of the outcome of interest was low and sample sizes were small. Consequently, the coverage probability of the *Q*-profile method is too small in these conditions even if the true sampling variances instead of estimated sampling variances are used. This deviation from the nominal coverage rate is caused by low frequencies in some of the cells of the observed frequency tables. Specific methods have been developed that perform better in such cases with sparse data by analyzing dichotomous data by means of generalized linear mixed-effects models. The sampling distributions in these methods are no longer assumed to be normal; instead, the exact likelihood based on binomial, Poisson, or hypergeometric distributions is used (Stijnen, Hamza, & Ozdemir, 2010). This approach is especially beneficial in case of a low probability of the outcome of interest, because no corrections (e.g., adding 0.5 to each cell) are required to deal with zero cells. However, future research is still needed to determine under which conditions the generalized linear mixed-effects models have better statistical properties for constructing CIs for τ^2 than the *Q*-profile and GENQ method.

A CI around the estimate of the between-study variance can also be used for computing a CI around the I^2 statistic (i.e., proportion of the total variance in a meta-analysis caused by the between-study variance) (Higgins & Thompson, 2002). Hence, the results presented in this chapter also apply to CIs around the I^2 statistic if constructed with the *Q*-profile or GENQ method. An advantage of quantifying between-study heterogeneity with the I^2 statistic is that it enables comparisons across meta-analyses (Higgins & Thompson, 2002; Higgins et al., 2003). CIs around the estimate of between-study variance and the I^2 statistic can also be used for testing the null hypothesis of homogeneous effect sizes in a meta-analysis. Software for applying the *Q*-profile and GENQ method for estimating a CI around the estimate of the between-study variance and the I^2 statistic are readily available in the R package "metafor" (Viechtbauer, 2010).

The commonly used *Q*-test (Cochran, 1954) for testing the null hypothesis of homogeneous effect sizes in a meta-analysis is also based on the assumptions of the random-effects model. The *Q*-statistic follows a χ^2 distribution if the assumptions underlying the random-effects model hold. Hence, inferences drawn by using the *Q*-test will also be affected by violations of these assumptions as is the case for the *Q*-profile and GENQ method, but the assumptions become more acceptable if the primary studies' sample size increase. Similar to our results with respect to the

generalized Q-statistic, Kulinskaya and Dollinger (2015) showed that the mean and variance of the distribution of the Q-statistic are too low when log odds ratios are used as effect size measure and the sample size is not sufficiently large. They propose to approximate the distribution of the Q-statistic by means of a gamma distribution and developed a new test for homogeneity based on this approximation. Future research may study whether the statistical properties of the Q-profile method improve if the distribution of the generalized Q-statistic is approximated by a gamma distribution instead of a χ^2 distribution.

Future research may also examine to what extent incorporating an estimate of the between-study variance in the weights of the GENQ method affects its CIs if the assumptions underlying the random-effects model do not hold. Using variance weights where an estimate of the between-study variance is also included corresponds to the standard weights that are commonly used in the random-effects model. However, the GENQ method is no longer exact if such an estimate is incorporated. Jackson (2013) already studied the statistical properties of the GENQ method when incorporating an estimate of the between-study variance in the weights, but only when the assumptions of the random-effects model hold; he concluded that the coverage probability only slightly deviated from the nominal coverage rate under these conditions.

One limitations of this chapter is that we only focus on one particular effect size measure (log odds ratio) in our simulation studies. Future research may therefore examine whether the statistical properties of the Q-profile and GENQ method depend on the effect size measure, because differences in estimates of the between-study variance can also be attributed to the effect size measure (Deeks, 2002; Engels, Schmid, Terrin, Olkin, & Lau, 2000; Friedrich, Adhikari, & Beyene, 2011). Deviations from the nominal coverage rate of the methods is expected to be less severe for effect size measures whose sampling distribution more closely follows a normal distribution.

To conclude, between-study heterogeneity is common in meta-analyses (Engels et al., 2000; Rhodes et al., 2015), and assessing heterogeneity is a crucial issue (Huedo-Medina et al., 2006). We recommend in line with others (Higgins et al., 2009; Ioannidis et al., 2007; Kepes et al., 2013; Langan et al., 2016) to include a CI around the estimate of between-study variance computed with the Q-profile or GENQ method in every meta-analysis. This illustrates imprecision in the estimate of the between-study variance and facilitates interpretation of the meta-analysis. Previous research has shown that the Q-profile and GENQ method have the best statistical properties, but the methods' coverage probabilities deviate from the nominal coverage rate if the probability of the outcome of interest is small. Hence, methods specifically developed for those situations should be considered to be used instead of the Q-profile or GENQ method.

CHAPTER 10

Epilogue

This dissertation focused on two issues that are crucial when conducting a meta-analysis: publication bias and heterogeneity in primary studies' true effect sizes. Publication bias is seen as one of the biggest threats to the validity of a meta-analysis (Dickersin & Min, 1993; Easterbrook et al., 1991; Rothstein et al., 2005a) generally resulting in overestimated effect sizes (e.g., Ioannidis, 2008b; Lane & Dunlap, 1978) and false positive results in the literature (e.g., Bakker et al., 2012; Fanelli, 2010b, 2012). There has been widespread attention for publication bias in a multitude of research fields (e.g., Driessen et al., 2015; Franco et al., 2014; Franco et al., 2016; Sterling et al., 1995) and multiple methods have been developed to test for publication bias and correct effect size estimation in a meta-analysis for publication bias (for an overview see Rothstein et al., 2005b). The unrealistic high prevalence of statistically significant results in the published literature (Fanelli, 2010b, 2012) that is probably caused by publication bias (but probably also by for instance *p*-hacking [Simmons et al., 2011]) also motivated researchers to replicate published studies. This resulted in several large groups of researchers of different disciplines verifying whether landmark studies of a research field could be confirmed in a replication (Begley & Ellis, 2012; Camerer et al., 2016; Errington et al., 2014; Open Science Collaboration, 2015).

The other prominent issue we focused on is heterogeneity in primary studies' true effect sizes. This between-study variance in true effect size is often present in meta-analyses (Higgins, 2008; Kontopantelis et al., 2013). Examining between-study variance may lead to relevant insights about the (in)consistency of the true effect size underlying primary studies' observed effect sizes (Higgins et al., 2009). An essential aspect in meta-analysis is accurately estimating this between-study variance in true effect sizes, which is also signified by the many different estimators for the between-study variance that have been developed (for an overview see Veroniki et al. [2016] and Langan et al. [2016]). Unfortunately, estimates of the between-study variance are usually imprecise (Chung et al., 2013; Sidik & Jonkman, 2007), and these estimates are therefore recommended to be reported together with a confidence interval (Ioannidis et al., 2007; Kepes et al., 2013; Langan et al., 2016). Similarly, many different methods to compute confidence intervals for the between-study variance in true effect size exist as well (for an overview see Veroniki et al., 2016).

We first discuss the results of the chapters related to publication bias (Chapters 2-7) and then those related to heterogeneity in true effect size. The key findings and recommendations of this dissertation are also listed in Table 10.1. These recommendations are built upon guidelines for conducting meta-analyses as the PRISMA (Moher et al., 2009) and MARS (American Psychological Association, 2010, Appendix) and are based on the results of the chapters in this dissertation for conducting meta-analyses, assessing publication bias, and meta-analyzing an original study and replication. Subsequently, I propose how the ideal system of science should look like such that the effects of publication bias and other biases are minimized. Next,

I suggest directions for future research and my dissertation ends with a general conclusion about the topics discussed in this dissertation.

Table 10.1. Key findings and recommendations of this dissertation.

1. A meta-analysis should always be accompanied with publication bias methods to get insight in the risk and severity of publication bias (Chapters 2-5, 10).

2. The trim-and-fill method should not be used to correct for publication bias, because the method does not adequately correct for the overestimation caused by publication bias and is outperformed by other methods (Chapter 2).

3. *P*-uniform accurately estimates effect size if between-study variance in true effect sizes is absent (Chapter 2 and Chapter 3).

4. *P*-hacking in primary studies causes an unpredictable bias in their effect size estimates, so researchers are advised to be reluctant in interpreting estimates of meta-analyses if strong indications of *p*-hacking are present in many primary studies (Chapter 3).

5. Hardly any evidence for publication bias was observed in a large-scale dataset of 83 meta-analysis published in Psychological Bulletin (71.1% of effect sizes was nonsignificant) and 499 systematic reviews from the Cochrane Database of Systematic Reviews (81.1% of effect sizes was nonsignificant) (Chapter 4).

6. Because *p*-uniform* and the selection model approach proposed by Hedges (1992) as implemented in the "weightr" package generally have good statistical properties, even if there are only 10 primary studies included in the meta-analysis, we generally recommend applying both methods to estimate effect size and between-study variance in true effect sizes (Chapter 5, but see key finding 7).

7. *P*-uniform* and the selection model approach proposed by Hedges (1992) are both not recommended to be used if there are only statistically significant results in a meta-analysis. In that situation, *p*-uniform is recommended to be used but only in the absence of heterogeneity in true effect sizes (Chapters 2-3, 5).

8. Meta-analysts should use multiple publication bias methods (i.e., triangulation) and report all their results, because no publication bias method performs best in all conditions (Chapters 4-5, 10).

9. Selecting a statistically significant effect size for replication is tantamount to the well-known "regression towards the mean problem"; the expected value of the statistically significant effect size exceeds the true effect size (Chapters 6-7).

10. When statistically combining a published original study and replication, the statistical significance of the original study should be taken into account when

estimating the common underlying true effect size to correct for the likely overestimation by the original study (see key finding 9; Chapters 6-7).

11. Computing the required sample size for a replication with the snapshot hybrid method while ignoring the information of the original study indicates that sample sizes of about 650 participants are needed to have a probability of 0.8 for observing a posterior model probability larger than 0.75 for a zero or small true effect size (Chapter 7).

12. Sample sizes in experimental economics and especially psychology are often too small to draw definite conclusions about the magnitude of effect size underlying one original study and replication (Chapters 6-7).

13. The multi-step estimator for estimating the between-study variance in a meta-analysis is an alternative justification for the use of the nowadays recommended Paule-Mandel estimator (Chapter 8).

14. Coverage probabilities of confidence intervals for the between-study variance in a meta-analysis can be substantially below the nominal coverage rate if the log odds ratio is used as effect size measure and the probability of the outcome of interest is low (Chapter 9).

15. The *Q*-profile and generalized *Q*-statistic method are recommended to be used for computing a confidence interval around the between-study variance in a meta-analysis when the effect size measure is the log odds ratio, but generalized linear mixed-effects models are advised to be used if the probability of the outcome of interest is low (Chapter 9).

16. Science in which all studies get published (including their materials, code, and data), where studies are pre-registered according to strict rules, and articles are reviewed with the consistency between pre-registration and article in mind, has the potential to eliminate most biases in science (Chapter 10).

17. The R package "puniform" and easy-to-use web applications were developed to apply the *p*-uniform, *p*-uniform*, hybrid method of meta-analysis, snapshot hybrid method, and to compute the required sample size for a replication with the snapshot hybrid method (Chapters 3, 5-7).

10.1 Publication bias

In Chapter 2, we introduced a new meta-analysis method, *p*-uniform, that deals with publication bias and is able to (i) test for publication bias, (ii) estimate effect size and compute a confidence interval, and (iii) test the null hypothesis of no effect. *P*-uniform only uses the statistically significant primary studies' effect sizes in a

meta-analysis and evaluates whether the *p*-values of these effect sizes conditional on being statistically significant are uniformly distributed at a particular effect size. For the effect size estimate of *p*-uniform, this means that the estimate is equal to the effect size for which a test statistic for assessing uniformity based on these conditional *p*-values equals its expected value. *P*-uniform was implemented using Fisher's (1925) method to assess uniformity of the conditional *p*-values, and a variant of Fisher's method that is less susceptible to outlying *p*-values.

We conducted Monte-Carlo simulations to compare the statistical properties of *p*-uniform with respect to effect size estimation with the trim-and-fill method (Duval & Tweedie, 2000a, 2000b), and the fixed-effect and random-effects model that do not correct for publication bias. Additionally, *p*-uniform's publication bias test was compared in the simulations with the test of excess significance (Ioannidis & Trikalinos, 2007b). The results of these simulations showed that estimates of the fixed-effect and random-effects model were accurate and their confidence intervals close to the nominal coverage rate in the absence of publication bias. However, statistical properties of the methods rapidly deteriorated as publication bias increased (key finding 1). Estimates of trim-and-fill were also severely biased if publication bias was present. This confirmed previous research that already concluded that trim-and-fill should not be used to correct for publication bias, because it was systematically outperformed by other methods (key finding 2; Moreno, Sutton, Ades, et al., 2009; Simonsohn et al., 2014a; Terrin et al., 2003). *P*-uniform's effect size estimates were accurate, and coverage probabilities of its confidence interval were close to the nominal coverage rate if the primary studies' true effect sizes were homogeneous (key finding 3). However, the method overestimated effect size if there was between-study variance in the true effect sizes, and this overestimation increased as a function of this between-study variance in true effect sizes. Statistical power of *p*-uniform's publication bias test was low, but generally larger than the power of the test of excess significance.

In Chapter 3, effect size estimation with *p*-uniform was compared to *p*-curve (Simonsohn et al., 2014a, 2014b), which is a highly similar method based on the same methodology as *p*-uniform. In Chapter 2, we used the Fisher's method for estimation, but we decided to use the Irwin-Hall estimator in Chapter 3. *P*-uniform's effect size estimate is in that case equal to the effect size for which the sum of the conditional *p*-values equals the expected value of the Irwin-Hall distribution. Advantages of the Irwin-Hall estimator over Fisher's method are that summing the conditional *p*-values is intuitive and that *p*-uniform's estimate using the Irwin-Hall distribution is smaller, equal, larger than zero if the average of the included statistically significant *p*-values in a meta-analysis is larger, equal, smaller than 0.025, respectively.

We examined the limitations of *p*-uniform and *p*-curve by assessing bias in the methods' estimates caused by *p*-hacking and overestimation of effect size if the between-study variance was moderate to large. The effects of three different types of

p-hacking (i.e., data peeking and collecting data for multiple conditions or dependent variables while not reporting all the results) were studied, and the results demonstrated that *p*-hacking caused a rather unpredictable bias in *p*-uniform and *p*-curve depending on the type of *p*-hacking (key finding 4). Based on our results with respect to estimating the effect size in the presence of between-study variance, we argued that *p*-uniform and *p*-curve did not accurately estimate the effect size if the between-study variance in primary studies' true effect sizes was larger than moderate. This is in contrast with what the developers of *p*-curve have repeatedly stated (Simmons, Nelson, & Simonsohn, 2018, January 8; Simonsohn, 2015, February 9; Simonsohn et al., 2014a). This disagreement boils down to a difference in the interpretation of the parameter that *p*-uniform and *p*-curve intend to estimate. Simonsohn and colleagues argue that *p*-uniform and *p*-curve intend to estimate the average true effect size underlying those studies yielding statistically significant effect sizes. That is, the methods answer the question what effect size will be obtained if we run exactly the same studies as those that yielded statistically significant results. We agree with Simonsohn et al. that *p*-uniform and *p*-curve provide an accurate answer to that question (i.e., an accurate estimate of the average true effect size underlying all statistically significant primary studies). However, we believe that estimating the true effect size of *all* primary studies (and not only those that gave statistically significant results) is of primary interest for applied researchers. This is actually the parameter that is estimated in the random-effects model (Borenstein et al., 2010). This parameter that we intend to estimate becomes upwardly biased as a function of the between-study variance.

In Chapter 4, we conducted a pre-registered study on the presence of publication bias and the possible overestimation caused by it in a large-scale data set consisting of 83 meta-analyses and 499 systematic reviews published in the fields of psychology and medicine, respectively. Data of all meta-analyses published in Psychological Bulletin between 2004 and 2014 that met the inclusion criteria were extracted to represent meta-analyses from the field of psychology. Systematic reviews published in the Cochrane Database of Systematic Reviews between 2004 and 2014 that met the inclusion criteria were sampled and data were extracted to represent meta-analyses from the field of medicine. The data of 83 meta-analyses representing research in psychology and 499 systematic reviews representing research in medicine were combined to create a large-scale data set. Before applying publication bias methods to this data set, we created homogeneous subsets per meta-analysis, because at the time of this study it was commonly believed that publication bias methods do not have good statistical properties in the presence of heterogeneity in primary studies' true effect sizes (e.g., Ioannidis & Trikalinos, 2007a; Ioannidis & Trikalinos, 2007b; McShane et al., 2016; Terrin et al., 2003). As a consequence, the majority of homogeneous subsets contained fewer than 10 primary studies' effect sizes, which are tough conditions for publication bias methods (e.g., Begg & Mazumdar, 1994; Sterne et

al., 2011).

Surprisingly, the results of Chapter 4 did not reveal evidence for the presence of publication bias in the large-scale data set; publication bias tests were not more often statistically significant than expected by chance and *p*-uniform's estimates hardly suggested overestimation caused by publication bias. Moreover, 71.1% of the primary studies' effect sizes representing the field of psychology and 81.1% representing the field of medicine were not statistically significant (key finding 5). These results are in sharp contrast with, for instance, research showing that 95% of the published research in psychology and psychiatry yields statistically significant results (Fanelli, 2010b, 2012). We observed that only 28.9% and 18.9% of the primary studies' effect sizes included in our large-scale data set were statistically significant in the meta-analyses published in the fields of psychology and medicine, respectively. A likely reason why publication bias was not observed in the large-scale data set is that meta-analyses studied relationships that were not the primary outcome for the majority of the primary studies. Publication bias most likely only influences publication of the outcome of a primary study that is of main interest, and secondary outcomes may remain relatively unaffected by publication bias.

We presented *p*-uniform*, which is an extension and major improvement of *p*-uniform, in Chapter 5. *P*-uniform* (i) does not overestimate the effect size in the presence of between-study variance in true effect size, (ii) is a more efficient estimator than *p*-uniform, and (iii) enables testing and drawing inferences for the average true effect size as well as the between-study variance. We compared the statistical properties of *p*-uniform* with the selection model approach proposed by Hedges (1992) implemented in the "weightr" package (Coburn & Vevea, 2016) in an analytical study with a meta-analysis only consisting of one statistically significant primary study's effect size and one nonsignificant effect size. Our analytical study revealed that *p*-uniform* and the selection model approach of Hedges (1992) did not suffer from severe convergence problems in case of a small number of primary studies in a meta-analysis, in contrast to what had been argued before (Field & Gillett, 2010; Terrin et al., 2003). Hence, both *p*-uniform* and the selection model approach proposed by Hedges (1992) can already be applied to meta-analyses with a very small number of studies, provided that these meta-analyses contain both statistically significant and nonsignificant effect sizes.

To examine the statistical properties of both methods we also conducted a Monte Carlo simulation study with conditions that were representative for meta-analyses in practice. Performance of the random-effects model was good if publication bias was absent, but its estimates were severely biased in case of extreme publication bias with only statistically significant primary studies' effect sizes in a meta-analysis. If the number of primary studies increased and the true effect size was zero, the test of no effect of the random-effects model always incorrectly concluded that the true effect size was larger than zero. This simulation study showed that the performance of *p*-

uniform* was generally comparable to the selection model approach proposed by Hedges (1992), yet superior to the performance of the random-effects model under publication bias (key finding 6). *P*-uniform* is preferred over the selection model approach if the main goal of a researcher is to estimate the average effect size and between-study variance, because the selection model approach had a systematic bias in estimating the average effect size and suffered more from convergence problems than *p*-uniform*. However, estimates of *p*-uniform* were, in general, more variable than estimates of the selection model approach proposed by Hedges (1992), and estimates of *p*-uniform* may be unrealistically highly negative if primary studies' *p*-values are close to the α-level. The simulations also revealed that statistical properties of *p*-uniform* and the selection model approach were both insufficient in cases with extreme publication bias resulting in meta-analyses with only included statistically significant primary studies' effect sizes (key finding 7). These results also imply that no publication bias method performs best in all possible conditions. Hence, we recommend in line with others (Coburn & Vevea, 2015; Kepes, Banks, McDaniel, et al., 2012) to use triangulation (key finding 8). That is, use multiple publication bias methods in a meta-analysis and report all their results.

In Chapter 6, we proposed a method (hybrid method of meta-analysis) to meta-analyze a statistically significant published original study and replication. The idea is that the original study is most likely statistically significant, because studies with statistically significant results are usually the ones that get published due to publication bias and/or are subsequently targeted for replication. Selecting a statistically significant effect size for replication is tantamount to the well-known "regression towards the mean problem"; the expected value of the statistically significant effect size exceeds the true effect size (key finding 9). In other words, this selection of statistically significant effect sizes introduces positive bias in the effect size of the original study, which is not necessarily caused by publication bias. Due to the positive bias, it is important that this statistical significance in the original study is taken into account, and this is exactly what the hybrid method of meta-analysis does. The hybrid method enables (i) estimating effect size, (ii) computing a confidence interval for the effect size, and (iii) testing the null hypothesis of no effect while combining a statistically significant original study and replication. The replication is assumed to be unbiased, so statistical significance is only taken into account in the original study. We implemented three variants of the hybrid method; the basic variant (hybrid), and two variants (hybrid^0 and hybrid^R) that avoid unrealistically highly negative estimates of the basic variant arising from original effect sizes with a *p*-value close to the α-level. Hybrid^0 sets the effect size equal to 0 if the hybrid method of meta-analysis yields a negative estimate while hybrid^R discards the original effect size if it suggests a negative effect size using *p*-uniform using the estimator based on the Irwin-Hall distribution. We implemented all methods using the Irwin-Hall distribution as estimator, and compared their performance to each other and to fixed-effect meta-

analysis in an analytical study. The results illustrated that it is inappropriate not to take statistical significance in the original study into account using the fixed-effect model, because it yields overestimated effect sizes and too large Type I errors (key finding 10).

We applied the three variants of the hybrid method of meta-analysis as well as the fixed-effect model to original studies and replications included in the Reproducibility Project: Psychology (Open Science Collaboration, 2015). The percentage of pairs original studies and replications where the null hypothesis of no effect was rejected was lower for the three variants of the hybrid method of meta-analysis (hybrid 28.4%, hybrid^0 28.4%, hybrid^R 34.3%) than fixed-effect meta-analysis (70.1%) and equal or lower than only considering the replication (34.3%). Estimates of the three variants of the hybrid method of meta-analysis were smaller than fixed-effect meta-analysis but slightly larger than only considering the replication. Hence, the smaller percentage of pairs of original studies and replications for hybrid and hybrid^0 where the null hypothesis was rejected was caused by more uncertainty in the estimates of these methods. Nevertheless, estimates of the three variants of the hybrid method of meta-analysis were usually too imprecise (i.e., large standard error) to draw definite conclusion about the existence of an effect, which is in line with a Bayesian reanalysis of the data of the Reproducibility Project: Psychology (Etz & Vandekerckhove, 2016).

Another method for evaluating an original study and replication was developed in Chapter 7. This snapshot hybrid method again takes statistical significance in the original study into account, and uses Bayesian analysis to compute posterior model probabilities at different snapshots of effect sizes (e.g., a zero, small, medium, and large effect size), thereby quantifying the magnitude of the effect size underlying an original study and replication. Furthermore, we illustrated how the snapshot hybrid method can be used for computing the required sample size for a replication using the observed effect size in the original study but also taking into account its statistical significance. This procedure is analogous to power analysis in null hypothesis testing, and resembles a frequentist method proposed by S. F. Anderson, Kelley, and Maxwell (2017) to conduct a power analysis while taking publication bias and uncertainty in effect size estimates into account. Interestingly, computing the required sample size for a replication with the snapshot hybrid method for a nonzero expected true effect size may result in a larger required sample size than when the original study was ignored. This occurs if the original study has a *p*-value close to the α-level, which is in line with a negative true effect size (see Chapters 2 and 3). Computing the required sample size for a replication with the snapshot hybrid method while ignoring the information of the original study indicates that sample sizes of about 650 participants are needed to have a probability of 0.8 for observing a posterior model probability larger than 0.75 for a zero or small true effect size (key finding 11).

Applying the snapshot hybrid method to the Reproducibility Project: Psychology (Open Science Collaboration, 2015) confirmed the findings of Chapter 6 and Etz and Vandekerckhove (2016) that it often was difficult to draw strong conclusions with respect to the underlying true effect size based on the original study and replication because of their relatively small sample sizes (key finding 12). That is, strong conclusions (i.e., posterior model probability of at least 0.75 for a true effect size) could not be drawn with respect to 62.7% of the pairs of original studies with the snapshot hybrid method. The snapshot hybrid method was also applied to the Experimental Economics Replication Project (Camerer et al., 2016) and, in contrast with the Reproducibility Project: Psychology, strong conclusions could not be drawn for only 18.8% of the pairs of original study and replication Comparing the results of the Reproducibility Project: Psychology and the Experimental Economics Replication Project also revealed that studied effect sizes were generally larger in the original studies and replications in the Experimental Economics Replication Project. However, the effect sizes in the original study were also likely overestimated since the observed effect size in the original study was larger than in the replication in 81.3% of the replicated effects.

10.2 Estimating heterogeneity in meta-analysis

In Chapter 8, we proposed a new multi-step estimator for estimating the between-study variance in a meta-analysis. This multi-step estimator is a natural extension of the two-step estimators as suggested by DerSimonian and Kacker (2007). We proved that, if the multi-step estimator converges, it converges to the Paule-Mandel estimator (Paule & Mandel, 1982) if the number of steps of the multi-step estimator is large. We also show that this proof holds for the random-effects meta-regression model where covariates are included in the model to explain the between-study variance. The established relationship between the multi-step estimator and Paule-Mandel estimator is important since it provides an alternative justification for the Paule-Mandel estimator (key finding 13) that is nowadays one of the recommend methods for estimating the between-study variance in primary studies' true effect sizes (Langan et al., 2016; Veroniki et al., 2016). We illustrated the multi-step estimator using three example data sets of meta-analyses and showed that for each example the multi-step and Paule-Mandel estimate were equal to each other.

Estimates of the between-study variance are usually imprecise (Chung et al., 2013; Kontopantelis et al., 2013; Sidik & Jonkman, 2007) making it desirable to report a confidence interval for the between-study variance. In Chapter 9, we studied the statistical properties of the Q-profile and generalized Q-statistic method that are recommended methods for computing such a confidence interval (Langan et al., 2016; Veroniki et al., 2016). These methods are exact if the assumptions underlying the random-effects model hold, but these assumptions are usually violated in practice (Biggerstaff & Tweedie, 1997; Hardy & Thompson, 1998; Hoaglin, 2016a, 2016b). For

example, the primary studies' sampling variances are usually not known whereas this is an assumption of the random-effects model. Statistical properties of the *Q*-profile and generalized *Q*-statistic method were examined in two Monte-Carlo simulation studies with log odds ratios as effect size measure. In the first simulation study, we assessed the methods' coverage probabilities and the width of the confidence intervals for conditions that were representative for meta-analyses in practice. Coverage probabilities of both methods were substantially below the nominal coverage rate if the assumptions of the random-effects model were violated especially if the probability of the outcome of interest was low (key finding 14). In the second simulation study, the true sampling variances of the primary studies were computed in order to examine whether this would affect the performance of the *Q*-profile and generalized *Q*-statistic method. Using the true sampling variances instead of estimated sampling variances in simulations improved the coverage probabilities of the methods, but coverage still deviated from the nominal coverage rate if the probability of the outcome of interest was low. This was caused by the distribution of generalized *Q*-statistics, which was not well approximated by a χ^2 distribution if the primary studies' sample sizes were small. Hence, we advise meta-analysts to be careful with interpreting these confidence intervals if the effect size measure is the log odds ratio and the probability of the outcome of interest is low. We recommend to use general linear mixed-effects models (for an overview see Jackson, Law, Stijnen, Viechtbauer, & White, 2018), because these methods have better statistical probabilities if log odds ratio is the effect size measure and the probability of the outcome of interest is low (key finding 15).

10.3 The ideal system of science

The problem of publication bias would be absent in an ideal system of science where all studies are published (or become at least publicly accessible) regardless of their statistical significance. A step in this direction can be set by fully disclosing all the results, data, and research materials of a study either in a paper or in its supplements, or if a study is not deemed suitable for publication, in an online repository (e.g., Open Science Framework, http://osf.io). This makes the scientific system as a whole more effective (van Assen et al., 2014), without bias, and enables more precisely estimating the between-study variance in true effect size in a meta-analysis. However, this only holds if studies are properly conducted, analyzed, and reported (i.e., no *p*-hacking). Fully disclosing the data and research materials also enables other researchers to reproduce results, thereby increasing the trust in research findings.

Another step in the right direction focusing on preventing *p*-hacking is by requiring researchers to pre-register all studies such that all steps that are planned in the design, conduct, analysis, and reporting of studies are specified before the researcher collects the data (e.g., Asendorpf et al., 2013; de Groot, 1956/2014; Nosek

et al., 2015; Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012; Wicherts et al., 2016). To prevent *p*-hacking, pre-registrations should be specific, precise, and exhaustive (Wicherts et al., 2016). For instance, researchers can include computer code for all planned statistical analyses in their pre-registration. Including computer code for analyses is specific (these analyses will be conducted), precise (with these options), and exhaustive (no other planned analyses are carried out). Unfortunately, current systems of pre-registrations are not so strict, thereby leaving ample room for researcher degrees of freedom (Veldkamp et al., 2017). Being specific, precise, and exhaustive in a pre-registration is strenuous, as it is very difficult to foresee all the issues that may arise during data collection. This is also what we experienced when pre-registering the analysis plan for Chapter 4. However, I believe that it is possible for researchers, especially if they are using common experimental designs (i.e., factorial designs), to learn how to effectively pre-register their studies (being specific, precise, and exhaustive). This is a necessary and essential step to create an unbiased scientific system. Finally, I note that this does not imply that all analyses reported in a paper should be planned. Deviations from a preregistration are possible as long as this is fully disclosed in a paper. It is then up to the editor, reviewers, and the scientific community to evaluate whether such a deviation is acceptable or not. Additionally, researchers should, of course, also be able to conduct "exploratory" studies (studies without planned analyses), but researchers should clearly distinguish between planned/confirmatory and exploratory analyses in their research.

Previous research comparing articles with their pre-registered plans revealed many differences and inconsistencies (Chan, 2008; Chan & Altman, 2005; Chan, Hróbjartsson, Haahr, Gøtzsche, & Altman, 2004; Chan, Hróbjartsson, Jørgensen, Gøtzsche, & Altman, 2008; Goldacre, 2016). This implies that pre-registration of planned analyses is not sufficient, and that reviewing should enforce the consistency of a paper with its pre-registration. This could be done by using registered reports (Chambers, 2013) where pre-registrations are peer-reviewed and researchers can adjust their pre-registration according to the comments made by the reviewers and editors. If reviewers and editors accept the pre-registration, the researchers can actually conduct the research, and it will be published irrespective of the results, provided that the researchers follow the pre-registration. Hence, these registered reports make it impossible for researchers to deviate from their preregistration, thereby tackling both the issues of *p*-hacking and publication bias.

Science in which all studies get published (including their materials, code, and data), where studies are pre-registered according to strict rules, and articles are reviewed with the consistency between pre-registration and article in mind, has the potential to eliminate most biases in science (key finding 16). Ideally, this setup is accompanied by a larger focus in research on finding determinants or potential moderators of effects. Currently the fundamental problem of science is that a majority

of the published research is not replicable, as many original effects in psychology, economics, biology and medicine do not hold or hold to a lesser extent in other contexts. This paradoxical state of affairs can be addressed while planning the research, for instance by large-scale collaborations examining a research question using the same or similar methods in many labs simultaneously. Some of these Many-Lab projects (Ebersole et al., 2016; Klein et al., 2014; Klein et al., 2017) are already conducted and meta-analyses methods are required to estimate the average effect size and between-study variance and test whether there are differences across labs. I hope to see many of these projects in the near future, because these projects provide relevant insights into the generalizability of a phenomenon.

Unfortunately, the described ideal system of science is not (yet) implemented. Hence, my hope is that the increased use of the promising methods that were developed and examined in this dissertation to correct for publication bias and statistically combine an original study and replication will help science moving forward. We have developed easy-to-use software for researchers that want to apply *p*-uniform, *p*-uniform*, the hybrid meta-analysis method, and snapshot hybrid method (key finding 17). R functions are available in the R package "puniform" that is available on GitHub (https://github.com/RobbievanAert/puniform), and we have developed web applications for researchers that are not familiar with R: *p*-uniform (https://rvanaert.shinyapps.io/p-uniform/), *p*-uniform* (https://rvanaert.shinyapps.io/p-uniformstar/), the hybrid meta-analysis method (https://rvanaert.shinyapps.io/hybrid/), and snapshot method (https://rvanaert.shinyapps.io/snapshot/).

10.4 Future research

A relevant option for future research is to develop estimators and confidence intervals for the between-study variance that deals with the fact that the *Q*-statistic is not well approximated by a χ^2 distribution if the primary studies' sample sizes are small as in most current applied research (Chapter 9). Kulinskaya and colleagues (Kulinskaya & Dollinger, 2015; Kulinskaya, Dollinger, & Bjørkestøl, 2011a, 2011b) obtained similar results as we did and showed that the distribution of the *Q*-statistic can be better approximated with the general gamma distribution for effect size measures log odds ratio, risk difference, and standardized mean differences (Cohen's *d*). Estimators of the between-study variance may be less biased and corresponding confidence intervals could be closer to the nominal coverage rate if a gamma distribution is used to approximate the distribution of the *Q*-statistic rather than a χ^2 distribution.

More research is also needed on the statistical properties of publication bias methods. Many publication bias methods have been developed, but there is not enough insight into what the most appropriate method is in different conditions. For instance, research on the statistical properties of PET-PEESE showed that the method

has good statistical properties for meta-analyses that are typical for the field of economics (Stanley & Doucouliagos, 2014), but the statistical properties of the method deteriorate for meta-analysis that are typical for the field of psychology (Carter et al., 2017; Stanley, 2017) due to the low variability in primary studies' sample sizes. More research is also needed on whether bias is introduced in *p*-uniform* and the selection model approach proposed by Hedges (1992) if the probability of publishing a statistically nonsignificant primary study's effect size depends on its effect size or *p*-value. *P*-uniform* and also the selection model approach assume that this probability is the same for each statistically nonsignificant primary study's effect size. However, this assumption may be violated in practice (see Simonsohn et al., 2017, December 20 for a discussion about this assumption). The selection model approach proposed by Hedges (1992) enables specifying more than two intervals making the method more robust to a violation of this assumption.

Moreover, the effect of *p*-hacking on publication bias methods was not widely studied whereas results in Chapter 2 provided ample evidence that at least *p*-uniform, *p*-curve, and also the fixed-effect model are biased by different types of *p*-hacking. Future research may also consider studying publication bias in existing meta-analyses with methods that have been developed in recent years. Publication bias methods were often not applied in these meta-analyses (Aguinis et al., 2010; Aytug et al., 2012; Banks, Kepes, & Banks, 2012) or suboptimal methods were used (Ioannidis, 2008a; Ioannidis & Trikalinos, 2007a). Hence, reanalyzing data analogous as we did in Chapter 4 with, for instance, *p*-uniform*, the selection model approach of Hedges (1992), or other methods (e.g., PET-PEESE [Stanley & Doucouliagos, 2014] and Bayesian methods proposed by Guan and Vandekerckhove [2015] and Du et al. [2017]) will likely yield relevant insights.

As a reviewer of multiple manuscripts that conducted a meta-analysis, I have often encountered the desire of researchers to include dependent primary studies' effect sizes in a meta-analysis. However, including dependent effect sizes in the same meta-analysis violates the assumption of independence (Borenstein et al., 2009) and using the fixed-effect or random-effects model is no longer suitable. Multivariate meta-analysis models (e.g., Jackson et al., 2011; Mavridis & Salanti, 2013) can be used to take the dependence of the primary studies' effect sizes into account, but no publication bias methods exist for these models. Future research may focus on to the consequences of publication bias in a multivariate meta-analysis as well as considering developing publication bias methods with good statistical properties for these models.

Some psychology journals (e.g., Psychological Science, Lindsay, 2017) encourage authors to share data, which may offer interesting directions for future research. Meta-analyses based on primary studies' effect sizes can be replaced by so-called individual participant data (or individual patient data, IPD) meta-analysis (Borenstein et al., 2009; Higgins & Green, 2011) if the data of primary studies

becomes available. The raw data of the primary studies is analyzed rather than the aggregated data in an individual participant data meta-analysis. This approach has become more popular in the field of medical research (Riley, Lambert, & Abo-Zaid, 2010) and has, for instance, the advantage over meta-analyzing aggregated primary studies' data that researchers can also examine the effect of moderating variables at the participant level (for an overview of advantages see Stewart & Tierney, 2002). However, publishing primary studies' data or sharing data (Wicherts, Borsboom, Kats, & Molenaar, 2006) is not (yet) the norm within psychology research, so it may take time before large-scale individual participant data meta-analyses can be conducted in psychology.

Another opportunity for future research is extending the hybrid and snapshot methods such that they can handle multiple published statistically significant original studies and replications. The original studies and replications can then be combined in the extended versions of the hybrid and snapshot method, where the replications are considered as unbiased replications by the methods. This will yield a more efficient estimator of the hybrid method and better inferences for both the hybrid and snapshot method. This extension is straightforward by combining on the one hand the methodology developed for *p*-uniform* (Chapter 5) and on the other hand the hybrid method of meta-analysis or the snapshot hybrid (Chapter 6 and Chapter 7). Another possible extension is developing Bayesian alternatives of methods that were proposed in this dissertation. In collaboration with Hilde Augusteijn, Quentin Gronau, Eric-Jan Wagenmakers, and Marcel van Assen, I already started to work on developing a Bayesian alternative of *p*-uniform, but this is also possible for *p*-uniform* and the hybrid method of meta-analysis. Moreover, future research may also consider implementing maximum likelihood estimation for the hybrid method of meta-analysis instead of the estimator based on the Irwin-Hall distribution.

10.5 General conclusion

Meta-analysis has become more and more popular the last couple of decades. Its increasing popularity is partially due to the information explosion with respect to the number of papers published. Other reasons are the relevant insights and potentially important implications for practice that a meta-analysis may reveal. However, meta-analysis is not a magic tool that always leads to conclusions that are both precise and valid. The quality of the results of a meta-analysis are highly dependent on the quantity and quality of the included primary studies. I hope to have increased awareness for the shortcomings of meta-analysis, and that the statistical methods developed in this dissertation will be used in the future to both improve the quality of meta-analyses and advance research on meta-analysis.

References

Ackerman, J. M., Nocera, C. C., & Bargh, J. A. (2010). Incidental haptic sensations influence social judgments and decisions. *Science, 328*(5986), 1712-1715. doi:http://dx.doi.org/10.1126/science.1189993

Adair, J. G., & Vohra, N. (2003). The explosion of knowledge, references, and citations. Psychology's unique response to a crisis. *American Psychologist, 58*(1), 15-23. doi: 10.1037/0003-066X.58.1.15

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

Addis, D. R., Wong, A. T., & Schacter, D. L. (2008). Age-related changes in the episodic simulation of future events. *Psychological Science, 19*(1), 33-41. doi:10.1111/j.1467-9280.2008.02043.x

Agresti, A. (2013). *Categorical data analysis*. Hoboken, N.J.: Wiley-Interscience.

Aguinis, H., Dalton, D. R., Bosco, F. A., Pierce, C. A., & Dalton, C. M. (2010). Meta-analytic choices and judgment calls: Implications for theory building and testing, obtained effect sizes, and scholarly impact. *Journal of Management, 37*(1), 5-38. doi:10.1177/0149206310377113

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

Albarracín, D., Handley, I. M., Noguchi, K., McCulloch, K. C., Li, H., Leeper, J., . . . Hart, W. P. (2008). Increasing and decreasing motor and cognitive output: A model of general action and inaction goals. *Journal of Personality and Social Psychology, 95*(3), 510-523. doi:10.1037/a0012833

Alvarez, G. A., & Oliva, A. (2008). The representation of simple ensemble visual features outside the focus of attention. *Psychological Science, 19*(4), 392-398. doi:10.1111/j.1467-9280.2008.02098.x

American Psychological Association. (2010). *Publication manual* (6th ed.). Washington, DC: American Psychological Association.

Amir, Y., & Sharon, I. (1990). Replication research: A "must" for the scientific advancement of psychology. *Journal of Social Behavior and Personality, 5*(4), 51-69.

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology, 94*(1), 60-74. doi:10.1037/0022-3514.94.1.60

Anderson, C. J., Bahník, Š., Barnett-Cowan, M., Bosco, F. A., Chandler, J., Chartier, C. R., . . . Zuni, K. (2016). Response to comment on “Estimating the reproducibility of psychological science”. *Science, 351*(6277), 1037-1037. doi:10.1126/science.aad9163

Anderson, S. F., Kelley, K., & Maxwell, S. E. (2017). Sample-size planning for more accurate statistical power: A method adjusting sample effect sizes for publication bias and uncertainty. *Psychological Science, 28*(11), 1547-1562. doi:10.1177/0956797617723724

Anderson, S. F., & Maxwell, S. E. (2016). There's more than one way to conduct a replication study: Beyond statistical significance. *Psychological Methods, 21*(1), 1-12. doi:10.1037/met0000051

Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association, 49*(268), 765-769.

Armor, D. A., Massey, C., & Sackett, A. M. (2008). Prescribed optimism: Is it right to be wrong about the future? *Psychological Science, 19*(4), 329-331. doi:10.1111/j.1467-9280.2008.02089.x

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., . . . Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality, 27*(2), 108-119. doi:http://dx.doi.org/10.1002/per.1919

Augusteijn, H. E. M., van Aert, R. C. M., & van Assen, M. A. L. M. (2017). *The effect of publication bias on the Q-test and assessment of heterogeneity*. Manuscript submitted for publication.

Aytug, Z. G., Rothstein, H. R., Zhou, W., & Kern, M. C. (2012). Revealed or concealed? Transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods, 15*(1), 103-133. doi:10.1177/1094428111403495

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543-554. doi:10.1177/1745691612459060

Bakker, M., & Wicherts, J. M. (2014a). Outlier removal and the relation with reporting errors and quality of psychological research. *PLoS ONE, 9*(7), : e103360. doi:10.1371/journal.pone.0103360

Bakker, M., & Wicherts, J. M. (2014b). Outlier removal, sum scores, and the inflation of the Type I error rate in independent samples t tests: the power of alternatives and recommendations. *Psychological Methods, 19*(3), 409-427. doi:10.1037/met0000014

Bangert-Drowns, R. L., Hurley, M. M., & Wilkinson, B. (2004). The effects of school-based writing-to-learn interventions on academic achievement: A meta-analysis. *Review of Educational Research, 74*(1), 29-58. doi:10.3102/00346543074001029

Banks, G. C., Kepes, S., & Banks, K. P. (2012). Publication bias: The antagonist of meta-analytic reviews and effective policymaking. *Educational Evaluation and Policy Analysis, 34*(3), 259-277. doi:10.3102/0162373712446144

Banks, G. C., Kepes, S., & McDaniel, M. A. (2012). Publication bias: A call for improved meta-analytic practice in the organizational sciences. *International Journal of Selection and Assessment, 20*(2), 182-197. doi:10.1111/j.1468-2389.2012.00591

Bassok, M., Pedigo, S. F., & Oskarsson, A. T. (2008). Priming addition facts with semantic relations. *Journal of Experimental Psychology, 34*(2), 343-352. doi:10.1037/0278-7393.34.2.343

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

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics, 50*(4), 1088-1101.

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature, 483*(7391), 531-533. doi:doi:10.1038/483531a

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science: Improving the standard for basic and preclinical research. *Circulation research, 116*(1), 116-126. doi:10.1161/CIRCRESAHA.114.303819

Berkey, C. S., Hoaglin, D. C., Mosteller, F., & Colditz, G. A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine, 14*(4), 395-411. doi:10.1002/sim.4780140406

Bhaumik, D. K., Amatya, A., Normand, S. L. T., Greenhouse, J., Kaizar, E., Neelon, B., & Gibbons, R. D. (2012). Meta-analysis of rare binary adverse event data. *Journal of the American Statistical Association, 107*(498), 555-567. doi:10.1080/01621459.2012.664484

Biggerstaff, B. J., & Jackson, D. (2008). The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. *Statistics in Medicine, 27*(29), 6093-6110. doi:10.1002/sim.3428

Biggerstaff, B. J., & Tweedie, R. L. (1997). Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine, 16*(7), 753-768.

Bland, M. (2013). Do baseline p-values follow a uniform distribution in randomised trials? *PLoS ONE, 8*(10). doi:10.1371/journal.pone.0076010

Blankenship, K. L., & Wegener, D. T. (2008). Opening the mind to close it: Considering a message in light of important values increases message processing and later resistance to change. *Journal of Personality and Social Psychology, 94*(2), 196-213. doi:10.1037/0022-3514.94.2.94.2.196

Borenstein, M. (2009). Effect sizes for continuous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (pp. 221-236). New York: Russell Sage Foundation.

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods, 1*(2), 97-111. doi:10.1002/jrsm.12

Bouter, L. M., Tijdink, J., Axelsen, N., Martinson, B. C., & ter Riet, G. (2016). Ranking major and minor research misbehaviors: results from a survey among participants of four World Conferences on Research Integrity. *Research Integrity and Peer Review, 1*(1). doi:https://doi.org/10.1186/s41073-016-0024-5

Bressan, P., & Stranieri, D. (2008). The best men are (not always) already taken: Female preference for single versus attached males depends on conception risk. *Psychological Science, 19*(2), 145-151. doi:10.1111/j.1467-9280.2008.02060.x

Bruns, S. B., & Ioannidis, J. P. (2016). P-curve and p-hacking in observational research. *PLoS ONE, 11*(2), :e0149144. doi:10.1371/journal.pone.0149144

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376. doi:10.1038/nrn3475

Byyny, R. L. (2012). The data deluge: The information explosion in medicine and science. *The Pharos, 75*(2), 2-5.

Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., . . . Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science, 351*(6280), 1433-1436. doi:10.1126/science.aaf0918

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

Casella, G., & Berger, R. L. (2002). *Statistical Inference*. Belmont, CA: Duxbury.

Centerbar, D. B., Schnall, S., Clore, G. L., & Garvin, E. D. (2008). Affective incoherence: When affective concepts and embodied reactions clash. *Journal of Personality and Social Psychology, 94*(4), 560-578. doi:10.1037/0022-3514.94.4.560

Chambers, C. D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*(3), 609-610. doi:10.1016/j.cortex.2012.12.016

Chan, A.-W. (2008). Bias, spin, and misreporting: Time for full access to trial protocols and results. *PLoS Med, 5*(11), e230. doi:10.1371/journal.pmed.0050230

Chan, A.-W., & Altman, D. G. (2005). Identifying outcome reporting bias in randomised trials on PubMed: Review of publications and survey of authors. *BMJ, 330*(7494), 753. doi:10.1136/bmj.38356.424606.8F

Chan, A.-W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *JAMA, 291*(20), 2457-2465.

Chan, A.-W., Hróbjartsson, A., Jørgensen, K. J., Gøtzsche, P. C., & Altman, D. G. (2008). Discrepancies in sample size calculations and data analyses reported in randomised trials: Comparison of publications with protocols. *BMJ, 337*. doi:10.1136/bmj.a2299

Chandler, J. J., Reinhard, D., & Schwarz, N. (2012). To judge a book by its weight you need to know its content: Knowledge moderates the use of embodied cues. *Journal of Experimental Social Psychology, 48*(4), 948-953. doi:10.1016/j.jesp.2012.03.003

Chung, Y., Rabe-Hesketh, S., & Choi, I. H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine, 32*(23), 4071-4089. doi:10.1002/sim.5821

Citkowicz, M., & Vevea, J. L. (2017). A parsimonious weight function for modeling publication bias. *Psychological Methods, 22*(1), 28-41. doi:http://dx.doi.org/10.1037/met0000119

Cleary, R. J., & Casella, G. (1997). An application of Gibbs sampling to estimation in meta-analysis: Accounting for publication bias. *Journal of Educational and Behavioral Statistics, 22*(2), 141-154.

Coburn, K. M., & Vevea, J. L. (2015). Publication bias as a function of study characteristics. *Psychological Methods, 20*(3), 310-330. doi:http://dx.doi.org/10.1037/met0000046

Coburn, K. M., & Vevea, J. L. (2016). weightr: Estimating weight-function models for publication bias. Retrieved from https://CRAN.R-project.org/package=weightr

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304-1312.

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), 447-452. doi:10.1037/1082-989X.2.4.447

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009a). Foreword. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (pp. xi-xiii). New York: Russell Sage Foundation.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009b). *The Handbook of Research Synthesis and Meta-Analysis*. New York: Russell Sage Foundation.

Copas, J. B. (1999). What works?: Selectivity models and meta-analysis. *Journal of the Royal Statistical Society. Series A, 162*(1), 95-109.

Copas, J. B., & Li, H. G. (1997). Inference for non-random Samples. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 59*(1), 55-95. doi:https://doi.org/10.1111/1467-9868.00055

Copas, J. B., & Shi, J. Q. (2000). Meta-analysis, funnel plots and sensitivity analysis. *Biostatistics, 1*(3), 247-262. doi:https://doi.org/10.1093/biostatistics/1.3.247

Copas, J. B., & Shi, J. Q. (2001). A sensitivity analysis for publication bias in systematic reviews. *Statistical Methods in Medical Research, 10*(4), 251-265.

Correll, J. (2008). 1/f noise and effort on implicit measures of bias. *Journal of Personality and Social Psychology, 94*(1), 48-59. doi:10.1037/0022-3514.94.1.48

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: A note on meta-analysis bias. *Professional Psychology: Research and Practice, 17*(2), 136-137. doi:10.1037/0735-7028.17.2.136

Cox, C. R., Arndt, J., Pyszczynski, T., Greenberg, J., Abdollahi, A., & Solomon, S. (2008). Terror management and adults' attachment to their parents: The safe haven remains. *Journal of Personality and Social Psychology, 94*(4), 696-717. doi:10.1037/0022-3514.94.4.696

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science, 3*(4), 286-300. doi:10.1111/j.1745-6924.2008.00079.x

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

de Groot, A. D. (1956/2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas]. *Acta Psychologica, 148*, 188-194. doi:10.1016/j.actpsy.2014.02.001

Dear, K. B. G., & Begg, C. B. (1992). An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science, 7*(2), 237-245.

Deeks, J. J. (2002). Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine, 21*(11), 1575-1600. doi:10.1002/sim.1188

DerSimonian, R., & Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials, 28*(2), 105-114.

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

Dessalegn, B., & Landau, B. (2008). More than meets the eye: The role of language in binding and maintaining feature conjunctions. *Psychological Science, 19*(2), 189-195. doi:10.1111/j.1467-9280.2008.02066.x

Dickersin, K. (2005). Publication bias: Recognizing the problem understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 11-33). Chichester, England: Wiley.

Dickersin, K., Chan, S., Chalmers, T. C., Sacks, H. S., & Smith, H. (1987). Publication bias and clinical trials. *Controlled Clinical Trials, 8*(4), 343-353.

Dickersin, K., & Min, Y.-I. (1993). Publication bias: The problem that won't go away. *Annals of the New York Academy of Sciences, 703*(1), 135-148. doi:10.1111/j.1749-6632.1993.tb26343.x

Dodson, C. S., Darragh, J., & Williams, A. (2008). Stereotypes and retrieval-provoked illusory source recollections. *Journal of Experimental Psychology, 34*(3), 460-477. doi:10.1037/0278-7393.34.3.460

Driessen, E., Hollon, S. D., Bockting, C. L. H., Cuijpers, P., & Turner, E. H. (2015). Does publication bias inflate the apparent efficacy of psychological treatment for major depressive disorder? A systematic review and meta-analysis of US National Institutes of Health-Funded Trials. *PLoS ONE, 10*(9), e0137864. doi:10.1371/journal.pone.0137864

Du, H., Liu, F., & Wang, L. (2017). A Bayesian "fill-In" method for correcting for publication bias in meta-analysis. *Psychological Methods, 22*(4), 799-817. doi:10.1037/met0000164

Duval, S. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assesment, and adjustments*. Chichester, UK: Wiley.

Duval, S., & Tweedie, R. L. (2000a). A nonparametric "trim and fill" method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89-98. doi:10.1080/01621459.2000.10473905

Duval, S., & Tweedie, R. L. (2000b). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455-463.

Dwan, K., Altman, D. G., Arnaiz, J. A., Bloom, J., Chan, A.-W., Cronin, E., . . . Siegfried, N. (2008). Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE, 3*(8), e3081. doi:10.1371/journal.pone.0003081

Easterbrook, P. J., Berlin, J. A., Gopalan, R., & Matthews, D. R. (1991). Publication bias in clinical research. *Lancet, 337*(8746), 867-872.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., . . . Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68-82. doi:10.1016/j.jesp.2015.10.012

Eddelbuettel, D. (2013). *Seamless R and C++ integration with Rcpp*. New York, NY: Springer.

Efron, B., & Tibshirani, R. (1993). *An introduction to the bootstrap*. New York: Chapman & Hall.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*, 629-634.

Eitam, B., Hassin, R. R., & Schul, Y. (2008). Nonconscious goal pursuit in novel environments: The case of implicit learning. *Psychological Science, 19*(3), 261-267. doi:10.1111/j.1467-9280.2008.02078.x

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

Else-Quest, N. M., Hyde, J. S., Goldsmith, H. H., & Van Hulle, C. A. (2006). Gender differences in temperament: A meta-analysis. *Psychological Bulletin, 132*(1), 33-72. doi:10.1037/0033-2909.132.1.33

Engels, E. A., Schmid, C. H., Terrin, N., Olkin, I., & Lau, J. (2000). Heterogeneity and statistical significance in meta-analysis: an empirical study of 125 meta-analyses. *Statistics in Medicine, 19*(13), 1707-1728.

Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *eLife, 3*, e04333. doi:10.7554/eLife.04333

Ersner-Hershfield, H., Mikels, J. A., Sullivan, S. J., & Carstensen, L. L. (2008). Poignancy: Mixed emotional experience in the face of meaningful endings. *Journal of Personality and Social Psychology, 94*(1), 158-167. doi:10.1037/0022-3514.94.1.158

Estes, Z., Verges, M., & Barsalou, L. W. (2008). Head up, foot down: Object words orient attention to the objects' typical location. *Psychological Science, 19*(2), 93-97. doi:10.1111/j.1467-9280.2008.02051.x

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the Reproducibility Project: Psychology. *PLoS ONE, 11*(2). doi:10.1371/journal.pone.0149794

Exline, J. J., Baumeister, R. F., Zell, A. L., Kraft, A. J., & Witvliet, C. V. (2008). Not so innocent: Does seeing one's own capacity for wrongdoing predict forgiveness? *Journal of Personality and Social Psychology, 94*(3), 495-515. doi:10.1037/0022-3514.94.3.495

Eysenck, H. J. (1978). An exercise in mega-silliness. *American Psychologist, 33*(5), 517-517.

Fanelli, D. (2010a). Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PLoS ONE, 5*(4). doi:doi:10.1371/journal.pone.0010271

Fanelli, D. (2010b). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE, 5*(4), e10068. doi:10.1371/journal.pone.0010068

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*(3), 891-904. doi:10.1007/s11192-011-0494-7

Farebrother, R. W. (1984). Algorithm AS 204: The distribution of a positive linear combination of χ^2 random variables. *Journal of the Royal Statistical Society, 33*(3), 332-339.

Farrell, S. (2008). Multiple roles for time in short-term memory: Evidence from serial recall of order and timing. *Journal of Experimental Psychology, 34*(1), 128-145. doi:10.1037/0278-7393.34.1.128

Farris, C., Treat, T. A., Viken, R. J., & McFall, R. M. (2008). Perceptual mechanisms that characterize gender differences in decoding women's sexual intent. *Psychological Science, 19*(4), 348-354. doi:10.1111/j.1467-9280.2008.02092.x

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. doi:10.3758/BF03193146

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods, 17*(1), 120-128. doi:10.1037/a0024445

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology, 63*(3), 665-694. doi:10.1348/000711010X502733

Fisher, R. A. (1921). On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron., 1*, 3-32.

Fisher, R. A. (1925). *Statistical methods for research workers* (1st ed.). London: Oliver & Boyd.

Fleiss, J. L., & Berlin, J. A. (2009). Effect sizes for dichotomous data. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (pp. 237-253). New York: Russell Sage Foundation.

Forti, S., & Humphreys, G. W. (2008). Sensitivity to object viewpoint and action instructions during search for targets in the lower visual field. *Psychological Science, 19*(1), 42-48. doi:10.1111/j.1467-9280.2008.02044.x

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review, 19*(6), 975-991. doi:10.3758/s13423-012-0322-y

Francis, G. (2013). Replication, statistical consistency, and publication bias. *Journal of Mathematical Psychology, 57*(5), 153-169. doi:10.1016/j.jmp.2013.02.003

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505. doi:10.1126/science.1255484

Franco, A., Simonovits, G., & Malhotra, N. (2016). Underreporting in psychology experiments: Evidence from a study registry. *Social Psychological and Personality Science, 7*(1), 8-12. doi:10.1177/1948550615598377

Friedrich, J. O., Adhikari, N. K., & Beyene, J. (2011). Ratio of means for analyzing continuous outcomes in meta-analysis performed as well as mean difference methods. *Journal of Clinical Epidemiology, 64*(5), 556-564. doi:10.1016/j.jclinepi.2010.09.016

Ganor-Stern, D., & Tzelgov, J. (2008). Across-notation automatic numerical processing. *Journal of Experimental Psychology, 34*(2), 430-437. doi:10.1037/0278-7393.34.2.430

Gentle, J. E. (2004). *Random number generation and Monte Carlo methods* (2nd ed.). New York: Springer.

Gerber, A. S., Green, D. P., & Nickerson, D. (2001). Testing for publication bias in political science. *Political Analysis, 9*(4), 385-392.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *351*(6277). doi:10.1126/science.aad7243

Givens, G. H., Smith, D. D., & Tweedie, R. L. (1997). Publication bias in meta-analysis: A Bayesian data-augmentation approach to account for issues exemplified in the passive smoking debate. *Statistical Science, 12*(4), 221-250.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(10), 3-8.

Goff, P. A., Steele, C. M., & Davies, P. G. (2008). The space between us: stereotype threat and distance in interracial contexts. *Journal of Personality and Social Psychology, 94*(1), 91-107. doi:10.1037/0022-3514.94.1.91

Goldacre, B. (2016). Make journals report clinical trials properly. *Nature, 530*(7588). doi:10.1038/530007a

Goschke, T., & Dreisbach, G. (2008). Conflict-triggered goal shielding: Response conflicts attenuate background monitoring for prospective memory cues. *Psychological Science, 19*(1), 25-32. doi:10.1111/j.1467-9280.2008.02042.x

Greenhouse, J. B., & Iyengar, S. (2009). Sensitivity analysis and diagnostics. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 417-433). New York: Russell Sage Foundation.

Guan, M., & Vandekerckhove, J. (2015). A Bayesian approach to mitigation of publication bias. *Psychonomic Bulletin and Review*, Advance online publication. doi:10.3758/s13423-015-0868-6

Häfner, M. (2013). When body and mind are talking. Interoception moderates embodied cognition. *Experimental Psychology, 60*(4), 255-259. doi:10.1027/1618-3169/a000194

Hajcak, G., & Foti, D. (2008). Errors are aversive: defensive motivation and the error-related negativity. *Psychological Science, 19*(2), 103-108. doi:10.1111/j.1467-9280.2008.02053.x

Halevy, N., Bornstein, G., & Sagiv, L. (2008). In-group love and out-group hate as motives for individual participation in intergroup conflict: A new game paradigm. *Psychological Science, 19*(4), 405-411. doi:10.1111/j.1467-9280.2008.02100.x

Hardy, R. J., & Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine, 15*(6), 619-629.

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine, 17*(8), 841-856.

Hartgerink, C. H. J., van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & van Assen, M. A. L. M. (2015). Distributions of p-values smaller than .05 in psychology: What is going on? *PeerJ, 4:e1935*. doi:https://doi.org/10.7717/peerj.1935

Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal, 41*(8), 901-916.

Hartung, J., & Knapp, G. (2001a). On tests of the overall treatment effect in meta-analysis with normally distributed responses. *Statistics in Medicine, 20*(12), 1771-1782. doi:10.1002/sim.791

Hartung, J., & Knapp, G. (2001b). A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine, 20*(24), 3875-3889. doi:10.1002/sim.1009

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. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107-128.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin, 93*(2), 388-395. doi:10.1037//0033-2909.93.2.388

Hedges, L. V. (1984). Estimation of effect size under nonrandom sampling: The effects of censoring studies yielding statistically insignificant mean differences. *Journal of Educational Statistics, 9*(1), 61-85.

Hedges, L. V. (1992). Modeling publication selection effects in meta-analysis. *Statistical Science, 7*(2), 246-255.

Hedges, L. V., & Vevea, J. L. (1996). Estimating effect size under publication bias: Small sample properties and robustness of a random effects selection model. *Journal of Educational and Behavioral Statistics, 21*(4), 299-332.

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

Heine, S. J., Buchtel, E. E., & Norenzayan, A. (2008). What do cross-national comparisons of personality traits tell us? The case of conscientiousness. *Psychological Science, 19*(4), 309-313. doi:10.1111/j.1467-9280.2008.02085.x

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61-83. doi:10.1017/S0140525X0999152X

Higgins, J. P. T. (2008). Commentary: Heterogeneity in meta-analysis should be expected and appropriately quantified. *International Journal of Epidemiology, 37*(5), 1158-1160. doi:10.1093/ije/dyn204

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 www.cochrane-handbook.org.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine, 21*(11), 1539-1558. doi:10.1002/sim.1186

Higgins, J. P. T., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *British Medical Journal, 327*(7414), 557-560. doi:10.1136/bmj.327.7414.557

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society, 172*(1), 137-159.

Ho, M. S. K., & Lee, C. W. (2012). Cognitive behaviour therapy versus eye movement desensitization and reprocessing for post-traumatic disorder – is it all in the homework then? *Revue Européenne de Psychologie Appliquée/European Review of Applied Psychology, 62*(4), 253-260. doi:10.1016/j.erap.2012.08.001

Hoaglin, D. C. (2016a). Misunderstandings about Q and 'Cochran's Q test' in meta-analysis. *Statistics in Medicine, 35*(4), 485-495. doi:10.1002/sim.6632

Hoaglin, D. C. (2016b). Shortcomings of an approximate confidence interval for moment-based estimators of the between-study variance in random-effects meta-analysis. *Research Synthesis Methods*. doi:10.1002/jrsm.1205

Hopewell, S., Clarke, M., & Mallet, S. (2005). Grey literature and systematic reviews. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99-110).

Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society, Series B, 15*(2), 193-232.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychological Methods, 11*(2), 193-206. doi:10.1037/1082-989X.11.2.193

Hung, H. M., O'Neill, R. T., Bauer, P., & Köhne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics, 53*(1), 11-22.

Hunter, J. E., & Schmidt, F. L. (2015). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks, California: Sage.

IJzerman, H., Padiotis, N., & Koole, S. L. (2013). Replicability of social-cognitive priming: The case of weight as an embodiment of importance. Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2255726.

IntHout, J., Ioannidis, J. P., & Borm, G. F. (2014). The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. *BMC Medical Research Methodology, 14*. doi:10.1186/1471-2288-14-25

Ioannidis, J. P. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), e124. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. (2008a). Interpretation of tests of heterogeneity and bias in meta-analysis. *Journal of Evaluation in Clinical Practice, 14*(5), 951-957. doi:10.1111/j.1365-2753.2008.00986.x

Ioannidis, J. P. (2008b). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640-648. doi:10.1097/EDE.0b013e31818131e7

Ioannidis, J. P. (2011). Excess significance bias in the literature on brain volume abnormalities. *Archives of General Psychiatry, 68*(8), 773-780. doi:10.1001/archgenpsychiatry.2011.28

Ioannidis, J. P. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly 94*(3), 485-514. doi:10.1111/1468-0009.12210

Ioannidis, J. P., Patsopoulos, N. A., & Evangelou, E. (2007). Uncertainty in heterogeneity estimates in meta-analyses. *BMJ, 335*(7626), 914-916. doi:10.1136/bmj.39343.408449.80

Ioannidis, J. P., & Trikalinos, T. A. (2007a). The appropriateness of asymmetry tests for publication bias in meta-analyses: A large survey. *Canadian Medical Association Journal, 176*(8), 1091-1096. doi:10.1503/cmaj.060410

Ioannidis, J. P., & Trikalinos, T. A. (2007b). An exploratory test for an excess of significant findings. *Clinical Trials, 4*(3), 245-253. doi:10.1177/1740774507079441

Ioannidis, J. P., Trikalinos, T. A., & Zintzaras, E. (2006). Extreme between-study homogeneity in meta-analyses could offer useful insights. *Journal of Clinical Epidemiology, 59*(10), 1023-1032. doi:10.1016/j.jclinepi.2006.02.013

Iyengar, S., & Greenhouse, J. B. (1988a). Selection models and the file drawer problem. *Statistical Science, 3*(1), 109-117. doi:10.1214/ss/1177013012

Iyengar, S., & Greenhouse, J. B. (1988b). Selection models and the file drawer problem: Rejoinder. *Statistical Science, 3*(1), 133-135. Retrieved from http://www.jstor.org/stable/2245932

Jackson, D. (2006). The implications of publication bias for meta-analysis' other parameter. *Statistics in Medicine, 25*(17), 2911-2921. doi:10.1002/sim.2293

Jackson, D. (2007). Assessing the implications of publication bias for two popular estimates of between-study variance in meta-analysis. *Biometrics, 63*(1), 187-193. doi:10.1111/j.1541-0420.2006.00663.x

Jackson, D. (2013). Confidence intervals for the between-study variance in random effects meta-analysis using generalised Cochran heterogeneity statistics. *Research Synthesis Methods, 4*(3), 220-229. doi:10.1002/jrsm.1081

Jackson, D., Bowden, J., & Baker, R. (2010). How does the DerSimonian and Laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *Journal of Statistical Planning and Inference, 140*(4), 961-970. doi:http://dx.doi.org/10.1016/j.jspi.2009.09.017

Jackson, D., Bowden, J., & Baker, R. (2015). Approximate confidence intervals for moment-based estimators of the between-study variance in random effects meta-analysis. *Research Synthesis Methods, 6*(4), 372-382. doi:10.1002/jrsm.1162

Jackson, D., Law, M., Stijnen, T., Viechtbauer, W., & White, I. R. (2018). A comparison of 7 random-effects models for meta-analyses that estimate the summary odds ratio. *Statistics in Medicine*. doi:10.1002/sim.7588

Jackson, D., Riley, R., & White, I. R. (2011). Multivariate meta-analysis: Potential and promise. *Statistics in Medicine, 30*(20), 2481-2498. doi:10.1002/sim.4172

Jackson, D., Turner, R. M., Rhodes, K., & Viechtbauer, W. (2014). Methods for calculating confidence and credible intervals for the residual between-study variance in random effects meta-regression models. *BMC Medical Research Methodology, 14*(1). doi:10.1186/1471-2288-14-103

Jackson, D., Veroniki, A. A., Law, M., Tricco, A. C., & Baker, R. (2017). Paule-Mandel estimators for network meta-analysis with random inconsistency effects. *Research Synthesis Methods*. doi:10.1002/jrsm.1244

Jackson, D., White, I. R., & Riley, R. D. (2013). A matrix-based method of moments for fitting the multivariate random effects model for meta-analysis and meta-regression. *Biometrical Journal, 55*(2), 231-245. doi:10.1002/bimj.201200152

Jackson, D., White, I. R., & Thompson, S. G. (2010). Extending DerSimonian and Laird's methodology to perform multivariate random effects meta-analyses. *Statistics in Medicine, 29*(12), 1282-1297. doi:10.1002/sim.3602

Janiszewski, C., & Uy, D. (2008). Precision of the anchor influences the amount of adjustment. *Psychological Science, 19*(2), 121-127. doi:10.1111/j.1467-9280.2008.02057.x

Janssen, N., Alario, F. X., & Caramazza, A. (2008). A word-order constraint on phonological activation. *Psychological Science, 19*(3), 216-220. doi:10.1111/j.1467-9280.2008.02070.x

Jin, Z. C., Zhou, X. H., & He, J. (2014). Statistical methods for dealing with publication bias in meta-analysis. *Statistics in Medicine, 34*(2), 343-360. doi:10.1002/sim.6342

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532. doi:10.1177/0956797611430953

Jones, C. W., Handler, L., Crowell, K. E., Keil, L. G., Weaver, M. A., & Platts-Mills, T. F. (2013). Non-publication of large randomized clinical trials: Cross sectional analysis. *British Medical Journal, 347*. doi:10.1136/bmj.f6104

Jostmann, N. B., Lakens, D., & Schubert, T. W. (2009). Weight as an embodiment of importance. *Psychological Science, 20*(9), 1169-1174.

Jürgens, G., & Graudal, N. A. (2004). Effects of low sodium diet versus high sodium diet on blood pressure, renin, aldosterone, catecholamines, cholesterols, and triglyceride. *Cochrane Database of Systematic Reviews, (1)*. doi:10.1002/14651858.CD004022.pub2

Kaspar, K. (2013). A weighty matter: Heaviness influences the evaluation of disease severity, drug effectiveness, and side effects. *PLoS ONE, 8*(11), e78307. doi:10.1371/journal.pone.0078307

Kaspar, K., & Krull, J. (2013). Incidental haptic stimulation in the context of flirt behavior. *Journal of Nonverbal Behavior, 37*(3), 165-173. doi:10.1007/s10919-013-0154-0

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association, 90*(430), 791. doi:10.2307/2291091

Kavvoura, F. K., McQueen, M. B., Khoury, M. J., Tanzi, R. E., Bertram, L., & Ioannidis, J. P. (2008). Evaluation of the potential excess of statistically significant findings in published genetic association studies: application to Alzheimer's disease. *American Journal of Epidemiology, 168*(8), 855-865. doi:10.1093/aje/kwn206

Kepes, S., Banks, G. C., McDaniel, M., & Whetzel, D. L. (2012). Publication bias in the organizational sciences. *Organizational Research Methods, 15*(4), 624-662. doi:10.1177/1094428112452760

Kepes, S., Banks, G. C., & Oh, I. S. (2012). Avoiding bias in publication bias research: The value of null findings. *Journal of Business and Psychology*. doi:10.1007/s10869-012-9279-0

Kepes, S., McDaniel, M. A., Brannick, M. T., & Banks, G. C. (2013). Meta-analytic reviews in the organizational sciences: Two meta-analytic schools on the way to MARS (the Meta-analytic Reporting Standards). *Journal of Business and Psychology, 28*(2), 123-143.

Kicinski, M. (2013). Publication bias in recent meta-analyses. *PLoS ONE, 8*(11). doi:10.1371/journal.pone.0081823

Kirkham, J. J., Dwan, K. M., Altman, D. G., Gamble, C., Dodd, S., Smyth, R., & Williamson, P. R. (2010). The impact of outcome reporting bias in randomised controlled trials on a cohort of systematic reviews. *BMJ, 340*(c365). doi:10.1136/bmj.c365

Kisamore, J., & Brannick, M. (2008). An illustration of the consequences of meta-analysis model choice. *Organizational Research Methods, 11*(1), 35-53. doi:10.1177/1094428106287393

Klein, R. A., Ratliff, K. A., Brooks, B., Vianello, M., Galliani, E. M., Adams Jr, R. B., . . . Hasselman, F. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142-152. doi:10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B., Adams, R. B., Alper, S., . . . Nosek, B. A. (2017). Many Labs 2: Investigating variation in replicability across sample and setting. Manuscript submitted for publication.

Knapp, G., Biggerstaff, B. J., & Hartung, J. (2006). Assessing the amount of heterogeneity in random-effects meta-analysis. *Biometrical Journal, 48*(2), 271-285.

Koenker, R. (2005). *Quantile regression*. Cambridge: Cambridge University Press.

Kontopantelis, E., Springate, D. A., & Reeves, D. (2013). A re-analysis of the Cochrane Library data: The dangers of unobserved heterogeneity in meta-analyses. *PLoS ONE, 8*(7). doi:10.1371/journal.pone.0069930

Kouchaki, M., Gino, F., & Jami, A. (2014). The burden of guilt: Heavy backpacks, light snacks, and enhanced morality. *Journal of Experimental Psychology General, 143*(1), 414-424. doi:http://dx.doi.org/10.1037/a0031769

Kraemer, H. C., Gardner, C., Brooks, J., & Yesavage, J. A. (1998). Advantages of excluding underpowered studies in meta-analysis: Inclusionist versus exclusionist viewpoints. *Psychological Methods, 3*(1), 23-31. doi:10.1037/1082-989X.3.1.23

Kühberger, A., Fritz, A., & Scherndl, T. (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLoS ONE, 9*(9), e105825. doi:10.1371/journal.pone.0105825

Kulinskaya, E., & Dollinger, M. B. (2015). An accurate test for homogeneity of odds ratios based on Cochran's Q-statistic. *BMC Medical Research Methodology, 15*, 49. doi: 10.1186/s12874-015-0034-x

Kulinskaya, E., Dollinger, M. B., & Bjørkestøl, K. (2011a). On the moments of Cochran's Q statistic under the null hypothesis, with application to the meta-analysis of risk difference. *Research Synthesis Methods, 2*(4), 254-270. doi:10.1002/jrsm.54

Kulinskaya, E., Dollinger, M. B., & Bjørkestøl, K. (2011b). Testing for homogeneity in meta-analysis I. The one-parameter case: Standardized mean difference. *Biometrics, 67*(1), 203-212. doi:10.1111/j.1541-0420.2010.01442.x

Lakens, D. (2014). What p-hacking really looks like: A comment on Masicampo and LaLande (2012). *Quarterly Journal of Experimental Psychology*. doi:10.1080/17470218.2014.982664

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. P. T., & Simmonds, M. (2016). Comparative performance of heterogeneity variance estimators in meta-analysis: A review of simulation studies. *Research Synthesis Methods*(498). doi:10.1002/jrsm.1198

Larose, D. T., & Dey, D. K. (1998). Modeling publication bias using weighted distributions in a Bayesian framework. *Computational Statistics and Data Analysis, 26*(3), 279-302. doi:https://doi.org/10.1016/S0167-9473(97)00039-X

Larsen, J. T., & McKibban, A. R. (2008). Is happiness having what you want, wanting what you have, or both? *Psychological Science, 19*(4), 371-377. doi:10.1111/j.1467-9280.2008.02095.x

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

Lau, G. P., Kay, A. C., & Spencer, S. J. (2008). Loving those who justify inequality: The effects of system threat on attraction to women who embody benevolent sexist ideals. *Psychological Science, 19*(1), 20-21. doi:10.1111/j.1467-9280.2008.02040.x

Lemay, E. P., & Clark, M. S. (2008). "Walking on eggshells": How expressing relationship insecurities perpetuates them. *Journal of Personality and Social Psychology, 95*(2), 420-441. doi:10.1037/0022-3514.95.2.420

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

Lindsay, D. S. (2017). Sharing data and materials in Psychological Science. *Psychological Science, 28*(6), 699-702. doi:10.1177/0956797617704015

Lobue, V., & DeLoache, J. S. (2008). Detecting the snake in the grass: Attention to fear-relevant stimuli by adults and young children. *Psychological Science, 19*(3), 284-289. doi:10.1111/j.1467-9280.2008.02081.x

Loevinger, J. (1948). The technic of homogeneous tests compared with some aspects of scale analysis and factor analysis. *Psychological Bulletin, 45*(6), 507-529.

Macaskill, P., Walter, S. D., & Irwig, L. (2001). A comparison of methods to detect publication bias in meta-analysis. *Statistics in Medicine, 20*(4), 641-654.

Maglio, S. J., & Trope, Y. (2012). Disembodiment: Abstract construal attenuates the influence of contextual bodily state in judgment. *Journal of Experimental Psychology General, 141*(2), 211-216. doi:http://dx.doi.org/10.1037/a0024520

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(2), 161-175.

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537-542. doi:10.1177/1745691612460688

Makovski, T., Sussman, R., & Jiang, Y. V. (2008). Orienting attention in visual working memory reduces interference from memory probes. *Journal of Experimental Psychology, 34*(2), 369-380. doi:10.1037/0278-7393.34.2.369

Masicampo, E. J., & Baumeister, R. F. (2008). Toward a physiology of dual-process reasoning and judgment: Lemonade, willpower, and expensive rule-based analysis. *Psychological Science, 19*(3), 255-260. doi:10.1111/j.1467-9280.2008.02077.x

Massey, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American Statistical Association, 46*(253), 68-78.

Mavridis, D., & Salanti, G. (2013). A practical introduction to multivariate meta-analysis. *Statistical Methods in Medical Research, 22*(2), 133-158. doi:10.1177/0962280211432219

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist, 70*(6), 487-498. doi:10.1037/a0039400

McCall, R. B., & Carriger, M. S. (1993). A meta-analysis of infant habituation and recognition memory performance as predictors of later IQ. *Child Development, 64*(1).

McCrea, S. M. (2008). Self-handicapping, excuse making, and counterfactual thinking: Consequences for self-esteem and future motivation. *Journal of Personality and Social Psychology, 95*(2), 274-292. doi:10.1037/0022-3514.95.2.274

McDaniel, M. A., Rothstein, H. R., & Whetzel, D. L. (2006). Publication bias: A case study of four test vendors. *Personnel Psychology Personnel Psychology, 59*(4), 927-953.

McKinstry, C., Dale, R., & Spivey, M. J. (2008). Action dynamics reveal parallel competition in decision making. *Psychological Science, 19*(1), 22-24. doi:10.1111/j.1467-9280.2008.02041.x

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. doi:10.1177/1745691616662243

Mirman, D., & Magnuson, J. S. (2008). Attractor dynamics and semantic neighborhood density: Processing is slowed by near neighbors and speeded by distant neighbors. *Journal of Experimental Psychology, 34*(1), 65-79. doi:10.1037/0278-7393.34.1.65

Mittlböck, M., & Heinzl, H. (2006). A simulation study comparing properties of heterogeneity measures in meta-analyses. *Statistics in Medicine, 25*(24), 4321-4333.

Moeller, S. K., Robinson, M. D., & Zabelina, D. L. (2008). Personality dominance and preferential use of the vertical dimension of space: Evidence from spatial attention paradigms. *Psychological Science, 19*(4), 355-361. doi:10.1111/j.1467-9280.2008.02093.x

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The Prisma Group. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLOS Medicine, 6*(7), e1000097. doi:10.1371/journal.pmed.1000097

Moreno, S. G., Sutton, A. J., Abrams, K. R., Cooper, N. J., Turner, E. H., Palmer, T. M., & Ades, A. E. (2009). Novel methods to deal with publication biases: Secondary analysis of antidepressant trials in the FDA trial registry database and related journal publications. *BMJ, 339*(b2981). doi:10.1136/bmj.b2981

Moreno, S. G., Sutton, A. J., Ades, A. E., Stanley, T. D., Abrams, K. R., Peters, J. L., & Cooper, N. J. (2009). Assessment of regression-based methods to adjust for publication bias through a comprehensive simulation study. *BMC Medical Research Methodology, 9*(2). doi:10.1186/1471-2288-9-2

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association, 78*(381), 47-55.

Murayama, K., Pekrun, R., & Fiedler, K. (2014). Research practices that can prevent an inflation of false-positive rates. *Personality and Social Psychology Review, 18*(2), 107-118. doi:10.1177/1088868313496330

Murray, S. L., Derrick, J. L., Leder, S., & Holmes, J. G. (2008). Balancing connectedness and self-protection goals in close relationships: A levels-of-processing perspective on risk regulation. *Journal of Personality and Social Psychology, 94*(3), 429-459. doi:10.1037/0022-3514.94.3.429

Nairne, J. S., Pandeirada, J. N., & Thompson, S. R. (2008). Adaptive memory: The comparative value of survival processing. *Psychological Science, 19*(2), 176-180. doi:10.1111/j.1467-9280.2008.02064.x

Neuliep, J. W., & Crandall, R. (1993). Everyone was wrong - There are lots of replications out there. *Journal of Social Behavior and Personality, 8*(6), 1-8.

Niemeyer, H., van Aert, R. C. M., Schmid, S., Uelsmann, D., Knaevelsrud, C., & Schulte-Herbrueggen, O. (2018). Publication bias in meta-analyses of posttraumatic stress disorder interventions. Manuscript submitted for publication.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., . . . Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422-1425. doi:10.1126/science.aab2374

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 1-22. doi:10.3758/s13428-015-0664-2

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 estimates. *Review of General Psychology, 19*(2), 172-182. doi:10.1037/gpr0000034

Nurmsoo, E., & Bloom, P. (2008). Preschoolers' perspective taking in word learning: Do they blindly follow eye gaze? *Psychological Science, 19*(3), 211-215. doi:10.1111/j.1467-9280.2008.02069.x

O'Boyle, E. H., Gonzalez-Mule, E., & Banks, G. C. (2017). The Chrysalis effect: How ugly initial results metamorphosize into beautiful articles. *Journal of Management, 43*(2), 376-399. doi:10.1177/0149206314527133

Oberauer, K. (2008). How to say no: Single- and dual-process theories of short-term recognition tested on negative probes. *Journal of Experimental Psychology, 34*(3), 439-459. doi:10.1037/0278-7393.34.3.439

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi:10.1126/science.aac4716

Pacton, S., & Perruchet, P. (2008). An attention-based associative account of adjacent and nonadjacent dependency learning. *Journal of Experimental Psychology, 34*(1), 80-96. doi:10.1037/0278-7393.34.1.80

Palmer, S. E., & Ghose, T. (2008). Extremal edges: A powerful cue to depth perception and figure-ground organization. *Psychological Science, 19*(1), 77-84. doi:10.1111/j.1467-9280.2008.02049.x

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*(6), 531-536. doi:10.1177/1745691612463401

Paule, R. C., & Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards, 87*(5), 377-385.

Pawitan, Y. (2013). *In all likelihood : Statistical modelling and inference using likelihood*. Oxford: OUP Oxford.

Payne, B. K., Burkley, M. A., & Stokes, M. B. (2008). Why do implicit and explicit attitude tests diverge? The role of structural fit. *Journal of Personality and Social Psychology, 94*(1), 16-31. doi:10.1037/0022-3514.94.1.16

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2007). Performance of the trim and fill method in the presence of publication bias and between-study heterogeneity. *Statistics in Medicine, 26*(25), 4544-4562. doi:10.1002/sim.2889

Peters, J. L., Sutton, A. J., Jones, D. R., Abrams, K. R., & Rushton, L. (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology, 61*(10), 991-996. doi:10.1016/j.jclinepi.2007.11.010

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research, 86*(1), 207-236. doi:10.3102/0034654315582067

Popper, K. R. (1959/2005). *The logic of scientific discovery* (2nd ed.). New York: Routledge Classics.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10*(9). doi:10.1038/nrd3439-c1

Purdie-Vaughns, V., Steele, C. M., Davies, P. G., Ditlmann, R., & Crosby, J. R. (2008). Social identity contingencies: How diversity cues signal threat or safety for African Americans in mainstream institutions. *Journal of Personality and Social Psychology, 94*(4), 615-630. doi:10.1037/0022-3514.94.4.615

R Core Team. (2017). R: A language and environment for statistical computing. Retrieved from http://www.R-project.org/

Rabelo, A. L. A., 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. doi:10.1371/journal.pone.0134808

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.

Renkewitz, F., Fuchs, H. M., & Fiedler, S. (2011). Is there evidence of publication biases in JDM research? *Judgment and Decision Making, 6*(8), 870-881.

Rhodes, K. M., Turner, R. M., & Higgins, J. P. (2015). Predictive distributions were developed for the extent of heterogeneity in meta-analyses of continuous outcome data. *Journal of Clinical Epidemiology, 68*(1), 52-60. doi:10.1016/j.jclinepi.2014.08.012

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ, 340*. doi:10.1136/bmj.c221

Risen, J. L., & Gilovich, T. (2008). Why people are reluctant to tempt fate. *Journal of Personality and Social Psychology, 95*(2), 293-307. doi:10.1037/0022-3514.95.2.293

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641.

Rothstein, H. R., & Bushman, B. J. (2012). Publication bias in psychological science: Comment on Ferguson and Brannick (2012). *Psychological Methods, 17*(1), 129-136. doi:10.1037/a0027128

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

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

Röver, C., Knapp, G., & Friede, T. (2015). Hartung-Knapp-Sidik-Jonkman approach and its modification for random-effects meta-analysis with few studies. *BMC Medical Research Methodology, 15*. doi:10.1186/s12874-015-0091-1

Rufibach, K. (2015). selectMeta: Estimation of weight functions in meta analysis. Retrieved from https://CRAN.R-project.org/package=selectMeta

Rukhin, A. L. (2013). Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society: Series B, 75*(3), 451-469. doi:10.1111/j.1467-9868.2012.01047.x

Sahakyan, L., Delaney, P. F., & Waldum, E. R. (2008). Intentional forgetting is easier after two "shots" than one. *Journal of Experimental Psychology, 34*(2), 408-414. doi:10.1037/0278-7393.34.2.408

Salanti, G. (2012). Indirect and mixed-treatment comparison, network, or multiple-treatments meta-analysis: many names, many benefits, many concerns for the next generation evidence synthesis tool. *Research Synthesis Methods, 3*(2), 80-97. doi:10.1002/jrsm.1037

Sarewitz, D. (2012). Beware the creeping cracks of bias. *Nature, 485*(7397), 149-150. doi:10.1038/485149a

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. *British Journal of Mathematical and Statistical Psychology, 62*(1), 97-128. doi:10.1348/000711007X255327

Schmidt, J. R., & Besner, D. (2008). The Stroop effect: why proportion congruent has nothing to do with congruency and everything to do with contingency. *Journal of Experimental Psychology, 34*(3), 514-523. doi:10.1037/0278-7393.34.3.514

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90-100. doi:10.1037/a0015108

Schnall, S., Benton, J., & Harvey, S. (2008). With a clean conscience: Cleanliness reduces the severity of moral judgments. *Psychological Science, 19*(12), 1219-1222. doi:10.1111/j.1467-9280.2008.02227.x

Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Cambridge, Massachusetts: Hogrefe & Huber.

Schwarzer, G., Carpenter, J., & Rücker, G. (2016). metasens: Advanced statistical methods to model and adjust for bias in meta-analysis. R package version 0.3-1. https://CRAN.R-project.org/package=metasens.

Shadish, W. R., & Haddock, C. K. (2009). Combining estimates of effect size. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 257-277). New York: Russell Sage Foundation.

Shnabel, N., & Nadler, A. (2008). A needs-based model of reconciliation: Satisfying the differential emotional needs of victim and perpetrator as a key to promoting

reconciliation. *Journal of Personality and Social Psychology, 94*(1), 116-132. doi:10.1037/0022-3514.94.1.116

Sidik, K., & Jonkman, J. N. (2002). A simple confidence interval for meta-analysis. *Statistics in Medicine, 21*(21), 3153-3159. doi:10.1002/sim.1262

Sidik, K., & Jonkman, J. N. (2005). Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society, 54*(2), 367-384. doi:10.1111/j.1467-9876.2005.00489.x

Sidik, K., & Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine, 26*(9), 1964-1981. doi:10.1002/sim.2688

Silliman, N. P. (1997). Hierarchical selection models with applications in meta-analysis. *Journal of the American Statistical Association, 92*(439), 926-936.

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

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2018, January 8). P-curve handles heterogeneity just fine [Web log message]. Retrieved from: http://datacolada.org/67.

Simonsohn, U. (2015, February 9). "The" effect size does not exist [Web log message]. Retrieved from: http://datacolada.org/33.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014a). P-curve and effect size: Correcting for publication bias using only significant results. *Perspectives on Psychological Science, 9*(6), 666-681. doi:10.1177/1745691614553988

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014b). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General, 143*(2), 534-547. doi:10.1037/a0033242

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2017, December 20). Why p-curve excludes ps>.05 [Web log message]. Retrieved from http://datacolada.org/61.

Smith, T. C., Spiegelhalter, D. J., & Thomas, A. (1995). Bayesian approaches to random-effects meta-analysis: A comparative study. *Statistics in Medicine, 14*(24), 2685-2699.

Springate, D. A., & Kontopantelis, E. (2014). Cochrane_scraper: tools for downloading data from the Cochrane Library of Systematic Reviews (Version 1.1.0). Retrieved from https://github.com/DASpringate/Cochrane_scraper

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science, 8*(5), 581-591. doi:10.1177/1948550617693062

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods, 5*(1), 60-78.

Stanley, T. D., Doucouliagos, H., & Ioannidis, J. P. (2017). Finding the power to reduce publication bias. *Statistics in Medicine*. doi:10.1002/sim.7228

Stanley, T. D., Jarrell, S. B., & Doucouliagos, H. (2010). Could it be better to discard 90% of the data? A statistical paradox. *The American Statistician, 64*(1), 70-77. doi:10.1198/tast.2009.08205

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology, 94*(4), 672-695. doi:10.1037/0022-3514.94.4.672

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance-or vice versa. *Journal of the American Statistical Association, 54*(285), 30-34. doi:10.2307/2282137

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*(1), 108-112. doi:10.2307/2684823

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

Sterne, J. A. C., Bradburn, M. J., & Egger, M. (2001). Meta-analysis in Stata. In M. Egger, G. Davey Smith, & D. G. Altman (Eds.), *Systematic reviews in health care: Meta-analysis in context* (2nd ed.). London: BMJ Books.

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology, 54*(10), 1046-1055.

Sterne, J. A. C., & Egger, M. (2005). Regression methods to detect publication and other bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis: Prevention, assessment and adjustments* (pp. 99-110). Chichester, UK: Wiley.

Sterne, J. A. C., Gavaghan, D., & Egger, M. (2000). Publication and related bias in meta-analysis: Power of statistical tests and prevalence in the literature. *Journal of Clinical Epidemiology, 53*(11), 1119-1129. doi:10.1016/S0895-4356(00)00242-0

Sterne, J. A. C., Harbord, R. M., Sutton, A. J., Jones, D. R., Ioannidis, J. P., Terrin, N., . . . Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *British Medical Journal, 343*(7818), 1-8. doi:http://dx.doi.org/10.1136/bmj.d4002

Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation & the Health Professions, 25*(1), 76-97. doi:10.1177/016327870202500106

Stijnen, T., Hamza, T. H., & Ozdemir, P. (2010). Random effects meta-analysis of event outcome in the framework of the generalized linear mixed model with applications in sparse data. *Statistics in Medicine, 29*(29), 3046-3067. doi:10.1002/sim.4040

Straits, B. C., & Singleton, R. A. (2011). *Social research: Approaches and fundamentals*. New York: Oxford University Press.

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

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (pp. 435-452). New York: Russell Sage Foundation.

Sutton, A. J., Song, F., Gilbody, S. M., & Abrams, K. R. (2000). Modelling publication bias in meta-analysis: A review. *Statistical Methods in Medical Research, 9*(5), 421-445. doi:10.1177/096228020000900503

Switzer, F. S., Paese, P. W., & Drasgow, F. (1992). Bootstrap estimates of standard errors in validity generalization. *Journal of Applied Psychology Journal of Applied Psychology, 77*(2), 123-129. doi:10.1037//0021-9010.77.2.123

Terrin, N., Schmid, C. H., & Lau, J. (2005). In an empirical evaluation of the funnel plot, researchers could not visually identify publication bias. *Journal of Clinical Epidemiology, 58*(9), 894-901. doi:10.1016/j.jclinepi.2005.01.006

Terrin, N., Schmid, C. H., Lau, J., & Olkin, I. (2003). Adjusting for publication bias in the presence of heterogeneity. *Statistics in Medicine, 22*(13), 2113-2126. doi:10.1002/sim.1461

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine, 18*(20), 2693-2708.

Tsilidis, K. K., Papatheodorou, S. I., Evangelou, E., & Ioannidis, J. P. (2012). Evaluation of excess statistical significance in meta-analyses of 98 biomarker associations with cancer risk. *Journal of the National Cancer Institute, 104*(24), 1867-1878. doi:10.1093/jnci/djs437

Turk-Browne, N. B., Isola, P. J., Scholl, B. J., & Treat, T. A. (2008). Multidimensional visual statistical learning. *Journal of Experimental Psychology, 34*(2), 399-407. doi:10.1037/0278-7393.34.2.399

Turner, R. M., Jackson, D., Wei, Y., Thompson, S. G., & Higgins, J. P. T. (2015). Predictive distributions for between-study heterogeneity and simple methods for their

application in Bayesian meta-analysis. *Statistics in Medicine, 34*(6), 984-998. doi:10.1002/sim.6381

Turner, R. M., Omar, R. Z., Yang, M., Goldstein, H., & Thompson, S. G. (2000). A multilevel model framework for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine, 19*(24), 3417-3432.

Ueno, T., Fastrich, G. M., & Murayama, K. (2016). Meta-analysis to integrate effect sizes within an article: Possible misuse and Type I error inflation. *Journal of Experimental Psychology: General, 145*(5), 643-654. doi:10.1037/xge0000159

Ulrich, R., & Miller, J. (2015). p-hacking by post hoc selection with multiple opportunities: Detectability by skewness test?: Comment on Simonsohn, Nelson, and Simmons (2014). *Journal of Experimental Psychology: General, 144*(6), 1137-1145. doi:10.1037/xge0000086

United States Code (2007) US Public Law 110-85: Food and Drug Administration Amendments Act 2007. http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=110_cong_public_laws&docid=f:publ085.110.pdf.

van Aert, R. C. M., Maassen, E., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Estimating effect size in meta-analysis in the presence of publication bias and questionable research practices. Manuscript in preparation.

van Aert, R. C. M., & van Assen, M. A. L. M. (2016). *Examining reproducibility in psychology: A hybrid method for statistically combining a biased original study and replication*. Manuscript submitted for publication.

van Aert, R. C. M., & van Assen, M. A. L. M. (2017). Bayesian evaluation of effect size after replicating an original study. *PLoS ONE, 12*(4), e0175302. doi:10.1371/journal.pone.0175302

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2016). Conducting meta-analyses 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 Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L. M. (2018). Publication bias in meta-analyses from psychology and medicine: A meta-meta-Analysis. Manuscript submitted for publication.

van Assen, M. A. L. M., van Aert, R. C. M., Nuijten, M. B., & Wicherts, J. M. (2014). Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS ONE, 9*(1), e84896. doi:10.1371/journal.pone.0084896

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:http://dx.doi.org/10.1037/met0000025

van Dijk, E., van Kleef, G. A., Steinel, W., & van Beest, I. (2008). A social functional approach to emotions in bargaining: When communicating anger pays and when it backfires. *Journal of Personality and Social Psychology, 94*(4), 600-614. doi:10.1037/0022-3514.94.4.600

van Erp, S. J., Verhagen, J., Grasman, R. P. P. P., & Wagenmakers, E. J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin from 1990-2013. *Journal of Open Psychology Data, 5*(1). doi:https://doi.org/10.5334/jopd.33

van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine, 21*(4), 589-624.

Veldkamp, C. L. S., Bakker, M., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Soderberg, C. K., . . . Wicherts, J. M. (2017). Restriction of opportunistic use of researcher degrees of freedom in pre-registrations on the Open Science Framework. Retrieved from https://psyarxiv.com/g8cjq

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., . . . Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods, 7*(1), 55-79. doi:10.1002/jrsm.1164

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*(3), 419-435. doi:10.1007/bf02294384

Vevea, J. L., & Woods, C. M. (2005). Publication bias in research synthesis: Sensitivity analysis using a priori weight functions. *Psychological Methods, 10*(4), 428-443. doi:10.1037/1082-989X.10.4.428

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. doi:10.3102/10769986030003261

Viechtbauer, W. (2007a). 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. doi:10.3102/1076998606298034

Viechtbauer, W. (2007b). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine, 26*(1), 37-52.

Viechtbauer, W. (2007c). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology, 60*, 29-60. doi:10.1348/000711005X64042

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1-48.

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological Methods, 20*(3), 360-374. doi:10.1037/met0000023

Vohs, K. D., & Schooler, J. W. (2008). The value of believing in free will: Encouraging a belief in determinism increases cheating. *Psychological Science, 19*(1), 49-54. doi:10.1111/j.1467-9280.2008.02045.x

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science, 19*(7), 645-647. doi:10.1111/j.1467-9280.2008.02136.x

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review, 14*(5), 779-804. doi:10.3758/BF03194105

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632-638. doi:10.1177/1745691612463078

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology, 100*(3), 426-432. doi:10.1037/a0022790

Walter, S. D., & Cook, R. J. (1991). A comparison of several point estimators of the odds ratio in a single 2 x 2 contingency table. *Biometrics, 47*(3), 795-811. doi:10.2307/2532640

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 *t* tests. *Perspectives on Psychological Science, 6*(3), 291-298. doi:10.1177/1745691611406923

White, P. A. (2008). Accounting for occurrences: A new view of the use of contingency information in causal judgment. *Journal of Experimental Psychology, 34*(1), 204-218. doi:10.1037/0278-7393.34.1.204

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist, 61*(7), 726-728. doi:10.1037/0003-066X.61.7.726

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*(1832). doi:10.3389/fpsyg.2016.01832

Wiksten, A., Rücker, G., & Schwarzer, G. (2016). Hartung-Knapp method is not always conservative compared with fixed-effect meta-analysis. *Statistics in Medicine, 35*(15), 2503-2515. doi:10.1002/sim.6879

Winawer, J., Huk, A. C., & Boroditsky, L. (2008). A motion aftereffect from still photographs depicting motion. *Psychological Science, 19*(3), 276-283. doi:10.1111/j.1467-9280.2008.02080.x

Witkovský, V. (2013). A note on computing extreme tail probabilities of the noncentral t-distribution with large noncentrality parameter. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica, 52*(2), 131-143. doi:http://dml.cz//handle/10338.dmlcz/143545

Wolfram Research Inc. (2015). Mathematica. Champaign, Illinois: Wolfram Research, Inc.

Yap, M. J., Balota, D. A., Tse, C. S., & Besner, D. (2008). On the additive effects of stimulus quality and word frequency in lexical decision: Evidence for opposing interactive influences revealed by RT distributional analyses. *Journal of Experimental Psychology, 34*(3), 495-513. doi:10.1037/0278-7393.34.3.495

Zhang, M., & Li, X. (2012). From physical weight to psychological significance: The contribution of semantic activations. *Journal of Consumer Research, 38*(6), 1063-1075. doi:10.1086/661768

Dankwoord

In dit laatste hoofdstuk wil ik graag mensen bedanken die direct of indirect een bijdrage geleverd hebben aan de totstandkoming van mijn proefschrift. Ik lees weleens dankwoorden van proefschriften die een vrij zwaarmoedig karakter hebben, maar ik kan alvast verklappen dat daar bij mij geen sprake van is. Ik heb de afgelopen vier jaar en vijf maanden met veel plezier ervaren!

Ten eerste wil ik graag mijn promoteres Marcel en Jelte bedanken. Marcel, bedankt voor de grote hoeveelheid tijd die je in mij hebt gestoken. Ook met dingen die niet direct relevant waren voor het afronden van mijn proefschrift zoals de hulp die je me vooral tijdens de eerste jaren bij het zijn van AIO hebt gegeven bij het maken van sommetjes. Ik probeer nog steeds jouw advies op te volgen om iedere week wat sommetjes calculus te maken om hier beter in te worden! Vanzelfsprekend zouden zonder jouw hulp de hoofdstukken uit mijn proefschrift niet de kwaliteit hebben die ze nu hebben en zullen ongetwijfeld nog meer reviewers klagen dat we de literatuur vervuilen met ons werk mocht het gepubliceerd worden. Verder waardeer ik je eerlijkheid en persoonlijke adviezen die je me geeft, want werk is uiteindelijk toch maar (een belangrijke) bijzaak. Jelte, bedankt voor de velen afspraken om te praten over onderzoek en het reviseren van manuscripten. Jouw kritische blik naar de hoofdstukken in mijn proefschrift hebben deze zeker beter gemaakt. In het bijzonder wil ik je bedanken voor het nog een keer helemaal doornemen van mijn proefschrift vlak voor dat het ingeleverd moest worden. Dit moet een hels karwei geweest zijn.

I would also like to thank Wolfgang, because your course on meta-analysis caught my attention for this topic. We started our collaboration five years ago as part of an internship of my Research Master's studies. Although the project we initially started with is a very long-lasting project, I am still enjoying it and learned a lot from our collaboration. I would also like to thank you for bringing me in contact with Dan that gave me the opportunity to visit him for three months in Cambridge. Dan, thank you for the time you invested (and are still investing) in our research projects. I have never met a person who is as thorough as you are with respect to conducting research. This is a very good characteristic for a scientist! I would also like to thank all the others at the MRC Biostatistics Unit and especially the PhD students for involving me in the group and activities.

Ook wil ik graag Prof. Wim Van Den Noortgate, Prof. Eric-Jan Wagenmakers, Prof. Pim Cuijpers, Dr. Elise Dusseldorp en Dr. Marjan Bakker bedanken voor het zitting nemen in de promotiecommissie en de moeite nemen om hiervoor af te reizen naar Tilburg.

Mandy, bedankt dat je me wilde helpen bij het design van mijn proefschrift. Door jouw hulp ziet het er mooi en professioneel uit!

Het werken aan mijn proefschrift was zeker niet zo leuk geweest zonder de mooie onderzoeksgroep die we hebben. De groep begon vrij klein en bestond uit Jelte, Marcel, Coosje, Michèle toen Paulette en ik erbij kwamen, maar is nu flink uitgebreid met Chris, Marjan, Hilde, Andrea, Esther, Olmo, Amir en Leonie. In onze meetings (met

koekjes) komen we vaak niet toe aan het bespreken van de papers die op de planning staan, maar dit maakt deze meetings des te belangrijker. Ook wil ik jullie bedanken voor de feedback die jullie gegeven hebben op verschillende papers die nu hoofdstukken in mijn proefschrift zijn geworden.

De prettige werksfeer bij het MTO-departement heeft ook zeker bijgedragen bij het succesvol afronden van mijn proefschrift. Als eerste wil ik mijn (voormalige) kamergenoten bedanken: Fetene, Lianne, Chris, Niek en Paul. Chris, ik wil jou in het bijzonder ook bedanken voor de velen hamburgers die we samen gegeten hebben en de kamers die we gedeeld hebben bij het bezoeken van een congres of cursus. Zoals de meesten wellicht zullen weten ben ik niet de persoon die heel graag naar congressen gaat. Gelukkig ging er vaak een grote groep collega's naar hetzelfde congres wat het uiteindelijk toch nog een leuke ervaring maakte. Ook wil ik (zonder namen te noemen en het risico te lopen mensen te vergeten) alle collega's bedanken voor de leuke gesprekken tijdens de lunch. Daarnaast wil ik ook Marcel, Joris, Paulette, Sara, Guy en Luc bedanken voor de prettige samenwerking bij het geven van onderwijs. Guy, jou wil ik in het bijzonder bedanken voor je hulp en advies bij het maken van opdrachten/tentamens en het voorbereiden en geven van een aantal hoorcolleges. Ik heb hier erg veel van geleerd wat zeker van pas gaat komen in de toekomst. Bij het bedanken van collega's mogen ook zeker onze (voormalige) secretaresses Marieke, Liesbeth en Anne-Marie niet ontbreken. Jullie waren altijd in staat om bijvoorbeeld mijn vragen te beantwoorden over onderwerpen als formulieren en het boeken van tickets/hotels.

Tijdens de afgelopen vier jaar en vijf maanden heb ik ook met veel plezier deelgenomen aan verschillende sportieve activiteiten. Dit waren de meest uiteenlopende sporten namelijk van paintboksen tot het rennen van de Tilburg Ten Miles en van het Vroege Vogels tennistoernooi tot een beach soccer toernooi. Ik wil graag alle collega's bedanken die meegedaan hebben met deze activiteiten. Ook wil ik me nogmaals verontschuldigen voor alle collega's die hier niet op zitten te wachten, maar wel telkens door mij benaderd werden om deel te nemen. Verder wil ik alle mensen bedanken die regelmatig deelnemen aan de tennistrainingen op woensdag- en vrijdagmiddag. In het bijzonder wil ik Michael bedanken voor de verschillende tennistoernooien waar we soms succesvol en soms minder succesvol aan deelgenomen hebben de afgelopen jaren.

Ik wil ook graag Daphne, Esther en Jeannette bedanken waarmee ik samen heb gestudeerd tijdens mijn bacheloropleiding Personeelwetenschappen en nog steeds contact mee heb. Ook al zeggen jullie dat mijn werk eigenlijk verspilling van belastinggeld is en is steevast de vraag als we elkaar zien welke snoepreisjes ik nu weer heb gemaakt of van plan ben te maken, het blijft toch altijd leuk als we weer gaan kamperen, er een landendiner is of we een andere activiteit ondernemen.

Cindy en Niek bedankt dat jullie mijn paranimfen willen zijn. Cindy, ik waardeer je doorzettingsvermogen om te bereiken wat je graag wilt bereiken in het

leven en vind het fijn dat jij mijn oudere (maar kleinere) zus bent. Niek, ik vind het erg gezellig dat wij nu samen met Chris en Paul een kantoor delen. Ik kan je humor erg waarderen en ook je enthousiasme om deel te nemen aan hardloopwedstrijden (ook al realiseer jij je soms dat het na het maken van een trainingsschema toch wat krap wordt).

Ook wil ik graag Kees, Betsie, Karlijn en Pepijn bedanken voor de gastvrijheid voor al meer dan 10 jaar. Ik waardeer het erg als we langskomen en bordspellen spelen ook al is er soms wat kritiek op mijn speelwijze ;-).

Aan mijn moeder heb ik wellicht het meeste te danken en dan niet alleen de afgelopen vier jaar en vijf maanden, maar de afgelopen 28 jaar. Mamma, ik vind het erg knap dat je mij en ook Cindy voor het grootste gedeelte alleen hebt opgevoed en ons vooral hebt gestimuleerd om de dingen te doen die we leuk vonden. Ik denk dat we allebei uiteindelijk goed terecht gekomen zijn! Pappa, Marleen, Ramon en Shanti, het is goed dat we weer wat frequenter contact hebben.

Tenslotte wil ik de belangrijkste persoon in mijn leven bedanken. Lieve Karin, niet alleen de afgelopen 4 jaar en 5 maanden heb jij me ontzettend veel geluk gebracht, maar ook geholpen om te relativeren over hoe (on)belangrijk werk wel niet is. Ik denk dat we alle problemen kunnen overwinnen, omdat we een erg goed team zijn samen. Ook al lijkt het soms dat we het over de meeste dingen oneens zijn, in werkelijkheid denk ik dat we het toch over meer zaken eens dan oneens zijn. Ik hoop dat we nog velen jaren samen gelukkig en gezond door kunnen brengen.