Bootstrap-Based Between-Study Heterogeneity Tests in Meta-Analysis

Han Du\(^a\), Ge Jiang\(^b\), and Zijun Ke\(^c\)

\(^a\)Department of Psychology, University of California, Los Angeles; \(^b\)Department of Educational Psychology, University of Illinois at Urbana-Champaign; \(^c\)Department of Psychology, Sun Yat-sen University

**ABSTRACT**

Meta-analysis combines pertinent information from existing studies to provide an overall estimate of population parameters/effect sizes, as well as to quantify and explain the differences between studies. However, testing between-study heterogeneity is one of the most challenging tasks in meta-analysis research. Existing methods for testing heterogeneity, such as the Q test and likelihood ratio (LR) test, have been criticized for their failure to control Type I error rate and/or failure to attain enough statistical power. Although better reference distribution approximations have been proposed in the literature, their application is limited. Additionally, when the interest is to test whether the size of the heterogeneity is larger than a specific level, existing methods are far from mature. To address these issues, we propose new heterogeneity tests. Specifically, we combine bootstrap methods with existing heterogeneity tests (i.e., the maximum LR test, the restricted maximum LR test, and the Q test) to overcome the reference distribution issue and denote them as B-ML-LRT, B-REML-LRT, and B-Q, respectively. Simulation studies were conducted to examine and compare the performance of the proposed methods with the regular LR test, the regular Q test, and the Kulinskaya’s improved Q test in both random- and mixed-effects meta-analyses. Based on the results of Type I error rates and statistical power, B-REML-LRT is recommended. Additionally, the improved Q test is also recommended when it is applicable. An R package \texttt{boot.heterogeneity} is provided to facilitate the implementation of the proposed tests.

**KEYWORDS**

Meta-analysis; between-study heterogeneity; bootstrap method

---

**Introduction**

Meta-analysis is a popular statistical technique that combines results across multiple studies (Glass et al., 1984; Hedges & Olkin, 1985; Hunter et al., 1982). Besides providing a pooled estimate of the unknown population effect sizes, modern meta-analysis methods also have the capacity to quantify and explain the differences between studies. Fixed-effects meta-analysis assumes that all studies have the same true effect, whereas random-effects and mixed-effects models assume that there is variability (i.e., between-study heterogeneity) across study-specific population/true effects. Researchers have suggested that one should decide whether to include the between-study heterogeneity based on the assumption and intended generalization of the obtained results (Borenstein et al., 2009b; Coughlin et al., 2015). More specifically, if one would like to generalize the conclusion from a specific set of studies, a model with between-study heterogeneity is preferred (another option is an unrestricted weighted least squares model; WLS\(^1\)).

Although we cannot rely on heterogeneity tests to select between fixed-effects models and random-effects models, testing between-study heterogeneity is a focus of interest in some studies and widely used in practice. There are two types of between-study heterogeneity tests. In the first type of test, the null hypothesis is “true effects are the same (no heterogeneity; \(\tau^2 = 0\))” and the alternative hypothesis is “true effects vary (\(\tau^2 > 0\)).” We refer to this type of test as the heterogeneity test. In the second type of test, the null hypothesis is that “the size of heterogeneity is equal to a specific level (\(\tau^2 = \lambda\))” and the alternative hypothesis is that “the size of heterogeneity is larger than a specific level (\(\tau^2 > \lambda\)).” We refer to this type of test as the heterogeneity magnitude test. Existing

\(^1\)Stanley et al. (2018) reported the use of 200 WLS meta-analyses and this review is not included in the current paper since we focus on fixed-effects, random-effects, and mixed-effects models. We refer interested readers to Stanley et al. (2018) for more discussions of the application of WLS.
heterogeneity tests have been criticized for their failure to control Type I error rate and/or to attain sufficient power. Additionally, despite extensive studies on heterogeneity tests, not much effort has been devoted to developing heterogeneity magnitude tests with a few exceptions (Hedges & Schauer, 2019; Schauer, 2018). The existing heterogeneity magnitude tests either do not directly test $\chi^2$ (e.g., Hedges and Schauer’s test), which makes the statistical inference less straightforward, or fail to control Type I error rate (e.g., confidence intervals of $F$ and $H$).

The importance of the two types of between-study heterogeneity tests is three-fold. First, “in a meta-analysis, it is usual to conduct a homogeneity test to determine whether the effects measured by the included studies are sufficiently similar to justify their combination” (Kulinskaya et al., 2011a, p. 254). When the between-study heterogeneity is large, researchers may be more interested in the study-specific effect size (e.g., Al Khalaf et al., 2011). Field (2003) illustrated one example. The effect sizes of cognitive behavior therapy for social phobia in the United Kingdom, United States, and Netherlands were found to be different, but the overall effect size across countries might be zero. In this case, interpreting the overall effect size may not provide any practical meaning, whereas estimating the between-study heterogeneity is more informative and “it is sometimes useful to test the hypothesis that the effect size variance is zero in addition to estimating this quantity” (Hedges & Olkin, 1985, p. 197). Furthermore, in the heterogeneity magnitude test, with a significant test result, we can claim that the heterogeneity is larger than a specific level (e.g., a small, medium, or large effect size, or a scientifically or clinically meaningful size).2

Second, one may be interested in what factors could explain the between-study heterogeneity. To address this research question, relevant study-level covariates (also referred to as moderators) are included in a mixed-effects model (meta-regression) to explain whether there are systematic reasons why effect sizes are different (Field, 2003). For example, studies on gender differences in marital satisfaction may have discrepancies due to cohort effects. Marriage may look very different today than it did 10 or 20 years ago. Thus, the year in which each study was initiated can be considered as a moderator to explain why effect sizes vary across studies. We do not anticipate that the covariates can explain the entire residual heterogeneity, but “the size of this residual or unexplained heterogeneity is one consideration in evaluating the adequacy of the moderation model” (Card, 2015, p. 218). The test of (unexplained) heterogeneity in meta-regression allows for the evaluation of more complex moderation hypotheses. “For example, one can test interactive combinations of moderators by creating product terms. Similarly, one can evaluate nonlinear moderation by the creation of power polynomial terms” (Card, 2015, p. 218). Borenstein et al., (2009b) called this test of unexplained heterogeneity the goodness of fit test (p. 198).

Third, heterogeneity tests are particularly important in replication research. When one wants to know whether the replication is exact (studies have identical effect sizes), heterogeneity tests can be used (Hedges & Schauer, 2019). However, it is usually too strict to require exact replications in scientific practice, and therefore we can test whether the heterogeneity is negligible (approximate replication; Hedges & Schauer, 2019; Schauer & Hedges, 2020). "Small differences in the magnitude of effects may not lead to different interpretations of a finding", so that the replications can be treated “almost the same” (Hedges & Schauer, 2019, p. 559). In this case, the heterogeneity magnitude test can be used, although some researchers may prefer to estimate the between-study heterogeneity variance using a confidence interval rather than conduct a heterogeneity magnitude test (e.g., Van Aert et al., 2019).

Tests of between-study heterogeneity are widely used in real data analysis. We reviewed studies published in the flagship journal Psychological Bulletin. Specifically, we searched for articles through PsycINFO using the keywords “meta-regression”, “mixed-effects model”, and “meta-analysis” from January 1, 2014 to March 15, 2019. The initial search returned 160 articles, 34 of which were identified as reply articles, comment articles, or erratum announcements. Those 34 articles were set aside for reference, bringing down the total number of articles to 126. Out of the 126 articles, no article only used a fixed-effects model, 3 articles (2.4%) used both random-effects and fixed-effects models, and the remaining 123 articles (97.6%) all chose the random-effects models or mixed-effects models. Seventy-seven articles out of the 126 articles (61.1%) reported the results of testing between-study heterogeneity (with no declining trend from 2014 to 2019), suggesting that tests of between-study heterogeneity are widely used in real meta-analyses. Among the 77 articles, 75 used the Q test (Hedges, 1983; Hedges & Olkin, 1985), one used a likelihood ratio (LR) test, and one used a Wald test.

---

1We refer interested readers to Schauer and Hedges (2020) for more discussions of the criteria of negligible differences.
Despite the theoretical and practical significance of studying between-study heterogeneity, testing between-study heterogeneity is one of the most challenging tasks in meta-analysis research. Existing heterogeneity (magnitude) tests have three main limitations. First, some of the existing methods, such as LR tests (Viechtbauer, 2007), are limited to testing for the homogeneity of effect sizes ($H_0 : \tau^2 = 0$) and they cannot test whether the heterogeneity is larger than a specific level (i.e., heterogeneity magnitude test). Second, existing methods have been criticized for their failure to control Type I error rate and/or to attain enough statistical power (e.g., Borenstein et al., 2009a; Huedo-Medina et al., 2006). The main reason is that the reference distributions of the Q test and the LR test do not have exact analytic expressions, and the asymptotic approximation requires large study-specific sample sizes. Third, little research has studied the empirical performance of heterogeneity tests in mixed-effects meta-analysis (i.e., with covariates). To overcome these limitations, we propose bootstrap-based heterogeneity tests that combine the LR test or Q test with bootstrap simulation. That is, for each of the proposed tests, we will use bootstrap simulations to construct a reference distribution and obtain critical values from the empirical reference distribution.

The outline of this paper is as follows. In the “Meta-analysis models and effect sizes” section, we provide an overview of meta-analysis models and effect sizes. The “Heterogeneity testing in meta-analysis” section presents existing methods for testing between-study heterogeneity, some of which motivate the current study. In the “The proposed method: a bootstrap-based heterogeneity test” section, we present the proposed bootstrap-based heterogeneity tests and develop an R package to facilitate their implementation. In the “Simulation studies” section, we thoroughly examine the Type I error rate and statistical power of the bootstrap-based heterogeneity tests with different types of effect sizes via simulations. In the “Empirical illustration” section, we provide real data examples to illustrate the bootstrap-based heterogeneity tests. We end the paper with some concluding remarks.

Meta-analysis models and effect sizes

Suppose $K$ unbiased effect size estimates $x_j$ ($j = 1, ..., K$) are included in a meta-analysis. A typical random-effects meta-analysis model (e.g., Hedges & Olkin, 1985) is given by

\[ x_j = \delta_j + e_j, \quad e_j \sim N\left(0, \sigma_e^2\right) \]

\[ \delta_j = \mu + u_j, \quad u_j \sim N\left(0, \tau^2\right) \]

where $e_j$ is the deviation of the observed study effect size $x_j$ from the true/population study effect size $\delta_j$, and its variance $\sigma_e^2$ represents the within-study sampling variability of study $j$, $u_j$ is the deviation of the true study effect size $\delta_j$ from the true/population overall effect size $\mu$, and $\tau^2$ represents between-study heterogeneity ($\tau^2 = 0$ in a fixed-effects meta-analysis model). The calculation of $\sigma_e^2$ varies depending on the type of effect size. Between-study discrepancies ($\tau^2$) could be explained by accounting for study-level covariates (e.g., sampling and design characteristics) to some degree. This model is referred to as a mixed-effects model or a meta-regression model (Card, 2015; Hedges & Vevea, 1998). When all the between-study discrepancies can be explained by the covariates and hence $\tau^2 = 0$, the model becomes a fixed-effects meta-regression model. In addition to fixed- and random-effects meta-analysis models and fixed- and random-effects meta-regression models, Stanley and Doucouliagos (2015, 2017) proposed an unrestricted weighted least squares (WLS) model. The WLS model is neither the fixed-effects model nor the random-effects model, and it can accommodate covariates (Stanley et al., 2018; Stanley & Jarrell, 2005).3

In the current article, we focus on three types of effect sizes: standardized mean differences, Pearson correlations, and odds ratios. When the effects are standardized mean differences, the effect size estimate is $g_j = \frac{\bar{y}_j - \bar{y}}{s}$, where $\bar{y}_j$ and $\bar{y}$ are the sample means of the two groups in study $j$, and $s_j = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$ where $s_{1j}$ and $s_{2j}$ are the sample variances of the two groups of study $j$ and $n_{1j}$ and $n_{2j}$ are the sample sizes of the two groups. However, $g_j$ is biased, especially when the per-study sample size is small (Hedges, 1981; Hedges & Olkin, 1985). The unbiased effect size estimate is $d_j = (1 - \frac{3}{4(n_{1j}+n_{2j})-9})g_j$ (Hedges, 1981).4 Using this effect size, $d_j$

3The WLS model does not estimate $\tau^2$ and assumes that the study variance $v_j$ can be estimated up to some unknown multiplicative constant $\phi_j$ so that $v_j = \phi_j^2$. Stanley and Doucouliagos (2015, 2017) found that the WLS model outperformed the conventional random-effects meta-analysis when there is publication bias and outperformed the fixed-effects model when there is heterogeneity.

4We used the notation in the original work of Hedges (1981) and Hedges and Olkin (1985). The notations vary across different researchers. For example, Borenstein et al. (2009b) called the biased estimate the Cohen’s $d$ and the unbiased estimate the Hedge’s $g$. 
asymptotically follows $N(\delta_j, n_{j1}+n_{j2}/2(n_{j1}+n_{j2}))$ where $\delta_j$ is the study-specific true effect size (Hedges, 1981; Hedges & Olkin, 1985). In practice, we usually use $d_j$ to replace $\delta_j$ to calculate the within-study sampling variance.

When the effects are Pearson correlations, we first transform correlation in each study $r_j$ to a Fisher’s $z$ score $z_{r,j}$ by $z_{r,j} = \frac{1}{2} \ln \left( \frac{1+r_j}{1-r_j} \right)$ (Hedges & Olkin, 1985).

$z_{r,j}$ asymptotically follows $N\left( \delta_j, \frac{1}{n_j-3} \right)$ where $n_j$ is the per-study sample size and $\delta_j$ is the study-specific true effect size.

When the effects are odds ratios, the effect size estimate is $\hat{\alpha_j} = \frac{n_{11,j} / n_{01,j}}{n_{00,j} / n_{10,j}}$ where $n_{11,j}$ is the number of participants with $Y=1$ in Group 1, $n_{01,j}$ is the number of participants with $Y=0$ in Group 1, $n_{00,j}$ is the number of participants with $Y=1$ in Group 2, and $n_{10,j}$ is the number of participants with $Y=0$ in Group 2. Since odds ratios cannot be negative and hence the distribution of sample estimates is skewed, odds ratios are usually transformed by natural logarithm (i.e., $\log(\hat{\alpha}_j)$; log odds ratios). The approximated standard error of the log odds ratio is $\hat{\sigma}_j = \sqrt{1/n_{00,j} + 1/n_{10,j} + 1/n_{01,j} + 1/n_{11,j}}$ (Morris & Gardner, 1988).

### Heterogeneity testing in meta-analysis

We focus on the $Q$ test, the Kulinskaya’s improved $Q$ test, the LR test, and the bootstrap method in this section. There are also other heterogeneity tests and we illustrate them in the supplemental materials. To provide an overall picture of different methods, the strengths and limitations of each method are summarized in Table 1 and will be elaborated further in the following sections.

#### $Q$ test

The $Q$ test is historically the most widely used test. The test statistic $Q$ follows a $\chi^2$ distribution with $K-1$ degrees of freedom in the random-effects meta-analysis or with $K-P-1$ ($P$ is the number of covariates) degrees of freedom in the mixed-effects meta-analysis when the effect size is normally distributed and the sampling variance $\sigma_j^2$ is known (Hoaglin, 2016; Hedges, 1983; Hedges & Olkin, 1985). When $\sigma_j^2$ is unknown and a consistent estimator of $\sigma_j^2$ is used, $Q$ asymptotically approximates the $\chi^2$ distribution in large samples. However, “the approximation is not accurate for small and medium study-level sample sizes” (Kulinskaya et al., 2011a, p. 254). The Type I error rate and power of the $Q$ test have been widely studied with different types of effect size. Generally, when the study-level sample sizes are small and the number of studies is large, the Type I error rate and power deviate from the theoretical values (e.g., Chang, 1993; Harwell, 1997; Hedges & Olkin, 1985; Morris, 2000; Takkouche et al., 1999; Viechtbauer, 2007). The direction of deviation depends on the type of effect size. With a larger number of studies and larger study-level sample sizes, the Type I error rate of the $Q$ test approximates the nominal $z$ level (e.g., Harwell, 1997; Morris, 2000; Viechtbauer, 2007). Hedges and Pigott (2001) and Viechtbauer (2007) found that more studies, large study-level sample sizes, and larger $\tau^2$ led to higher power across different types of effect sizes; however, Harwell (1997) found that power values decreased with more studies, especially for smaller study-level sample sizes for standardized mean differences.

When $\tau^2 \neq 0$, the reference distribution of the $Q$ test is a weighted linear combination of chi-squared distributions (Hedges & Pigott, 2001) and can be approximated by a noncentral chi-squared distribution (Schauer, 2018). Hedges and Schauer (2019) and Schauer (2018) proposed to test the noncentrality parameter of the chi-squared distribution ($\omega = \sum_{j=1}^{K} (\hat{\delta}_j - \mu)^2/\sigma_j^2$). Hence, in a heterogeneity test, the null hypothesis is that $H_0: \omega = 0$; and in a heterogeneity magnitude test, the null hypothesis is that $H_0: \omega \leq \omega_0$. Note that $\omega$ is not $\tau^2$. When all studies have the same within-study sampling variance $\sigma^2$, $\tau^2 = \omega \sigma^2/\sigma_0^2$ and we can test $\tau^2$. When studies have different within-study sampling variances, linking $\omega$ with $\tau^2$ is more challenging.

Kulinskaya et al. (2004, 2011a, 2011b) and Kulinskaya and Dollinger (2015) pointed out that there are better approximations than $\chi^2_{K-1}$ for the $Q$ test in meta-analysis and that the approximated distribution of $Q$ depends on the type of effect size.\(^5\) When effect sizes are standardized mean differences and mean differences, Kulinskaya et al. (2004) and Kulinskaya et al. (2011b) either used a multiple of an $F$-distribution as a new reference distribution or approximated the degrees of freedom of the $\chi^2$ distribution using $E(Q)$. However, the improved $Q$ test

---

\(^5\) The improved approximations were proposed in the cases where the effects are not normally distributed (but can be asymptotically normally distributed) and the estimators of the weights are not independent of the estimators of the effects. But Kulinskaya still assumed that raw data are normally distributed.
needs the sample means and sample variances from each study, which may not be available in meta-analysis (as illustrated in the real data example). When effect sizes are risk differences and log odds ratios, Kulinskaya et al. (2011a) and Kulinskaya and Dollinger (2015) approximated the reference distribution of the Q test by matching Gamma distributions. Kulinskaya et al. (2004, 2011a, 2011b) and Kulinskaya and Dollinger (2015) found that the Type I error rate and statistical power from the improved Q test were much better than those from the regular Q test. The generalization of Kulinskaya’s approximation to other types of effect sizes needs to be resolved based on each specific effect size. As independent work, Breslow et al. (1980) proposed a $\chi^2$ test for odds ratios (no logarithm transformation needed) and the reference distribution is $\chi^2_{k-1}$.

### Table 1. Strengths and limitations of existing heterogeneity tests.

| Method                                | Strengths                                                                                      | Limitations                                                                                       |
|---------------------------------------|-------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
| Regular Q test                        | Widely used in different software and R packages                                               | Type I error rates are not appropriately controlled                                              |
|                                       |                                                                                                | Power can be low                                                                                 |
|                                       |                                                                                                | Can conduct heterogeneity magnitude test regarding a noncentrality parameter but not regarding $\tau^2$ |
| Kulinskaya’s improved Q test ($Q_2$) | Appropriately control Type I error rates                                                      | Require raw data information such as sample means and sample standard deviations                 |
|                                       | High power                                                                                     | Cannot handle effect size measures other than (standardized) mean differences, risk differences, and log odds ratios so far |
|                                       | R package available                                                                             | Cannot conduct heterogeneity magnitude test                                                     |
| Breslow–Day test                      | Appropriately control Type I error rates                                                      | Cannot conduct heterogeneity magnitude test                                                     |
|                                       | High power                                                                                     | Type I error rates are not appropriately controlled                                             |
|                                       | R package available                                                                             | Power can be low                                                                                 |
|                                       |                                                                                                | Cannot conduct heterogeneity magnitude test                                                     |
| Takkouche’s, Van den Noortgate–Onghena, and Sinha’s bootstrap tests | Appropriately control Type I error rates in the small scale simulations                         | Cannot be replicated given important details missing                                              |
|                                       |                                                                                                | Performance has not been examined by large scale simulations                                      |
| Bootstrap-based ML LR test (B-ML-LRT), bootstrap-based REML LR test (B-REML-LRT) | B-REML-LRT appropriately Type I error rates                                                   | Cannot incorporate covariates                                                                    |
|                                       | Higher power compared to the regular Q test                                                    | Cannot conduct heterogeneity magnitude test                                                     |
|                                       | Can incorporate covariates                                                                     | Can be less powerful than the improved Q test                                                    |
|                                       | Can be used in heterogeneity magnitude test                                                    | B-ML-LRT can fail to control Type I error rates                                                  |
|                                       | Can be easily generalized to other types of effect sizes                                       |                                                                                                |
|                                       | R package available                                                                             |                                                                                                |
| Bootstrap-based Q test (B-Q)          | Control Type I error rates relatively well                                                    | Can fail to control Type I error rates when the study-level sample sizes were small (SZ = 24) and the effect size was the log odds ratio |
|                                       | Higher power compared to the regular Q test                                                    | Can be less powerful than B-REML-LRT and the improved Q test                                    |
|                                       | Can incorporate covariates                                                                     |                                                                                                |
|                                       | Can be used in heterogeneity magnitude test                                                    |                                                                                                |
|                                       | Can be easily generalized to other types of effect sizes                                       |                                                                                                |
|                                       | R package available                                                                             |                                                                                                |

#### LR test

In meta-analysis, LR tests are not very widely used but frequently mentioned in the literature (Takkouche et al., 1999; Viechtbauer, 2007). LR tests for between-study heterogeneity are based on comparing two models: a full model in which $\tau^2$ is freely estimated (the random-effects model or the mixed-effects model) and a reduced model in which $\tau^2 = 0$ (the fixed-effects model or the fixed-effects meta-regression). There are two likelihood-based estimators: the maximum likelihood (ML) estimator and the restricted maximum likelihood (REML) estimator. Generally, REML is preferred over ML for variance component estimation because ML tends to underestimate variances especially when sample sizes are small (e.g., Fitzmaurice et al., 2012). Researchers such as Berkey et al. (1995) and Viechtbauer (2007) studied these two estimators
in meta-analysis and their simulation results echoed the findings in the general context.

There is a boundary (i.e., 0) in the variances’ parameter space when only nonnegative variances are allowed (i.e., constrained estimation). This is the so-called boundary issue. With constrained estimation, the reference distribution is a .5:.5 mixture of chi-squared distributions, and the critical value is 2.71 \((5\chi^2_0 + 5\chi^2_1)\) at \(z = .05\) (Stram & Lee, 1994, 1995; Stoele et al., 2006). Similar to the Q test, the simulation in Viechtbauer (2007) showed that a larger number of studies and larger study-level sample sizes, especially the latter, improved the Type I error rate, except in the Fisher-transformed correlation case (Viechtbauer, 2007). However, even with a large number of studies and large study-level sample sizes, the ML-based LR test (ML-LRT) still could be too conservative (i.e., smaller than the nominal level). In meta-analyses, the REML-based LR test (REML-LRT) has been found to be slightly better than the ML-based LR test (Viechtbauer, 2007). The statistical power of the Q test, ML-LRT, and REML-LRT has been found to be very similar (Viechtbauer, 2007).

**Bootstrap methods**

Because the approximation of the LR test is sensitive to the number of studies and study-level sample sizes, when the number of studies is small and/or the study-level sample sizes are small, the reference distribution may be far different from a .5:.5 mixture of \(\chi^2_0\) and \(\chi^2_1\). Takkouche et al. (1999) proposed a parametric bootstrap version of the LR test as follows: (1) draw with replacement from the original \(K\) studies; (2) estimate \(\mu\) and the sampling variance \(\sigma^2_j\) for each study within each bootstrap sample; (3) simulate \(K\) observed study effect sizes using a fixed-effects model based on the estimated \(\mu\) and \(\sigma^2_j\); (4) calculate the LR test statistic, \(LR^B\); (5) repeat steps 1–4 for \(B\) times; and (6) calculate the empirical \(p\)-value as the proportion of \(LR^B\)'s larger than the LR in the original sample. However, there are two things unclear in Takkouche et al. (1999). First, they did not describe which model was used to obtain \(\hat{\mu}\) (i.e., a fixed- or random-effects model). Second and more importantly, it is unclear whether ML-LRT or REML-LRT was used. For the odds ratio effect size, Takkouche et al. (1999) found that the Q test and the bootstrap-based LR test appropriately controlled the Type I error rate, but the regular LR test was too conservative. The power of the Q test and the bootstrap-based LR test were similar, whereas the LR test had lower power. Sinha et al. (2012) and Van den Noortgate and Onghena (2003) had independent work on parametric bootstrap tests for testing between-study heterogeneity. Their procedures were similar to that of Takkouche et al. (1999). We illustrate the details in the supplemental materials.

**The proposed method: a bootstrap-based heterogeneity test**

Given the importance of testing between-study heterogeneity, researchers have worked on proposing better heterogeneity tests. However, there are several limitations in the existing methods as summarized in Table 1. We briefly repeat these limitations below. First, existing methods focus on testing homogeneity (\(\tau^2 = 0\) versus \(\tau^2 \neq 0\)). More specifically, the .5:.5 mixture of chisquared distributions of ML-LRT and REML-LRT is not proposed for the null hypothesis of \(\tau^2 = \lambda\) (Stram & Lee, 1994, 1995) and the reference distribution is complex since we need to consider the boundary issue. The Q test can provide a heterogeneity magnitude test regarding the noncentrality parameter \(\omega\) but not regarding the between-study variance \(\tau^2\).

Second, despite the widespread use of different tests of heterogeneity, existing methods have been criticized for their failure to control Type I error rate and attain enough statistical power (e.g., Borenstein et al., 2009a; Huedo-Medina et al., 2006). Third, although Kulinskaya’s improved Q test performs better than the regular Q test, the improved approximations cannot handle effect size measures other than mean differences, standardized mean differences, risk differences, and log odds ratios. Additionally, the improved Q test requires the raw data information for (standardized) mean difference effect sizes, which is often unavailable in real meta-analyses. Fourth, the bootstrap version of the LR test in Takkouche et al. (1999) cannot be replicated given that important details are missing. In addition, the parametric bootstrap methods by Takkouche et al. (1999), Sinha et al. (2012), and Van den Noortgate and Onghena (2003) were only examined in small simulation studies. Fifth, the previous research on between-study heterogeneity tests mainly focuses on random-effects meta-analysis, and the mixed-effects meta-analysis has not received enough attention. In a mixed-effects model, the heterogeneity test detects the residual heterogeneity that is not accounted for by the moderators, which can be used to search for important moderators and to evaluate more complex moderation hypotheses.

We now propose parametric bootstrap-based heterogeneity tests. The basic procedure of our bootstrap method is to simulate the empirical reference
distribution of a chosen test statistic (i.e., ML-LRT, REML-LRT, or the Q test) to get a critical value for null hypothesis testing. This method can be used with or without covariates. When the bootstrap procedure is coupled with the ML-LRT, REML-LRT, or the Q test, we refer to it as B-ML-LRT, B-REML-LRT, or B-Q, respectively.\(^6\) The general procedure of testing \(\tau^2\) in a meta-analysis with \(K\) studies using the proposed method is as follows.

1. For the bootstrap-based LR tests, we calculate the ML- or REML-based LR test statistic using the original sample (i.e., \(LR^O\)). For the bootstrap-based Q test, we calculate the Q statistic using the original sample (i.e., \(Q^O\)).

2. We use the original sample sizes and covariate values, treat the REML estimates of \(\mu\) and the regression coefficients of covariates as the true parameters, and simulate \(K\) sample study effect sizes (\(\chi^2_j, j \in [1, K]\)) under the null hypothesis (e.g., \(H_0: \tau^2 = .05\) for a heterogeneity magnitude test). The sampling variances used to simulate \(\chi^2_j\) are the sampling variances in the original study.

3. Based on the simulated effect size \(\chi^2_j\), we re-estimate the sampling variances based on \(\chi^2_j\) in cases of the standardized mean difference and log odds ratio. In this way, the bootstrap test incorporates the randomness in the \(\sigma^2_j\) estimates and therefore it avoids the assumption in the regular heterogeneity tests that \(\sigma^2_j\) is either known or closely approximated. Then, we calculate the ML- or REML-based LR test statistic \(LR^B\) and the Q statistic \(Q^B\). Note that the heterogeneity magnitude test is a one-sided test (\(\tau^2 = \lambda\) versus \(\tau^2 > \lambda\)), and therefore when \(\tau^2\) in the simulated data was smaller than \(\lambda\), the difference between the log-likelihoods was counted as zero.\(^7\)

4. We repeat steps 2 and 3 \(B\) times.

5. The critical value \(LR_{critical}\) is the \((1-\alpha)\)-quantile of the \(B\) \(LR^B\)s where \(\alpha\) is the pre-specified Type I error rate, and the critical value \(Q_{critical}\) is the \((1-\alpha)\)-quantile of the \(B\) \(Q^B\)s.

6. In the heterogeneity test, the null assumption is rejected if \(LR^O\) is larger than \(LR_{critical}\) or \(Q^O\) is larger than \(Q_{critical}\). In the heterogeneity magnitude test, the null assumption is rejected if \(LR^O\) is larger than \(LR_{critical}\) or \(Q^O\) is larger than \(Q_{critical}\), and \(\tau^2\) in the original data is larger than \(\lambda\). When \(\tau^2\) in the original data is smaller than \(\lambda\), we fail to reject the null assumption.

Note that the log-likelihood function should match the LR test. That is, when the ML estimator is used to estimate \(\tau^2\), \(L_{\tau^2} = -2(L_{\hat{\mu}^O, \tau^2=0} - L_{\hat{\mu}, \tau^2})\) based on the regular log-likelihood function (\(L\) indicates the regular log-likelihood function, and \(\hat{\mu}\) and \(\tau^2\) indicate the ML estimates), and when the REML estimator is used to estimate \(\tau^2\), \(L_{\tau^2} = -2(L_{\hat{\tau}^2=0} - L_{\hat{\tau}^2})\) based on the restricted log-likelihood function (\(L^R\) indicates the restricted log-likelihood function, and \(\hat{\tau}^2\) indicates the REML estimate of \(\tau^2\)).

**R package for implementing the bootstrap-based between-study heterogeneity test**

We provide an R package for the bootstrap-based between-study heterogeneity test, `boot.heterogeneity`. This package can be downloaded from GitHub and CRAN (The Comprehensive R Archive Network). The code, example illustrations, and the help manual are in the supplemental materials. `boot.heterogeneity` builds on the metafor package by Viechtbauer (2010). It implements the heterogeneity test for standardized mean differences, Fisher-transformed Pearson correlations, and log odds ratios. For example, `boot.d` is a function for testing between-study heterogeneity with standardized mean differences.

```r
def(boot.d(n1, n2, est, model = "random", mods = NULL, nrep = 10^6, pcut = .05, adjust = FALSE, lambda = 0))
est in a vector of unbiased estimates of standardized mean differences from individual studies. If the biased estimates \(g_i\) are read in for \(est\), adjust = TRUE must be specified to obtain the corresponding unbiased standardized mean difference estimates \(d_i\). \(n1\) and \(n2\) read in two vectors of sample sizes from group 1 and group 2 in each of the
```

\(^6\)Different from Sinha et al. (2012) and Noortgate and Onghena (2003), we simulate bootstrap samples based on effect sizes’ asymptotic sampling distributions, whereas Sinha et al. (2012) and Noortgate and Onghena (2003) simulated raw data first and calculated the “simulated” effect sizes based on the simulated data.

\(^7\)When the effect size is the log odds ratio, the effect size is a function of four cell sizes. The four cell sizes in the original data will not be consistent with the simulated \(\chi^2\), therefore there is one more step. We keep three of the simulated cell sizes \(n_{00j}, n_{01j}, n_{11j}\), and \(n_{10j}\) the same as \(n_{00j}, n_{10j}, n_{11j}\), and \(n_{01j}\) in the original data, and the fourth one is updated based on \(\chi^2\). For example, we fix \(n_{00j}\) and \(n_{01j}\) to be the same as \(n_{00j}, n_{10j}, n_{11j}\), and \(n_{01j}\) is calculated as \(n_{11j} / \sum n_{11j} / \sum n_{00j}\). Each cell size has a 25% chance to be calculated across the permutations. Based on the calculated fourth cell size, the sampling variances of simulated studies are updated, and then we calculate the ML- or REML-based LR test statistic and the Q statistic.
Table 2. Levels of heterogeneity of $\tau^2$ used in the simulation.

| Sample size (SZ) | Levels of heterogeneity |
|------------------|-------------------------|
|                  | Small | Medium | Large |
| Standardized mean difference | 24   | .030   | .100  | .300  |
| Fisher-transformed Pearson correlation | 24   | .010   | .030  | .100  |
| Log odds ratio    | 91   | .006   | .020  | .050  |

Each function for testing between-study heterogeneity provides the results from the bootstrap-based REML-LRT (B-REML-LRT) and the Q test. In addition, for an odds ratio, its standard error will be infinite if any one of the four cells is zero. In this case, Haldane and Anscombe’s correction is used by adding .5 to each cell value automatically (Anscombe, 1956; Haldane, 1940).

Simulation studies

Simulation design

In this section, we explored the performance of the proposed bootstrap-based REML-LRT (B-REML-LRT), bootstrap-based ML-LRT (B-ML-LRT), and bootstrap-based Q test (B-Q). The proposed bootstrap-based methods were compared with the regular LR tests (ML-LRT and REML-LRT), the Q test, the improved Q test by Kulinskaya et al. (2011a, 2011b) and Kulinskaya and Dollinger (2015) ($Q_0$ was available for standardized mean differences and log odds ratios only), and the Breslow–Day test by Breslow et al. (1980) (BD was for odds ratios only). Specifically, results of ML-LRT and REML-LRT were compared to the critical value of 2.71.

Five factors that have been found to have an impact on the performance of the between-study heterogeneity tests were manipulated in the simulation study: (1) the type of effect size (standardized mean differences, Fisher-transformed Pearson correlations, or log odds ratios), (2) the number of studies ($K = 10, 20, 30, 50, 100, \text{ or } 300$), (3) the per-group study-level sample size (SZ = 24, 91, or 370), (4) the size of between-study heterogeneity $\tau^2$, and (5) whether covariates were included. The overall effect size $\mu$ was not manipulated because previous research has found that its influence on the performance of the between-study heterogeneity tests was negligible (e.g., Takkouche et al., 1999; Viechtbauer, 2005). Our pilot simulation results confirmed this. Please see the supplemental materials for more details. Therefore, we specified $\mu = 0$ in the simulation.

Sample sizes were chosen based on three real meta-analyses of standardized mean differences in order to mimic a realistic setting: a 55-study meta-analysis with per-group sample sizes ranging from 5 to 83 and a median of 24 (Deci et al., 1999), a 33-study meta-analysis with per-group sample sizes ranging from 10 to 489 and a median of 91 (Becker, 1986), and a 186-study meta-analysis with per-group sample sizes ranging from 12 to 96,267 and a median of 370 (Hyde et al., 1990). These chosen study-level sample sizes and numbers of studies are also representative values based on our literature review of the 126 studies in Psychological Bulletin.

The size of between-study heterogeneity $\tau^2$ was determined in the following way. Pigott (2012) provided benchmarks for the size of a relative heterogeneity index $I^2$, which is the proportion of total variation in the observed effect sizes that is due to the between-study heterogeneity (Higgins & Thompson, 2002): small (25%), medium (50%), and large (75%). We first calculated the average within-study variance per sample size set for each type of effect. Then, we computed $\tau^2$ based on the within-study variance and the above-mentioned benchmarks. Hence, the size of between-study heterogeneity is different for different types of effect size. The condition of $SZ = 370$ was not considered because the computed $\tau^2$ was too small based on the $I^2$ benchmarks (≤.001). We illustrate different levels of $\tau^2$ in Table 2. Additionally, we checked heterogeneity review papers to verify the appropriateness of $\tau^2$’s specification (Rhodes et al., 2015; van Erp et al., 2017 for standardized mean differences; Gunhan et al., 2020; Turner et al., 2015 for log odds ratios; van Erp et al., 2017 for correlations). These articles supported our specifications.

In the mixed-effects meta-analyses, the study-level covariates were simulated from a normal distribution $N(0, 1)$ and the regression coefficient for each study-level covariate $\beta$ was specified to .5. The number of replications was 1000 for each condition. The study-specific true effect size was simulated by $\delta_j \sim$
N(\mu, \tau^2) without covariates or \( \delta_j \sim N(\mu + \beta Z_j, \tau^2) \) with one covariate. The number of bootstrap samples, B, was set to 10,000 to find the critical value of the LR test or Q test.\(^8\) We consider a Type I error rate between .025 and .075 as satisfactory (Bradley, 1978).

In the simulation of standardized mean differences, we randomly drew K sample sizes from the aforementioned three real meta-analysis data sets. Based on each simulated \( \delta_j \), the raw data were simulated as \( y_{ij} \sim N(0, 1) \) and \( y_{2ij} \sim N(\delta_j, 1) \) where \( i \) indicates the \( i \)th individual and \( j \) indicates the \( j \)th study. Based on the sample means and sample variances of each study, we calculated the observed standardized mean differences.

In the simulation of Fisher-transformed Pearson correlations, we randomly drew K per-group sample sizes from one of the two groups in the real meta-analysis data. The simulated study-specific true Fisher-transformed Pearson correlations \( \delta_j \) were converted to Pearson correlations \( \rho_j \). For each simulated study, the raw data were simulated as \( \begin{pmatrix} y_{1ij} \\ y_{2ij} \end{pmatrix} \sim MV(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_j \\ \rho_j & 1 \end{pmatrix}) \). Based on the simulated raw data, we calculated the observed Fisher-transformed Pearson correlations \( r_j \).

In the simulation of log odds ratio, we used the total sample sizes of the two groups in the real meta-analysis data sets as the total sample sizes of the odds ratios (\( N_j \)). We randomly drew \( n_{00j} \) and \( n_{01j} \) \((n_{00j}/N_j \sim \text{unif}(2, 3))\) and \( n_{01j}/N_j = 0.5 - n_{00j}/N_j \) to mimic the scenario that there is almost no treatment effect in the control group, and control and treatment groups have equal sample sizes), and calculated \( n_{10j} \) and \( n_{11j} \) based on the simulated study-specific true log odds ratio \( \delta_j \). The current \( n_{00j}, n_{10j}, n_{01j}, \) and \( n_{11j} \) were the true cell sizes without sampling error. Accordingly, the estimated proportions of the four cells \( p_{00j} = n_{00j}/N_j, \)
\[ p_{01j} = n_{01j}/N_j, \]
\[ p_{10j} = n_{10j}/N_j, \]
\[ p_{11j} = n_{11j}/N_j \]
were calculated. We used binomial distributions to simulate the raw data, add sampling error, and update \( n_{00j}, n_{01j}, n_{10j}, n_{11j}, \) and \( n_{11j.update} \sim \text{binomial}(n_{00j} + n_{01j}, n_{01j}/n_{00j}) \),
\[ n_{00j.update} = n_{00j} + n_{10j} - n_{01j.update}, \]
\[ n_{11j.update} \sim \text{binomial}(n_{10j} + n_{11j}, n_{11j}/n_{10j}), \]
\[ n_{00j} + n_{01j} - n_{11j.update} \]. The updated cell sizes contained sampling errors. Based on the simulated raw data, we calculated the observed log odds ratios.

As pilot studies, we examined the Type I error rates using confidence intervals of the heterogeneity indexes \( I^2 \) and \( H \) coupled with different estimators of \( \tau^2 \). The Type I error rates were too low across all combinations. Please see the supplemental materials for more details.

**Type I error rates**

**Type I error rates in random-effects models for heterogeneity tests**

In this section, we evaluate the performance of the heterogeneity tests when the true size of heterogeneity is zero (i.e., \( \tau^2 = 0 \)). The Type I error rates of the bootstrap-based ML-LRT (B-ML-LRT), the bootstrap-based REMEL-LRT (B-REML-LRT), the bootstrap-based Q test (B-Q), the improved Q test by Kulinskaya and Dollinger (2015) (Q2), and the Breslow–Day test by Breslow et al. (1980) for odds ratios (BD), the regular ML-based LR test (ML-LRT), the regular REMEL-based LR test (REML-LRT), and the regular Q test are presented in Table 3 with their corresponding Monte Carlo standard errors. ML-LRT produced too conservative Type I error rates in most conditions. Consistent with the results from Viechtbauer (2007), more studies and larger study-level sample sizes improved the Type I error rates of ML-LRT, but even with a relatively large number of studies and study-level sample sizes (e.g., \( \text{SZ} = 370 \) and \( K = 50 \)), ML-LRT was too conservative in the examined conditions because ML tended to underestimate \( \tau^2 \). REMEL-LRT performed better than ML-LRT, but it also could provide too conservative Type I error rates, especially with log odds ratios. The Q test could fail to control Type I error rates when the study-level sample sizes were small (\( \text{SZ} = 24 \)). Especially in the case of the log odds ratio, with small study-level sample sizes, a large number of studies failed to improve the Type I error rates of the Q test, ML-LRT, and REMEL-LRT; in contrast, more studies tended to decrease the Type I error rates. The improved Q test appropriately controlled the Type I error rates across different conditions of study-level sample size, number of studies, and type of effect size. The Breslow–Day test also appropriately controlled the Type I error rates for odds ratios. The two bootstrap LR methods, B-ML-LRT and B-REML-LRT, appropriately controlled the Type I error rates regardless of the type of effect size, the number of studies, and the study-level sample sizes (but when \( \text{SZ} = 24, K = 10 \), and the effect

---

\(^8\)In our simulation and real data analyses, we found \( 10^4 \) bootstrap samples provided relatively stable critical values.
sizes were Fisher-transformed Pearson correlations, all of the bootstrap methods had inflated Type I error rates although still smaller than .075). Their performances were similar. The bootstrap-based Q test (B-Q) appropriately controlled the Type I error rates in most of the examined conditions except when effect sizes were log odds ratios and the study-level sample sizes were small (SZ = 24). Compared to the Monte Carlo standard errors that were smaller than .01, the methods that appropriately controlled the Type I error rates and the methods that failed to control Type I error rates yielded quite different Type I error rates.
The difference between their Type I error rates should not be just due to random errors.

**Type I error rates in mixed-effects models for heterogeneity tests**

Here we evaluate the performance of heterogeneity tests when the true size of residual heterogeneity is zero, i.e., $\tau^2 = 0$. The Type I error rates of B-ML-LRT, B-REML-LRT, B-Q, ML-LRT, REML-LRT, and the Q test with one covariate are presented in Table 4 with their corresponding Monte Carlo standard errors. The improved Q test and Breslow–Day test are not applicable in mixed-effects models. Similar to the model without covariates, ML-LRT was too conservative in most examined conditions, regardless of which type of effect size was of interest and the number of covariates. REML-LRT and the Q test also could be too conservative when effect sizes of interest were log odds ratios. A larger number of studies could further decrease the conservative Type I error rates of ML-LRT, REML-LRT, and the Q test when the effect sizes were log odds ratios and the study-level sample sizes were small ($SZ = 24$). B-Q generally performed well, except when the study-level sample sizes were small ($SZ = 24$) and the effect sizes were log odds ratios. Unlike the cases without covariates, B-ML-LRT could be too conservative when the study-level sample sizes were large ($SZ = 370$). Similar to ML-LRT, we think the underestimated $\tau^2$ causes the conservative performance of B-ML-LRT. B-REML-LRT outperformed B-ML-LRT and B-Q, and appropriately controlled the Type I error rates regardless of the type of effect size, the number of studies, and the study-level sample sizes. Since the Monte Carlo standard errors were all smaller than .01, the difference between the methods that appropriately controlled the Type I error rates and the methods that failed to control Type I error rates was not due to random errors.

**Type I error rates in random-effects models for heterogeneity magnitude tests**

We explore the Type I error rates of the proposed methods in testing whether the heterogeneity is larger than a specific level ($\tau^2 = \lambda$ versus $\tau^2 > \lambda$) in this section. The null hypothesis was $H_0 : \tau^2 = \lambda$ and $\lambda$ was specified as the small level of heterogeneity in Table 2 to mimic real data. The Type I error rates of B-ML-LRT, B-REML-LRT, and B-Q are presented in Table 5 with their corresponding Monte Carlo standard errors. B-Q appropriately controlled the Type I error rates, except when the study-level sample sizes were small ($SZ = 24$) and effect sizes were log odds ratios. B-ML-LRT and B-REML-LRT appropriately controlled the Type I error rates across all conditions.

**Summary of the performance of the tests on Type I error rates**

In sum, across different conditions and types of effect sizes, although the performance of REML-LRT and Q test was better than ML-LRT, the three methods could be too conservative in the examined conditions and they are not applicable in the heterogeneity magnitude test for testing $\tau^2$. Without covariates, the improved Q test and the Breslow–Day test appropriately controlled the Type I error rates in the heterogeneity test but they are not applicable in the heterogeneity magnitude test. The bootstrap methods are applicable in both the heterogeneity test and the heterogeneity magnitude test. Among the three bootstrap methods, B-REML-LRT was superior to the other two, and B-Q could be too liberal when effect sizes were log odds ratios. B-REML-LRT was competitive with the improved Q test and the Breslow–Day test when the improved Q test and the Breslow–Day test were applicable.

**Statistical power**

REML-LRT, ML-LRT, and B-Q will not be discussed further due to their failure to control Type I error rates. We varied the between-study heterogeneity ($\tau^2$) as mentioned above (see Table 2 for the small, medium, and large levels of heterogeneity) and considered the same simulation conditions as in examining Type I error rates, except the cases where the power values from B-ML-LRT, B-REML-LRT, the improved Q test, the Breslow–Day test, and the Q test were all nearly 1.

**Statistical power in random-effects models for heterogeneity tests**

In random-effects meta-analyses, the statistical power values of B-ML-LRT, B-REML-LRT, the improved Q test ($Q_2$), the Breslow–Day test, and the Q test are presented in Table 6, with the highest power values under each condition bolded and the Monte Carlo standard errors inside the parentheses. More studies, larger study-level sample sizes, and/or larger $\tau^2$ increased power, which supported the findings from Hedges and Pigott (2001) and Viechtbauer (2007), but not Harwell (1997). Regarding the relative performance, the improved Q test, B-REML-LRT, and B-ML-LRT were the three methods that provided higher statistical power. Among these three methods, however, no single method uniformly outperformed the other two methods.
In some conditions, as a parametric method, the improved Q test was more powerful than B-REML-LRT and B-ML-LRT, whereas in the other conditions, B-REML-LRT and B-ML-LRT outperformed the improved Q test. The proposed bootstrap-based methods provided heterogeneity tests with enhanced power when compared to the regular Q test. Specifically, when the effect of interest was the standardized mean difference, the maximum power increments of B-ML-LRT and B-REML-LRT compared to the regular Q test were .117 and .119, respectively, which were much larger than the maximum power increments of B-ML-LRT and B-REML-LRT outperformed the improved Q test. The proposed bootstrap-based methods provided heterogeneity tests with enhanced power when compared to the regular Q test.
increments of B-ML-LRT and B-REML-LRT were 34.0% and 31.6%, respectively (the medium increments were 14.3% and 14.5%, respectively). When effect sizes were Fisher-transformed Pearson correlations, the maximum power increments of B-ML-LRT and B-REML-LRT were .076 and the maximum percentage increments of B-ML-LRT and B-REML-LRT were 17.6% (the medium increments were 6.3%). When effect sizes were log odds ratios, the maximum power increments of B-ML-LRT and B-REML-LRT were .148 and .149, respectively, and the maximum percentage increments of B-ML-LRT and B-REML-LRT were 67.1% and 66.4%, respectively (the medium increments were 11.1% and 11.3%, respectively).

Statistical power in mixed-effects models for heterogeneity tests
In mixed-effects meta-analyses, the statistical power values of B-ML-LRT, B-REML-LRT, and the Q test with one covariate are presented in Table 7. The highest power values under each condition are bolded and the Monte Carlo standard errors are presented inside the parentheses. Similar to the cases without covariates, more studies, larger study-level sample sizes, and/or larger $\tau^2$ produced higher power. Regarding the relative performance, the proposed bootstrap-based methods outperformed the regular Q test. Specifically, when the effect of interest was the standardized mean difference, the maximum power increments of B-ML-LRT and B-REML-LRT compared to the Q test were .098 and .1, respectively, and the maximum percentage increments of B-ML-LRT and B-REML-LRT were 25.8% and 26.6%, respectively (the medium increments were 10.2% and 8%, respectively). When effect sizes were Fisher-transformed Pearson correlations, the maximum power increments of both B-ML-LRT and B-REML-LRT were .108 and .107, respectively, and the maximum percentage increments of both B-ML-LRT and B-REML-LRT were 23.8% and 23.6%, respectively (the medium increments were 5.2% and 5.4%, respectively). When effect sizes were log odds ratios, the maximum power increments of B-ML-LRT and B-REML-LRT were .176 and .177, respectively, and the maximum percentage increments of B-ML-LRT and B-REML-LRT were 54.5% and 54.8%, respectively (the medium increments were 11.3% and 11.6%, respectively).

| Sample size (SZ) | Number of studies (K) | B-ML-LRT | B-REML-LRT | B-Q | B-ML-LRT | B-REML-LRT | B-Q |
|-----------------|-----------------------|----------|------------|-----|----------|------------|-----|
|                 | Standardized mean difference | Log odds ratio |
| 24              | 10                     | .049 (.007) | .048 (.007) | .048 (.007) | .056 (.007) | .056 (.007) | .066 (.008) |
|                 | 20                     | .047 (.007) | .045 (.007) | .048 (.007) | .059 (.007) | .059 (.007) | .063 (.008) |
|                 | 30                     | .057 (.007) | .059 (.007) | .058 (.007) | .043 (.006) | .043 (.006) | .075 (.008) |
|                 | 50                     | .05 (.007)  | .050 (.007) | .059 (.007) | .036 (.006) | .036 (.006) | .062 (.008) |
|                 | 100                    | .047 (.007) | .048 (.007) | .046 (.007) | .051 (.007) | .051 (.007) | .085 (.009) |
|                 | 300                    | .056 (.007) | .056 (.007) | .055 (.007) | .043 (.006) | .043 (.006) | .118 (.010) |
| 91              | 10                     | .038 (.006) | .039 (.006) | .040 (.006) | .049 (.007) | .049 (.007) | .049 (.007) |
|                 | 20                     | .053 (.007) | .052 (.007) | .057 (.007) | .040 (.006) | .041 (.006) | .044 (.006) |
|                 | 30                     | .045 (.007) | .044 (.006) | .039 (.006) | .053 (.007) | .054 (.007) | .062 (.008) |
|                 | 50                     | .054 (.007) | .055 (.007) | .046 (.007) | .051 (.007) | .052 (.007) | .060 (.008) |
|                 | 100                    | .061 (.008) | .062 (.008) | .067 (.008) | .055 (.007) | .054 (.007) | .070 (.008) |
|                 | 300                    | .043 (.006) | .043 (.006) | .053 (.007) | .060 (.008) | .060 (.008) | .084 (.009) |

Note: Monte Carlo standard errors are presented in the parentheses. All unacceptable Type I error rates (i.e., <.025 or >.075) are bold. ML or REML estimation sometimes did not have converged results, but the nonconvergence rates across all conditions were within .3%.

Statistical power in random-effects models for heterogeneity magnitude tests
We explore the power of the proposed methods in testing whether the heterogeneity is larger than a specific level ($\tau^2 = \lambda$ versus $\tau^2 > \lambda$) in this section. $\tau^2$ under the null hypothesis ($\lambda$) was specified as the small level of heterogeneity in Table 2 and the true $\tau^2$ that generated data was specified as the medium level of heterogeneity. The power values are presented in Table 8 with the highest power values under each conditionbolded.
Table 6. Statistical power of the bootstrap-based ML-LRT (B-ML-LRT), the bootstrap-based REML-LRT (B-REML-LRT), the improved Q test (Q₂), the Breslow–Day test (BD), and the regular Q test in random-effects models for heterogeneity tests.

| Number of studies | B-ML-LRT | B-REML-LRT | Q₂ | BD | Q-ML-LRT | Q-REML-LRT | Q₂ | BD | Q |
|-------------------|----------|------------|----|----|----------|------------|----|----|---|
|                   |          |            |    |    |          |            |    |    |   |
| **Standardized mean difference** |          |            |    |    |          |            |    |    |   |
| **SZ = 24**       |          |            |    |    |          |            |    |    |   |
| 10                | .174 (.012) | .174 (.012) | .214 (.013) | .147 (.011) | .542 (.016) | .541 (.016) | .530 (.016) | .514 (.016) |   |
| 20                | .304 (.015) | .305 (.015) | .341 (.015) | .248 (.014) | .795 (.013) | .796 (.013) | .780 (.013) | .766 (.013) |   |
| 30                | .342 (.015) | .342 (.015) | .390 (.015) | .279 (.014) | .885 (.01) | .886 (.010) | .875 (.010) | .852 (.011) |   |
| 50                | .503 (.016) | .507 (.016) | .560 (.016) | .420 (.016) |          |            |    |    |   |
| 100               | .742 (.014) | .743 (.014) | .794 (.013) | .625 (.015) |          |            |    |    |   |
|                   | .877 (.010) | .880 (.01) | .873 (.011) | .866 (.011) |          |            |    |    |   |
| **SZ = 91**       |          |            |    |    |          |            |    |    |   |
| 10                | .153 (.011) | .153 (.011) | .149 (.011) | .125 (.01) | .474 (.016) | .478 (.016) | .460 (.016) | .446 (.016) |   |
| 20                | .284 (.014) | .279 (.014) | .243 (.014) | .212 (.013) | .715 (.014) | .715 (.014) | .689 (.015) | .672 (.015) |   |
| 30                | .346 (.015) | .347 (.015) | .341 (.015) | .300 (.014) | .841 (.012) | .842 (.012) | .816 (.012) | .809 (.012) |   |
| 50                | .457 (.016) | .458 (.016) | .441 (.016) | .400 (.015) | .957 (.006) | .960 (.006) | .955 (.007) | .941 (.007) |   |
| 100               | .694 (.015) | .695 (.015) | .653 (.015) | .577 (.016) |          |            |    |    |   |
|                   | .799 (.013) | .798 (.013) | .791 (.013) | .791 (.013) |          |            |    |    |   |
| **Fisher-transformed Pearson correlation** |          |            |    |    |          |            |    |    |   |
| **SZ = 24**       |          |            |    |    |          |            |    |    |   |
| 10                | .161 (.012) | .158 (.012) |          | .145 (.011) | .396 (.015) | .398 (.015) |          | .370 (.015) |   |
| 20                | .225 (.013) | .223 (.013) |          | .195 (.013) | .575 (.016) | .578 (.016) |          | .548 (.016) |   |
| 30                | .286 (.014) | .282 (.014) |          | .256 (.014) | .737 (.014) | .740 (.014) |          | .690 (.015) |   |
| 50                | .348 (.015) | .348 (.015) |          | .296 (.014) | .887 (.010) | .887 (.010) |          | .849 (.011) |   |
| 100               | .615 (.015) | .615 (.015) |          | .539 (.016) | .994 (.002) | .994 (.002) |          | .986 (.004) |   |
|                   | .815 (.012) | .815 (.012) |          | .802 (.013) |            |            |    |    |   |
| 20                | .965 (.006) | .965 (.006) |          | .963 (.006) |          |            |    |    |   |
| **SZ = 91**       |          |            |    |    |          |            |    |    |   |
| 10                | .318 (.015) | .315 (.015) |          | .289 (.014) | .755 (.014) | .756 (.014) |          | .735 (.014) |   |
| 20                | .512 (.016) | .512 (.016) |          | .461 (.016) | .942 (.007) | .942 (.007) |          | .931 (.008) |   |
| 30                | .624 (.015) | .626 (.015) |          | .586 (.016) | .985 (.004) | .985 (.004) |          | .980 (.004) |   |
| 50                | .791 (.013) | .791 (.013) |          | .744 (.014) |            |            |    |    |   |
| 100               | .970 (.005) | .970 (.005) |          | .945 (.007) |          |            |    |    |   |
|                   | .946 (.007) | .948 (.007) |          | .942 (.007) |          |            |    |    |   |
| **Log odds ratio** |          |            |    |    |          |            |    |    |   |
| **SZ = 24**       |          |            |    |    |          |            |    |    |   |
| 10                | .115 (.010) | .115 (.010) | .132 (.011) | .110 (.010) | .087 (.009) | .308 (.015) | .310 (.015) | .341 (.015) | .140 (.011) | .245 (.014) |
| 20                | .141 (.011) | .142 (.011) | .182 (.012) | .173 (.012) | .110 (.010) | .430 (.016) | .432 (.016) | .480 (.016) | .216 (.013) | .355 (.015) |
| 30                | .192 (.012) | .190 (.012) | .226 (.013) | .196 (.013) | .125 (.01) | .560 (.016) | .565 (.016) | .627 (.015) | .256 (.014) | .481 (.016) |
| 50                | .249 (.014) | .248 (.014) | .284 (.014) | .249 (.014) | .149 (.011) | .725 (.014) | .725 (.014) | .794 (.013) | .355 (.015) | .652 (.015) |
| 100               | .396 (.015) | .397 (.015) | .463 (.016) | .426 (.016) | .248 (.014) | .898 (.010) | .898 (.010) | .938 (.008) | .914 (.009) | .818 (.012) |
|                   | .658 (.015) | .662 (.015) | .718 (.014) | .686 (.015) | .601 (.015) |            |    |    |   |
| 20                | .856 (.011) | .854 (.011) | .899 (.010) | .881 (.010) | .822 (.012) |          |    |    |   |
| 30                | .941 (.007) | .942 (.007) | .963 (.006) | .951 (.007) | .915 (.009) |          |    |    |   |

(Continued)
condition bolded and the Monte Carlo standard errors inside the parentheses. Based on Monte Carlo standard errors, B-ML-LRT and B-REML-LRT did not provide noticeably different power.

**Summary of the performance of the tests on statistical power**

In sum, regardless of the method and type of effect size, more studies, larger study-level sample sizes, and/or larger \( \tau^2 \) provided higher power. The bootstrap-based methods boosted power compared to the regular Q test. The increments of power were the highest when log odds ratios were of interest. The improved Q test demonstrated the highest power in a number of conditions when it was applicable.

**Empirical illustrations**

We applied the proposed bootstrap-based heterogeneity test in three real meta-analyses. Our null hypothesis is \( \tau^2 = 0 \). We considered the bootstrap-based REML-LRT (B-REML-LRT) since it generally appropriately controlled Type I error rates. Data sets used in the three examples along with the R codes are provided in the supplemental materials.

The first meta-analysis consists of 13 studies that examined the correlation between sensation seeking scores and levels of monoamine oxidase (Zuckerman, 1994). The sample sizes ranged from 10 to 125 with a median of 40. In this example, the REML estimate of the overall Fisher-transformed Pearson correlation was -.26 and the REML estimate of the between-study variance was .03. The Q test result was \( Q(df = 12) = 29.06, p = .004 \). B-REML-LRT also rejected the assumption of homogeneity with \( p = .004 \).

The second meta-analysis consists of 18 studies, in which the effect of open versus traditional education on students’ self-concept was examined (Hedges et al., 1981). The per-group sample sizes ranged from 13.5 to 140 with a median of 64.5. In this example, the REML estimate of the overall standardized mean difference was .01 and the REML estimate of the between-study variance was .02. The Q test result was \( Q(df = 17) = 23.39, p = .137 \), and thus the test failed to reject the assumption of homogeneity when \( \alpha = .05 \). B-REML-LRT failed to reject the assumption of homogeneity with \( p = .053 \). The improved Q test is not applicable because the sample mean and sample variance information was not available in this meta-analysis.

The third meta-analysis consists of 26 studies on nicotine replacement therapy (nicotine chewing gum)
for smoking cessation (Silagy et al., 2004). The numbers of participants in the control group and treatment group (i.e., nicotine replacement therapy) who stopped smoking for at least 6 months after treatment were recorded. The total sample sizes for the control and treatment groups ranged from 47 to 1217 with a median of 175. The REML estimate of the overall log odds ratio was .56 and the REML estimate of the between-study variance was .05. The Q test result was $Q(df = 25) = 34.87, p = .091$, and thus the test failed to reject the assumption of homogeneity when $z = .05$. The improved Q test also failed to reject the assumption of homogeneity with $p = .073$. In contrast, B-REML-LRT rejected the assumption of homogeneity with $p = .037$. This real data analysis result was consistent with the simulation results: when effect sizes were log odds ratios and $SZ = 91$, B-REML-LRT could be the most powerful method.

**Discussion**

Testing between-study heterogeneity is of interest in some studies and is widely used in practice. We considered two types of between-study heterogeneity tests. In the heterogeneity test, we test $\tau^2 = 0$ versus $\tau^2 \neq 0$. In the heterogeneity magnitude test, we test $\tau^2 = \lambda$. 

---

**Table 7. Statistical power of the bootstrap-based ML-LRT (B-ML-LRT), the bootstrap-based REML-LRT (B-REML-LRT), and the regular Q test in mixed-effects models for heterogeneity tests.**

| Number of studies | B-ML-LRT | B-REML-LRT | Q | B-ML-LRT | B-REML-LRT | Q |
|------------------|----------|------------|---|----------|------------|---|
| $SZ = 24$ | $10.173 (.012)$ | $.169 (.012)$ | $.144 (.011)$ | $.457 (.016)$ | $.459 (.016)$ | $.425 (.016)$ |
| 20 | $.265 (.014)$ | $.263 (.014)$ | $.220 (.013)$ | $.720 (.014)$ | $.726 (.014)$ | $.696 (.015)$ |
| 30 | $.365 (.015)$ | $.365 (.015)$ | $.303 (.015)$ | $.852 (.011)$ | $.851 (.011)$ | $.832 (.012)$ |
| 50 | $.473 (.016)$ | $.476 (.016)$ | $.376 (.015)$ | |
| 100 | $.701 (.014)$ | $.701 (.014)$ | $.603 (.015)$ | |

---

**Note:** Monte Carlo standard errors are presented in the parentheses. $SZ$ is the average per-group sample sizes. The largest power values under each condition are highlighted as bold. ML or REML estimation sometimes did not have converged results, but the nonconvergence rates across all conditions were within 4%.
versus $\tau^2 > \lambda$. The heterogeneity test can be viewed as a special case of heterogeneity magnitude test with $\lambda = 0$. Researchers may be interested in the heterogeneity test and the heterogeneity magnitude test for three reasons. First, with a large between-study heterogeneity, interpreting study-specific effect size is more practically meaningful than the overall effect size. Second, researchers may also be interested in exploring factors that can explain the systematic between-study heterogeneity. The relevant test can be viewed as the goodness of fit test in an exploratory data analysis. If the unexplained heterogeneity in a mixed-effects model is significant, we may consider interactive combinations of moderators by creating product terms. Third, in replication research, the heterogeneity test can be used in an exact replication test and the heterogeneity magnitude test can be used in an approximate replication test. For both the heterogeneity test and the heterogeneity magnitude test, failure of controlling Type I error rates and low power will lead to misleading testing results. Hence, the goal of the current study is to propose a powerful and reliable method for both hypothesis tests.

Given the strengths and limitations of the existing methods in Table 1, we propose bootstrap-based heterogeneity tests combining the LR tests (ML- and REML-based) or the Q test with parametric bootstrap procedures. The proposed methods have several advantages over the existing approaches. First, the proposed bootstrap-based heterogeneity tests can be applied to the heterogeneity magnitude test while the existing heterogeneity magnitude tests either do not directly test $\tau^2$ or fail to control Type I error rates. Second, based on our simulation results, B-REML-LRT outperformed ML-LRT, REML-LRT, and the Q test in terms of controlling Type I error rate. B-REML-LRT appropriately controlled the Type I error rate and was competitive with the improved Q test. Third, the proposed bootstrap-based heterogeneity test increased statistical power compared to the regular Q test. The increment in power was the largest with log odds ratios.

Another strength of the proposed methods is their flexibility. Unlike the improved Q test, which requires an analytical form for each type of effect size and information of raw data that is often unavailable, the bootstrap-based heterogeneity tests can be easily extended to cases where effect sizes of interest are not discussed in the current paper, such as standardized mean changes. The only thing that would need

---

**Table 8.** Statistical power of the bootstrap-based ML-LRT (B-ML-LRT) and the bootstrap-based REML-LRT (B-REML-LRT) in random-effects models for heterogeneity magnitude tests.

| Number of studies | B-ML-LRT | B-REML-LRT | Number of studies | B-ML-LRT | B-REML-LRT |
|-------------------|----------|------------|-------------------|----------|------------|
|                   | $\lambda = 0.3$, $\tau^2 = 0.1$ | | $\lambda = 1$, $\tau^2 = 0.3$ |
| **SZ = 24**       |          |            |                  |          |            |
| 10                | 0.284 (.014) | 0.285 (.014) | 10                | 0.284 (.014) | 0.285 (.014) |
| 20                | 0.476 (.016) | 0.477 (.016) | 20                | 0.476 (.016) | 0.477 (.016) |
| 30                | 0.566 (.016) | 0.565 (.016) | 30                | 0.566 (.016) | 0.565 (.016) |
| 50                | 0.750 (.014) | 0.751 (.014) | 50                | 0.750 (.014) | 0.751 (.014) |
| **SZ = 91**       |          |            |                  |          |            |
| 10                | 0.243 (.014) | 0.247 (.014) | 10                | 0.284 (.014) | 0.284 (.014) |
| 20                | 0.418 (.016) | 0.418 (.016) | 20                | 0.450 (.016) | 0.450 (.016) |
| 30                | 0.528 (.016) | 0.528 (.016) | 30                | 0.607 (.015) | 0.608 (.015) |
| 50                | 0.702 (.014) | 0.704 (.014) | 50                | 0.775 (.013) | 0.776 (.013) |
| 100               | 0.911 (.009) | 0.911 (.009) | 100               | 0.955 (.007) | 0.955 (.007) |
| **SZ = 24**       |          |            |                  |          |            |
| 10                | 0.211 (.013) | 0.213 (.013) | 10                | 0.284 (.014) | 0.284 (.014) |
| 20                | 0.314 (.015) | 0.315 (.015) | 20                | 0.450 (.016) | 0.450 (.016) |
| 30                | 0.425 (.016) | 0.422 (.016) | 30                | 0.607 (.015) | 0.608 (.015) |
| 50                | 0.535 (.016) | 0.535 (.016) | 50                | 0.775 (.013) | 0.776 (.013) |
| 100               | 0.815 (.012) | 0.815 (.012) |                  |          |            |
| **SZ = 91**       |          |            |                  |          |            |
| 10                | 0.391 (.015) | 0.391 (.015) | 10                | 0.450 (.016) | 0.450 (.016) |
| 20                | 0.613 (.015) | 0.617 (.015) | 20                | 0.607 (.015) | 0.608 (.015) |
| 30                | 0.753 (.014) | 0.754 (.014) | 30                | 0.775 (.013) | 0.776 (.013) |
| 50                | 0.909 (.009) | 0.910 (.009) | 50                | 0.955 (.007) | 0.955 (.007) |

Note: Monte Carlo standard errors are presented in the parentheses. $\lambda$ indicates the heterogeneity level under the null hypothesis. $\tau^2$ is the true data generating heterogeneity level. SZ is the average per-group sample sizes. The largest power values under each condition are highlighted as bold. ML or REML estimation sometimes did not have converged results, but the nonconvergence rates across all conditions were within 4%.

### Footnotes

1. Fisher indicates the heterogeneity level under the null hypothesis.

2. $s^2_1$, $s^2_2$ are the true data generating heterogeneity level. SZ is the average per-group sample sizes. The largest power values under each condition are highlighted as bold. ML or REML estimation sometimes did not have converged results, but the nonconvergence rates across all conditions were within 4%.

3. $s^2_1 = \lambda$, $s^2_2 = \lambda$: The heterogeneity test can be viewed as a special case of heterogeneity magnitude test with $\lambda = 0$. Researchers may be interested in the heterogeneity test and the heterogeneity magnitude test for three reasons. First, with a large between-study heterogeneity, interpreting study-specific effect size is more practically meaningful than the overall effect size. Second, researchers may also be interested in exploring factors that can explain the systematic between-study heterogeneity. The relevant test can be viewed as the goodness of fit test in an exploratory data analysis. If the unexplained heterogeneity in a mixed-effects model is significant, we may consider interactive combinations of moderators by creating product terms. Third, in replication research, the heterogeneity test can be used in an exact replication test and the heterogeneity magnitude test can be used in an approximate replication test. For both the heterogeneity test and the heterogeneity magnitude test, failure of controlling Type I error rates and low power will lead to misleading testing results. Hence, the goal of the current study is to propose a powerful and reliable method for both hypothesis tests.

4. Given the strengths and limitations of the existing methods in Table 1, we propose bootstrap-based heterogeneity tests combining the LR tests (ML- and REML-based) or the Q test with parametric bootstrap procedures. The proposed methods have several advantages over the existing approaches. First, the proposed bootstrap-based heterogeneity tests can be applied to the heterogeneity magnitude test while the existing heterogeneity magnitude tests either do not directly test $\tau^2$ or fail to control Type I error rates. Second, based on our simulation results, B-REML-LRT outperformed ML-LRT, REML-LRT, and the Q test in terms of controlling Type I error rate. B-REML-LRT appropriately controlled the Type I error rate and was competitive with the improved Q test. Third, the proposed bootstrap-based heterogeneity test increased statistical power compared to the regular Q test. The increment in power was the largest with log odds ratios.
modification is the analytical form of sampling variances. We have summarized the strengths and limitations of the proposed bootstrap methods in Table 1.

Overall, based on both Type I error rate and statistical power, we recommend B-REML-LRT for the heterogeneity test and heterogeneity magnitude test. One also can consider the improved Q test when it is applicable. Future work could look into developing the bootstrap-based heterogeneity test for multivariate meta-analysis. The generalization is feasible when the analytical form of the sampling distribution of multivariate meta-analysis is available. Furthermore, published studies may not be truly representative of all valid studies undertaken on a research topic, which is known as publication bias or the file-drawer problem (Hunter & Schmidt, 2004; Iyengar & Greenhouse, 1988; Rosenthal, 1979). Conclusions based on the meta-analyses of published studies without correcting for publication bias compromise the validity of a systematic review (Du et al., 2017; Klein et al., 2018; Kvarven et al., 2020; Open Science Collaboration, 2015; Sutton et al., 2000). Future research should look into developing the bootstrap-based heterogeneity test while reducing the bias caused by the file-drawer problem.

### Article information

**Conflict of interest disclosures:** Each author signed a form for disclosure of potential conflicts of interest. No authors reported any financial or other conflicts of interest in relation to the work described.

**Ethical principles:** The authors affirm having followed professional ethical guidelines in preparing this work. These guidelines include obtaining informed consent from human participants, maintaining ethical treatment and respect for the rights of human or animal participants, and ensuring the privacy of participants and their data, such as ensuring that individual participants cannot be identified in reported results or from publicly available original or archival data.

**Funding:** This work was not supported.

**Role of the funders/sponsors:** None of the funders or sponsors of this research had any role in the design and conduct of the study; collection, management, analysis, and interpretation of data; preparation, review, or approval of the manuscript; or decision to submit the manuscript for publication.

**Acknowledgments:** The ideas and opinions expressed herein are those of the authors alone, and endorsement by the author’s institutions is not intended and should not be inferred.

### ORCID

Han Du [http://orcid.org/0000-0002-3174-6935](http://orcid.org/0000-0002-3174-6935)

### References

Anscombe, F. J. (1956). On estimating binomial response relations. *Biometrika*, 43(3–4), 461–464. [https://doi.org/10.1093/biomet/43.3-4.461](https://doi.org/10.1093/biomet/43.3-4.461)

Becker, B. J. (1986). Influence again: An examination of reviews and studies of gender differences in social influence. In J. S. Hyde & M. C. Linn (Eds.), *The psychology of gender: Advances through meta-analysis*. Johns Hopkins University Press.

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. [https://doi.org/10.1002/sim.4780140406](https://doi.org/10.1002/sim.4780140406)

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009a). Fixed-effect versus random-effects models. In M. Borenstein, L. V. Hedges, J. Higgins, & H. R. Rothstein (Eds.), *Introduction to meta-analysis* (pp. 77–86). Wiley.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2009b). *Introduction to meta-analysis*. Wiley.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, 31(2), 144–152. [https://doi.org/10.1111/j.2044-8317.1978.tb00581.x](https://doi.org/10.1111/j.2044-8317.1978.tb00581.x)

Breslow, N. E., Day, N. E., & Hoselton, E. (1980). *Statistical methods in cancer research* (Vol. 1). International Agency for Research on Cancer.

Card, N. A. (2015). *Applied meta-analysis for social science research*. Guilford Publications.

Chang, L. (1993). A power analysis of the test of homogeneity in effect-size meta-analysis [Unpublished doctoral dissertation]. Michigan State University.

Coughlin, C. G., Cohen, S. C., Mulqueen, J. M., Ferracioli-Oda, E., Stuckelman, Z. D., & Bloch, M. H. (2015). Meta-analysis: Reduced risk of anxiety with psychostimulant treatment in children with attention-deficit/hyperactivity disorder. *Journal of Child and Adolescent Psychopharmacology*, 25(8), 611–617.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6), 627–667.

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.

Field, A. P. (2003). The problems in using fixed-effects models of meta-analysis on real-world data. *Understanding Statistics*, 2(2), 105–124. [https://doi.org/10.1207/s15328031us0202_02](https://doi.org/10.1207/s15328031us0202_02)

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). *Applied longitudinal analysis* (Vol. 998). John Wiley & Sons.

Glass, G. V., MacGaw, B., & Smith, M. L. (1984). *Meta-analysis in social research*. SAGE.

Gunhan, B. K., Rover, C., & Friede, T. (2020). Random-effects meta-analysis of few studies involving rare events. *Research Synthesis Methods*, 11(1), 74–90.
Stanley, T. D., & Jarrell, S. B. (2005). Meta-regression analysis: A quantitative method of literature surveys. *Journal of Economic Surveys, 19*(3), 299–308. https://doi.org/10.1111/j.0950-0804.2005.00249.x

Stoel, R. D., Garre, F. G., Dolan, C., & Van Den Wittenboer, G. (2006). On the likelihood ratio test in structural equation modeling when parameters are subject to boundary constraints. *Psychological Methods, 11*(4), 439–455.

Stram, D., & Lee, J. (1995). Corrections: Variance component testing in the longitudinal mixed effects model. *Biometrics, 51*(3), 1196–1196.

Stram, D., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics, 50*(4), 1171–1177.

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.

Takkouche, B., Cadarso-Suarez, C., & Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology, 150*(2), 206–215. https://doi.org/10.1093/oxfordjournals.aje.a009981

Al Khalaf, M. M., Thalib, L., & Doi, S. A. R. (2011). Combining heterogenous studies using the random-effects model is a mistake and leads to inconclusive meta-analyses. *Journal of Clinical Epidemiology, 64*(2), 119–123. https://doi.org/10.1016/j.jclinepi.2010.01.009

Turner, R. M., Jackson, D., Wei, Y., Thompson, S. G., & Higgins, J. P. (2015). Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in Medicine, 34*(6), 984–998.

Van Aert, R. C., Van Assen, M. A., & Viechtbauer, W. (2019). Statistical properties of methods based on the Q-statistic for constructing a confidence interval for the between-study variance in meta-analysis. *Research Synthesis Methods, 10*(2), 225–239.

Van den Noortgate, W., & Onghena, P. (2003). A parametric bootstrap version of Hedges’ homogeneity test. *Journal of Modern Applied Statistical Methods, 2*(1), 73–79. https://doi.org/10.22237/jmasm/1051747620

van Erp, S., Verhagen, J., Grasman, R. 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), 4. https://doi.org/10.5334/jopd.33

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. https://doi.org/10.3102/10769986030003261

Viechtbauer, W. (2007). Hypothesis tests for population heterogeneity in meta-analysis. *The British Journal of Mathematical and Statistical Psychology, 60*(Pt 1), 29–60. https://doi.org/10.1348/000711007X4042

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Zuckerman, M. (1994). *Behavioral expressions and biosocial bases of sensation seeking*. Cambridge University Press.