Guidelines for Using the *Q* Test in Meta-Analysis

Yukiko Maeda
Purdue University

Michael R. Harwell
University of Minnesota

The Q test is regularly used in meta-analysis to examine variation in effect sizes. However, the assumptions of Q are unlikely to be satisfied in practice prompting methodological researchers to conduct computer simulation studies examining its statistical properties. Narrative summaries of this literature are available but a quantitative synthesis of study findings for using the Q test has not appeared. We quantitatively synthesized estimated Type I error rates and power values of a sample of computer simulation studies of the Q test. The results suggest that Q should not be used for standardized mean difference effect sizes like Hedges' g unless the number of studies and primary study sample sizes are at least 40. Use of the Fisher's r-to-z transformed effect size, on the other hand, resulted in Q performing well in almost all conditions studied. We summarize our findings in a table that provides guidelines for using this important test.

Meta-analysis is widely used in education and psychology to quantitatively synthesize study findings for a common topic. A survey of published articles in the journals *Psychological Bulletin, Journal of Applied Psychology*, and the *Review of Educational Research* from 2000 to 2014 showed that 349 of 2,505 articles (14%) utilized a meta-analytic technique to synthesize study results. In a meta-analysis, the *Q* test (Hedges, 1981, 1982a, 1982b) is typically used to test the homogeneity of effect sizes as well as the impact of moderators. The fact that 47.3% of the meta-analyses cited above utilized *Q* suggests that when a meta-analysis is done the *Q* test is frequently used.

To test the variability among *K* (k = 1, …, *K*) independent effect sizes (ES) using *Q* requires meta-analysts to choose a fixed effect or random effects model (we assume one ES per study but in practice multiple ESs are possible). Borenstein, Hedges, Higgins, and Rothstein (2009) argued the choice should be driven by whether we assume there is one true effect that every available ES is estimating in which case all ESs share a common population mean (fixed effect), or whether ESs are assumed to vary across studies implying they represent a random sample from a distribution of ESs (random effects). In practice the random effects model is frequently easier to justify (Borenstein et al., 2009).

To delineate the underlying model for ESs it's useful to employ a two-level hierarchical (random effects) model in which level 1 (within-study) has the form $\hat{\theta}_k = \theta_k + e_k, \;\; e_k \sim N(0, \sigma_k^2)$ where $\hat{\theta}_k$is an estimated ES (e.g., standardized mean difference) assumed to follow a normal distribution with mean $\theta_k$and variance σ_k^2, and e_k represents error. At level 2 (between-studies) the unconditional (no predictors) model is simply $\theta_k = \theta + u_k$, $u_k \sim N(0, \tau)$ where θ is the mean effect size (mean of the θ_k) and u_k is a level 2 random effect representing the difference

between the θ and the θ_k that is assumed to be normally-distributed with variance τ (Raudenbush & Bryk, 2002). If predictors representing study features like sample size or the year a study appeared are used then the level 2 model is conditional on those predictors. Substituting the level 2 expression into the level 1 expression produces the composite random effects unconditional model $\hat{\theta}_k = \theta_k + u_k + e_k$where $\hat{\theta}_k \sim N(\theta, \tau + \sigma_k^2)$. If $\tau = 0$, this is a fixed effect model.

To test the hypothesis of homogeneity of ESs, Ho: $\tau = 0$ (random effects) or, equivalently, Ho: All $\theta_k = \theta$ (fixed effect) (Raudenbush, 2009), the *Q* test (Hedges, 1982a, 1982b) is used:

$$Q = \sum_{k=1}^{K} \frac{\left(\hat{\theta}_k - \hat{\bar{\theta}}\right)^2}{\hat{\sigma}^2_{\hat{\theta}_k}}, \text{ where } \hat{\bar{\theta}} = \frac{\sum_{k=1}^{K} w_k \hat{\theta}_k}{\sum_{k=1}^{K} w_k}. \quad (1)$$

In equation (1) $\hat{\bar{\theta}}$ is an estimated grand mean, $\hat{\sigma}^2_{\hat{\theta}_k} = \hat{\tau} + \hat{\sigma}^2_k$ represents the variance of $\hat{\theta}_k$whose precise form depends on the nature of the ES, and w_k is a weight defined as $[\hat{\tau} + \hat{\sigma}^2_k]^{-1}$. The *Q* test is approximately distributed as a chi-square variable with degrees of freedom of *K*-1 assuming Ho is true. Retention of Ho: $\tau = 0$ implies that all ESs estimate a common mean, whereas rejection of Ho usually triggers additional analyses to identify moderators that account for variation among ESs meaning that the level 2 model now contains predictors. The *Q* test is also used to compute the percentage of total variation in ESs due to between-study variance that is potentially explainable (I^2) (Higgins & Thompson, 2002) and to estimate τ (Konstantopoulos & Hedges, 2004).

Although θ can represent several kinds of ESs for continuous outcomes, two frequently used indices are the standardized mean difference for two independent groups and the transformed (normalized) Pearson's correlation coefficient (McGrath & Meyer, 2006). The standardized mean difference and its variance are estimated as

$$\hat{\theta}_k = \frac{\bar{Y}_{1k} - \bar{Y}_{2k}}{S_{within}}, \; \hat{\sigma}^2_{\hat{\theta}_k} = \frac{n_1 + n_2}{n_1 n_2} + \frac{\hat{\theta}^2_k}{2(n_1 + n_2)}, \quad (2)$$

where $\bar{Y}_{1k}$ and $\bar{Y}_{2k}$ are sample means obtained from two independent groups for the *k*th study and S_{within} is a standardizer that traditionally has been a pooled within-groups standard deviation (Borenstein, 2009). Hedges (1981) modified $\hat{\theta}_k$to provide an (approximately) unbiased estimator for the standardized mean difference often called Hedges' *g*:

$$\hat{\theta}_{gk} = \hat{\theta}_k \left(1 - \frac{3}{n_1 + n_2 - 2}\right),$$

$$\hat{\sigma}^2_{\hat{\theta}_{gk}} = \hat{\tau} + \hat{\sigma}^2_k = \hat{\tau} + \left[\frac{n_1 + n_2}{n_1 n_2} + \frac{\hat{\theta}^2_k}{2(n_1 + n_2)}\right]\left[1 - \frac{3}{4(n_1 + n_2) - 1}\right]^2. \quad (3)$$

Pearson's correlations (r_k) also serve as ESs and are often transformed (normalized) using the Fisher's *r*-to-*z* transformation:

$$\hat{\theta}_{z_k} = \frac{1}{2} ln\left[\frac{1 + r_k}{1 - r_k}\right], \hat{\sigma}^2_{\hat{\theta}_{z_k}} = \hat{\tau} + \frac{1}{n_k - 3} \quad . (4)$$

(Borenstein, 2009). The use of $\hat{\theta}_{z_k}$ has been criticized on statistical grounds (e.g., Hunter & Schmidt, 2004) and some meta-analyses simply use r_k as an ES:

$$\hat{\theta}_{r_k} = r_k, \; \hat{\sigma}^2_{\hat{\theta}_{r_k}} = \left[\frac{\left(1 - r_k^2\right)^2}{n_k - 1}\right]. (5)$$

(Borenstein, 2009). The $\hat{\theta}$ values in equations (3), (4), and (5) and their variances are used in the *Q* test in equation (1).

Strictly speaking, valid inferences using the *Q* test are most likely when its underlying assumptions are satisfied. These assumption include: (a) ESs are independent, (b) e_k and $u_k$are normally-distributed, (c) observations within primary studies are independent, normally-distributed, and homoscedastic, and (d) sample sizes are large enough to justify the chi-square approximation (Hedges & Olkin, 1985). In practice, it is likely that one or more of the assumptions underlying *Q* will be violated (see Greenhouse & Iyengar, 2009; Micceri, 1989), raising questions about the impact of violations of assumptions on the statistical behavior of *Q*. In response, methodological researchers have used computer simulation studies to investigate the Type I error rate and statistical power of the *Q* test and to identify data conditions for which inferences based on *Q* are expected to be correct (e.g., Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006). A Type I error represents the probability of incorrectly rejecting Ho: $\tau = 0$ whereas power is the probability of correctly rejecting Ho: $\tau = 0$. These studies typically simulate data reflecting realistic data conditions like small numbers of studies and non-normal data and use these data to investigate the Type I error rate and power of *Q*. The goal is to identify data conditions in which *Q* controls its Type I error rate at a nominal value like $\alpha = 0.5$ and demonstrates adequate statistical power to detect heterogeneity among ESs (e.g., .85).

Review of the Computer Simulation Literature for *Q*

Homogeneous effect sizes: Type I error rate of the *Q* test. Several simulation studies have provided evidence that *Q* does not control its Type I error rate at α = .05 even for normally-distributed and homoscedastic data, and within-study sample sizes (N_k) of at least 40 and $K \geq$ 30 (e.g., Aaron, 2003; Harwell, 1997; Sánchez-Meca & Marín-Martínez, 1997); estimated Type I error rates ($\hat{\alpha}$) under these conditions sometimes differed noticeably from .05 (e.g., 02, .12). On the other hand, Hedges (1982b) found that *Q* showed good control of Type I error rates when data were normally-distributed and homoscedastic even for *K* = 5. Unequal within-study sample sizes and normally-distributed data tend to produce somewhat conservative $\hat{\alpha}$ rates (e.g., .03). Inflated $\hat{\alpha}$ rates (e.g., .07) have often appeared for *Q* when smaller (unequal) within-study variances were paired with larger (unequal) within-study sample sizes even for normally-distributed data (Cornwell, 1993; Harwell, 1997), a well-documented pattern in primary studies (Box, 1954). For non-normal and heteroscedastic primary study data $\hat{\alpha}$ is frequently conservative especially for smaller values of *K* but there is evidence that it moves toward .05 for $N_k \geq 40$ and $K \geq 40$. Sackett, Harris, and Orr (1986) and Cornwell (1993) also found that $\hat{\alpha}$ was insensitive to measurement error, but Sagie and Koslowsky (1993) reported that measurement error produced inflated $\hat{\alpha}$ values.

Heterogeneous effect sizes: The power of the *Q* test. As expected, the power of the *Q* test increases as heterogeneity of ESs across studies increases for a fixed value of *K* (Aaron, 2003; Hardy & Thompson, 1998; Harwell, 1997; Schulze, 2004). There is also agreement that in maximizing the power of the *Q* test larger values of *K* are typically more important than larger values of N_k (Schulze, 2004) regardless of the type of ES. However, larger values of *K* do not guarantee high power and in practice the power of *Q* depends on a complicated mix of factors including *K*, N_k, heterogeneity of ESs, noncentrality patterns of ESs, as well as the assumptions of normality and homoscedasticity within primary studies (Chang, 1993). The power of *Q* also

appears to decrease when the reliability of the variables decreases (Cornwell, 1993; Sackett et al., 1986).

Collectively, these simulation studies provide information about the behavior of *Q* under various data conditions and, to this point, their findings have been summarized qualitatively as part of a literature review in a computer simulation study or in characterizations of study findings. For example, Huedo-Medina et al. (2006) qualitatively summarized the results of several Monte Carlo studies of *Q* by commenting that "A shortcoming of the *Q* statistic is that it has poor power to detect true heterogeneity among studies when the meta-analysis includes a small number of studies and excessive power to detect negligible variability with a high number of studies" (p. 194). Similarly, Higgins, Thompson, Deeks, and Altman (2003) commented about *Q* that "The test is known to be poor at detecting true heterogeneity among studies as significant. Meta-analyses often include small numbers of studies, and the power of the test in such circumstances is low" (p. 557; see also Aaron, 2003; Harwell, 1997; Chang, 1993). In short, the primary point of agreement among qualitative summaries of simulation results seems to be that the power of *Q* is adversely affected by small numbers of studies; a consensus of the impact of assumption violations like non-normality and within-study heteroscedasticity has not been reached in qualitative summaries.

Qualitative summaries are helpful but can be problematic for at least three reasons. One is that the volume of findings can create significant challenges because a single computer simulation study can produce dozens and sometimes hundreds of findings, making a qualitative synthesis difficult for a sample of these studies. Second, the specific impact of conditions like numbers of studies (e.g., $K = 10$ versus 20) is usually difficult to quantitatively characterize in a qualitative review. Third, mixed and contradictory nature of simulation findings complicates attempts to qualitatively synthesize information accurately. Thus, the Sutton and Higgins' (2008) comment that "Simulation studies of the properties of tests for heterogeneity are plentiful, yet no overview of their findings seems to have been produced" (p. 628) still holds. Below we report a quantitative synthesis of a sample of computer simulation studies to provide guidelines for using *Q* for realistic data conditions that simultaneously address many of the difficulties of qualitative reviews and increases the likelihood of this test being appropriately used.

Meta-Analysis of Computer Simulation Results for the *Q* Test

Several researchers have used meta-analytic methods to summarize the results of computer simulation studies of a statistical test (e.g., Hoogland & Boomsma, 1998; Newman, Hall, & Fraase, 2003; Powell & Schafer, 2001). In this method, outcomes of simulation studies such as $\hat{\alpha}$ and 1- $\hat{\beta}$ (power) serve as ESs and study features such as number of studies as independent variables. This approach shares many of the advantages of meta-analysis over qualitative literature summaries, including the use of systematic methods to aggregate study findings, incorporating how sampling of individual studies was done into the modeling, and assessing the role of moderating variables (DeCoster, 2005). The result is that valid inferences about the impact of various conditions on the statistical behavior of the *Q* test can be made based on a quantitative summary of simulation results.

We used a meta-analytic method to summarize simulation results for the *Q* test to identify data conditions for which inferences based on *Q* are appropriate. Our goal was to quantitatively synthesize $\hat{\alpha}$ and $1-\hat{\beta}$ values for *Q* for a sample of computer simulation studies and to use our findings to provide guidelines for using the *Q* test. These guidelines should increase the likelihood that inferences based on *Q* will be correct.

Methodology

Sampling of Studies

The population consisted of computer simulation studies in which the Type I error and/or power of the *Q* test was investigated. Studies fitting this description were included in the meta-analysis regardless of the research domain in which they appeared. Scholarly works were identified using the *Educational Resources Information Center (ERIC)* database, *Dissertation Abstracts International*, reference lists of obtained studies, and *Google Scholar*. Examples of keywords used for identifying studies were: *Hedges, Q test, heterogeneity, homogeneity, independence, effect size(s), meta-analysis, simulation, Type I error, power,* and *Monte Carlo*. Reference lists of the identified studies were also examined. Through this process, we initially identified a sample of 21 studies of the *Q* test. However, a study by Hardy and Thompson (1998) was excluded because these authors did not report values for Type I error or power, and Viechtbauer (2007) was excluded because $\hat{\alpha}$ and $1-\hat{\beta}$ values for *Q* were reported using graphical displays and could not be accurately retrieved. Finally, a study by Takkochue, Cadarso-Suárez, and Spiegeiman (1999) was removed because it used odds ratios as ESs. As a result, eighteen studies met the inclusion criteria for quantitative synthesis and included a book chapter, 13 published journal articles, two unpublished conference or technical papers, and two dissertations. Effect sizes for the 18 studies included standardized mean differences and transformed and untransformed correlation coefficients. The sample of 18 studies produced 1,872 $\hat{\alpha}$ values and 3,087 $1-\hat{\beta}$ values for the *Q* test. These values can be treated as independent even if they were generated within the same simulation study because the simulated data values are independent based on conventional statistical methods to assess independence like serial correlation (Ripley, 1987). In 12 of the 18 studies, only normally-distributed data were simulated and only three studies modeled non-normal θ_k.

Coding of Studies

Coded study features included estimated Type I error rates ($\hat{\alpha}_k$) and power values ($1-\hat{\beta}_k$), the number of studies simulated (*K*), within-study sample size (N_k) for each simulated study, within-study variances (equal, unequal), distribution of the simulated primary study data or ES distribution where available, between-study parameter variance (τ) for power, measurement error, and range restrictions. In all studies α =.05.

Coding of studies was mainly done by the first author. The accuracy of the coded values was checked by coding all studies twice about three months apart and performing descriptive analyses of the coded values to identify unusual values. Few discrepancies emerged between the two rounds of coding. When a discrepancy was observed the coded value was compared with the information reported in the original simulation study. Any remaining ambiguity of coding was

resolved by consensus among the authors, helping to ensure that study features and ESs were accurately coded.

Data Analysis

We examined patterns in the ESs ($\hat{\alpha}_k$,$1-\hat{\beta}_k$) and their relation to study features using plots and descriptive statistics. The non-normal distributions of $\hat{\alpha}_k$ and $1-\hat{\beta}_k$ led us to transform these values using the arcsine transformation and to use the transformed values in our analyses. Proportions possess several statistical shortcomings including means and variances that are dependent and a distribution that is not necessarily well approximated by a normal distribution for small sample sizes. The arcsine transformation is a linearizing transformation to a scale in which the mean and variance of the transformed statistic are independent and approximately normally-distributed for the full range of transformed values even for small samples (Cox, 1970). The transformed and untransformed values (e.g., $\hat{\alpha}_k$, and $\widehat{\alpha'}_k$) share a monotonic relationship. In our analyses, the transformed values ($\widehat{\alpha'}_k$,$\left[1-\hat{\beta}_k\right]'$) served as outcome variables and were analyzed separately for each kind of ES modeled in the sample of Monte Carlo studies (i.e., $\hat{\theta}_{g_k}$, $\hat{\theta}_{z_k}$, $\hat{\theta}_{r_k}$). The transformed values ($\widehat{\alpha'}_k$,$\left[1-\hat{\beta}_k\right]'$) were transformed back to the original metric ($\hat{\alpha}_k$, $1-\hat{\beta}_k$) for reporting the results.

Variables used for the analysis were: (a) N_k based on four categories suggested by values coded for the studies (0 = Less than 30, 1 = 30 to less than 50, 2 = 50 to less than 100, and 3 = 100 or more), (b) *K* based on four categories suggested by values coded for the studies (0 = Less than 10, 1 = 10 to less than 30, 2 = 30 to less than 50, 3 = 50 or more), (c) normality of primary study distributions (1 = normally-distributed, 0 = Not normally-distributed) (d) heterogeneity of within-study variances for two groups (1 = yes, 0 = no), measurement error (1 = reliabilities of variables (X, Y) for computing $\hat{\theta}_k$ in primary studies is less than 1, 0 = reliabilities are 1), and range restriction (1 = yes, 0 = no). The range restriction was operationalized differently among studies. For example, Cornwell (1993) restricted the range of one variable used for computing $\hat{\theta}_k$, while Sagie and Koslowsky (1993) manipulated the ratio of sample to population SD. We also modeled between-study variance (τ) using quartiles suggested by the coded values (0 = Less than 0.01, 1 = 0.01 to less than 0.025, 2 = 0.025 to less than 0.067, and 3 = 0.067 and more). More specific coding of simulation factors and assumption violations was not possible because of a lack of variation. For example, most conditions in the sample of studies involved normally-distributed data, which limited the investigation of non-normal primary study data on $\widehat{\alpha'}_k$ and $\left(1-\hat{\beta}_k\right)'$.

Although we intended to code within-study sample sizes we found that most studies simulated unequal within-study sample sizes that varied across values of *K*. For example, Field (2001) used four levels of *N* ($\bar{N}$ = 20, 40, 80, or 160) for a given *K*, which presumably resulted in most within-study sample sizes being unequal, but it's impossible to know for certain because only average within-study sample sizes were reported. Also, none of the simulation studies simulated non-normal u_k for random effect models. The summary of simulation conditions and findings of each study is available from the first author upon request.

Results

The value of *K* modeled in the simulation studies ranged from 2 to 256 with 12 of the 18 studies simulating *K* being equal to or less than 50. There was also considerable variation in N_k that ranged from 7 to 5,000. Similarly, researchers employed a variety of τ values, which ranged from 0 (i.e., homogeneous effect sizes) to 1 with the median of 0.025. Perhaps the most striking finding in this literature is lack of investigation of the effect of combinations of realistic data conditions on the Type I error rate and power of *Q*. For example, as previously mentioned, three out of 18 studies used primary study data that are non-normal. Only one study (i.e., Cornwell, 1993) combined a non-normal primary study data distribution with a range restriction for small numbers of studies; and only a few Monte Carlo studies of *Q* (Cornwell, 1993; Sackett, Harris, & Orr, 1986; Sagie & Koslowsky, 1993) have examined the impact of range restrictions or measurement error which preliminary evidence suggests have a significant impact on the power of the *Q* test.

Table 1 reports average $\hat{\alpha}$ values for *Q* as a function of *K*, N_k, and whether underlying assumptions were satisfied versus at least one assumption was violated (i.e., non-normal distributions, within-study variance heterogeneity, measurement error, range restriction) for each type of ES. Under the condition of no assumption violation $\hat{\alpha}$ for $\hat{\theta}_{g_k}$tended to be conservative across all levels of *K* and N_k, ranging from 0.016 to 0.046. While *Q* maintained its Type I error rate at 0.05 for $\hat{\theta}_{z_k}$ with no assumption violations $\hat{\alpha}$ was always inflated unless N_k is larger than 100 when using untransformed Pearson's correlations $\hat{\theta}_{r_k}$. When at least one underlying assumption was violated $\hat{\alpha}$ for $\hat{\theta}_{g_k}$ was near α = .05 whereas for $\hat{\theta}_{Z_k}$the reported $\hat{\alpha}$ once again was near .05 even for small N_k and *K*, while $\hat{\theta}_{r_k}$ produced estimated error rates that tended to be inflated.

Table 1
Average estimated Type I error rates as a function of simulation factors across studies

ES type	Study feature		Violation: No J	M	95%CI LL_M	95%CI UL_M	Violation: Yes J	M	95%CI LL_M	95%CI UL_M
$\hat{\theta}_g$	K	Less than 10	91	0.039	0.039	0.040	114	0.054	0.053	0.054
		10 to less than 30	74	0.032	0.032	0.032	267	0.049	0.049	0.050
		30 to less than 50	25	0.042	0.041	0.043	267	0.053	0.052	0.053
		50 or more	15	0.016	0.016	0.017	--	--	--	--
$\hat{\theta}_Z$		Less than 10	52	0.050	0.050	0.052	14	0.050	0.047	0.054
		10 to less than 30	142	0.051	0.051	0.051	28	0.050	0.048	0.053
		30 to less than 50	40	0.051	0.050	0.051	--	--	--	--
		50 or more	47	0.049	0.048	0.050	--	--	--	--
$\hat{\theta}_r$		Less than 10	29	0.079	0.078	0.081	43	0.053	0.051	0.056
		10 to less than 30	52	0.059	0.058	0.060	22	0.059	0.056	0.062
		30 to less than 50	17	0.062	0.058	0.066	21	0.059	0.056	0.062
		50 or more	41	0.077	0.076	0.078	44	0.070	0.068	0.073
$\hat{\theta}_g$	N_k	Less than 30	67	0.024	0.024	0.024	277	0.046	0.045	0.046
		30 to less than 50	53	0.032	0.031	0.032	178	0.056	0.056	0.057
		50 to less than 100	36	0.041	0.040	0.041	8	0.047	0.044	0.051
		100 or more	49	0.046	0.045	0.046	185	0.056	0.056	0.057
$\hat{\theta}_Z$		Less than 30	37	0.052	0.051	0.052	42	0.050	0.048	0.052
		30 to less than 50	68	0.051	0.051	0.051	--	--	--	--
		50 to less than 100	73	0.051	0.050	0.051	--	--	--	--
		100 or more	103	0.050	0.050	0.050	--	--	--	--
$\hat{\theta}_r$		Less than 30	--	--	--	--	--	--	--	--
		30 to less than 50	35	0.083	0.082	0.084	4	0.087	0.052	0.130
		50 to less than 100	46	0.077	0.076	0.078	42	0.056	0.054	0.059
		100 or more	58	0.055	0.054	0.056	84	0.063	0.061	0.065

Note: $\hat{\theta}_g$= standardized mean difference, $\hat{\theta}_z$= Fisher r-to-z transformed correlation, $\hat{\theta}_r$= untransformed Pearson correlation, K = the number of effect sizes for a study, J = the number of effect sizes for the meta-analysis; M = weighted mean, LL_M = 95% lower limit for M, and UL_M = 95% upper limit for M. Violation = yes includes non-normal distributions, within-study variance heterogeneity, range restriction, or measurement error. -- means results could not be produced because of no data or lack of variation.

Table 2 reports power for Q as a function of K, N_k, and between-study variance τ by type of ES. We note that interpreting power when the corresponding Type I error rate is inflated or conservative must be done cautiously. The power of Q for $\hat{\theta}_{g_k}$ ranged from 0.220 to 0.819 for no assumption violations and generally increased as K, N_k, or τ increased except for N_k between 30 and 50 or τ between 0.01 and 0.025. The power associated with $\hat{\theta}_{Z_k}$ was quite large when no violation existed except τ less than 0.01, ranging from 0.453 to 0.998. The power of Q associated with $\hat{\theta}_{r_k}$ for the no violation case ranged from 0.189 to .870 and increased with increases in K, N_k, or τ but didn't exceed .70 unless K is larger than 50 and N_k is larger than 100. For the

assumption violated conditions $\hat{\theta}_{z_k}$ appeared to produce a *Q* test that was less sensitive than its competitors, although this inference is based on a relatively modest amount of data.

Table 2
Average estimated power values as a function of simulation factors across studies

			Violation							
			No				Yes			
ES					*95%CI*				*95%CI*	
type	Study feature		*J*	*M*	LL_M	UL_M	*J*	*M*	LL_M	UL_M
$\hat{\theta}_g$	K	Less than 10	145	0.345	0.344	0.346	112	0.362	0.361	0.363
		10 to less than 30	166	0.540	0.539	0.540	120	0.445	0.444	0.447
		30 to less than 50	124	0.647	0.645	0.648	123	0.562	0.561	0.564
		50 or more	52	0.819	0.819	0.820	7	0.703	0.698	0.708
$\hat{\theta}_Z$		Less than 10	130	0.910	0.910	0.911	14	0.723	0.716	0.730
		10 to less than 30	264	0.998	0.998	0.998	28	0.910	0.907	0.910
		30 to less than 50	115	0.998	0.998	0.999	--	--	--	--
		50 or more	115	0.911	0.91	0.912	--	--	--	--
$\hat{\theta}_r$		Less than 10	58	0.447	0.445	0.450	127	0.350	0.348	0.353
		10 to less than 30	98	0.526	0.522	0.529	64	0.537	0.533	0.541
		30 to less than 50	46	0.733	0.729	0.737	63	0.673	0.669	0.676
		50 or more	43	0.870	0.866	0.874	128	0.841	0.839	0.843
$\hat{\theta}_g$	N_k	Less than 30	164	0.717	0.716	0.718	184	0.216	0.215	0.217
		30 to less than 50	40	0.374	0.373	0.374	59	0.517	0.516	0.519
		50 to less than 100	126	0.558	0.557	0.558	28	0.397	0.393	0.401
		100 or more	157	0.741	0.74	0.743	91	0.933	0.932	0.933
$\hat{\theta}_Z$		Less than 30	109	0.960	0.960	0.960	42	0.857	0.854	0.860
		30 to less than 50	115	0.993	0.993	0.993	--	--	--	--
		50 to less than 100	125	0.997	0.997	0.997	--	--	--	--
		100 or more	275	0.998	0.998	0.998	--	--	--	--
$\hat{\theta}_r$		Less than 30	3	0.473	0.469	0.477	--	--	--	--
		30 to less than 50	46	0.314	0.310	0.318	4	0.805	0.748	0.857
		50 to less than 100	74	0.559	0.555	0.562	126	0.379	0.376	0.381
		100 or more	122	0.754	0.751	0.757	252	0.718	0.717	0.720
$\hat{\theta}_g$	τ	Less than 0.01	110	0.408	0.406	0.409	21	0.103	0.101	0.105
		0.01 to less than 0.025	60	0.220	0.269	0.270	12	0.468	0.462	0.474
		0.025 to less than 0.067	126	0.505	0.504	0.506	54	0.384	0.382	0.387
		0.067 or more	191	0.746	0.746	0.747	275	0.496	0.495	0.497
$\hat{\theta}_Z$		Less than 0.01	171	0.453	0.452	0.455	--	--	--	--
		0.01 to less than 0.025	191	0.763	0.762	0.764	24	0.731	0.725	0.736
		0.025 to less than 0.067	223	0.996	0.996	0.996	18	0.969	0.966	0.971
		0.067 or more	39	0.982	0.981	0.982	--	--	--	--
$\hat{\theta}_r$		Less than 0.01	89	0.189	0.187	0.192	126	0.224	0.221	0.226
		0.01 to less than 0.025	124	0.686	0.684	0.689	252	0.791	0.790	0.793
		0.025 to less than 0.067	7	0.669	0.663	0.675	--	--	--	--
		0.067 or more	25	0.870	0.866	0.874	--	--	--	--

Note: $\hat{\theta}_g$ = standardized mean difference, $\hat{\theta}_z$ = Fisher's transformed correlation, $\hat{\theta}_r$= untransformed Pearson correlation, *K* = the number of effect sizes for a study, *J* = the number of effect sizes for the meta-analysis, *M* = weighted mean, LL_M = 95% lower limit for *M*, and UL_M = 95% upper limit for *M*. -- means results could not be produced because of no data or lack of variation.

We further explored patterns in $\hat{\alpha}$ and $1-\hat{\beta}$ by specific assumption violations for each type of ES and report these descriptive results in Table 3. Because of a relative paucity of simulation work with conditions where data assumptions were violated, a consistent pattern of the impact of assumption violations on power was hard to discern. However, $\hat{\theta}_{Z_k}$again outperformed $\hat{\theta}_{g_k}$ for normal and non-normal distributions for Type I error and power, and offered better control of Type I error rates than $\hat{\theta}_{r_k}$ in the presence of measurement error or a range restriction. Heterogeneity of variance was also associated with a substantial decrease in power for $\hat{\theta}_{g_k}$. The presence of a range restriction produced an inflated $\hat{\alpha}$ for $\hat{\theta}_{r_k}$ but not $\hat{\theta}_{Z_k}$, and decreased the power of *Q* for the latter.

Table 3
Average estimated Type I error rates and power values as a function of specific assumption violations in the simulation

			Type I				Power			
					95% CI				95% CI	
Violation		*ES*	*J*	*M*	LL_M	UL_M	*J*	*M*	LL_M	UL_M
Distribution	Normal	$\hat{\theta}_g$	236	0.057	0.056	0.057	243	0.623	0.622	0.624
		$\hat{\theta}_z$	110	0.049	0.049	0.050	153	0.759	0.758	0.760
	Non-normal	$\hat{\theta}_g$	468	0.050	0.050	0.051	218	0.534	0.533	0.535
		$\hat{\theta}_z$	18	0.050	0.047	0.053	18	0.813	0.808	0.819
Within-study variances	Equal	$\hat{\theta}_g$	199	0.049	0.048	0.049	215	0.523	0.522	0.524
	Unequal	$\hat{\theta}_g$	504	0.053	0.053	0.053	176	0.338	0.337	0.339
RR	No	$\hat{\theta}_z$	299	0.051	0.051	0.051	642	0.990	0.990	0.990
		$\hat{\theta}_r$	265	0.069	0.068	0.070	623	0.589	0.588	0.591
	Yes	$\hat{\theta}_z$	24	0.049	0.046	0.051	24	0.731	0.725	0.736
		$\hat{\theta}_r$	4	0.087	0.052	0.130	4	0.805	0.748	0.857
ME	No	$\hat{\theta}_z$	300	0.051	0.051	0.051	642	0.990	0.990	0.990
		$\hat{\theta}_r$	139	0.071	0.070	0.071	245	0.564	0.562	0.566
	Yes	$\hat{\theta}_z$	23	0.051	0.048	0.054	24	0.728	0.722	0.734
		$\hat{\theta}_r$	130	0.061	0.059	0.062	382	0.609	0.607	0.611

Note: RR = range restriction, ME = measurement error, *K* = the number of cases for a study, $\hat{\theta}_g$= standardized mean difference, $\hat{\theta}_z$ = Fisher's transformed correlation, $\hat{\theta}_r$ = untransformed Pearson correlation, *J* = the number of effect sizes for the meta-analysis *M* = weighted mean, LL_M = 95% lower limit for *M*, and UL_M = 95% upper limit for *M*.

Discussion

Testing the homogeneity of effect sizes with Hedges' (1982b) *Q* test remains a common practice in meta-analysis. In using *Q* it is important to identify the data conditions for which this test is appropriately applied, i.e. controls its Type I error rates at nominal values and produces adequate power to detect heterogeneity among effect sizes. This requires understanding the impact of violating underlying assumptions such as normality, homoscedasticity, and no measurement error or range restriction, as well as the conditions under which the large sample approximation of *Q* is justified.

Computer simulation studies have been used to identify which (if any) assumption violations affect the Type I error rate and power of *Q*, as well as conditions for which the large sample approximation is justified. As Sutton and Higgins (2008) noted a summary of this literature is needed to help ensure that *Q* is used when inferences based on this test are valid. The simulation literature for the *Q* test contains qualitative summaries but not a quantitative summary that can provide guidelines to meta-analysts for appropriately using *Q*. Our study adds to the meta-analytic literature of *Q* by quantitatively synthesizing a sample of simulation studies of this test. This approach draws on the many strengths of this methodology and our findings (summarized in Table 4) provide guidelines for using the *Q* test

Two key findings emerged. First, the *Q* test based on Fisher's *r*-to-*z* transformation ($\hat{\theta}_z$) overall showed excellent control of Type I error rates and strong power for the conditions modeled that were clearly superior to those linked to standardized mean differences ($\hat{\theta}_g$) and untransformed correlation effect sizes ($\hat{\theta}_r$). For example, Table 4 shows that the Type I error rate of *Q* stayed near the nominal level for $\hat{\theta}_z$ regardless of whether assumptions were satisfied even for small numbers of studies ($K < 10$) and small study sample sizes ($N_k < 10$), and was insensitive to non-normal data. *Q* also generally showed excellent power ($1 - \beta = .91$) for normally-distributed and homoscedastic data for small numbers of studies ($K < 10$) and small study sample sizes ($N_k < 30$) for $\hat{\theta}_z$.

There were instances in which $\hat{\theta}_g$ was associated with good control of Type I error rates and good power for *Q*, but there simply were not as many as for $\hat{\theta}_z$. In general, the most negative effects on *Q* with $\hat{\theta}_g$ appeared for smaller numbers of studies and smaller study sample sizes. For example, with distributional assumptions (i.e., normality, homoscedasticity) satisfied and the number of studies less than 10 the Type I error rate of *Q* was .039 ($\alpha = .05$) and power was .345, both of which fell short of those linked to $\hat{\theta}_z$ for the same conditions (.050 and .910). Unequal within-study variances had little effect on the Type I error rate for $\hat{\theta}_g$but were associated with less power. Of course, for power it's possible that studies using $\hat{\theta}_g$consistently modeled less effect size variability than those that studied $\hat{\theta}_z$ which, other things being equal, would produce smaller power values for the former. It is difficult to ignore the almost uniformly superior performance of *Q* for $\hat{\theta}_z$. The *Q* test based on $\hat{\theta}_z$ can also be used with fewer 10 studies and primary study sample sizes of 30 or less and still show excellent power for detecting heterogeneity unless the data show range restriction.

Our results also suggest that untransformed correlations ($\hat{\theta}_r$) not be used with *Q* because of frequent inflation of Type I error rates for a range of conditions. This finding adds to the literature of the ongoing debate on $\hat{\theta}_r$versus $\hat{\theta}_z$ (Field, 2001; Hafdahl, 2010).

A second key finding present in Table 4 is the sparseness of the simulation literature for *Q*. While the sample of Type I error and power values generated in simulation studies is large the effect of combinations of conditions on *Q* is still unclear. For example, there has been little work done of the impact of primary study data that are non-normal and heteroscedastic and show measurement error. Still, the results in Table 4 provide guidelines for correctly using *Q* and thus increase the likelihood of valid inferences based on this test.

Table 4
Guidelines for Using the Q Test ($\alpha = .05$)

	Effect on α			Effect on $1-\beta$		
Factor	$\hat{\theta}_g$	$\hat{\theta}_z$	$\hat{\theta}_r$	$\hat{\theta}_g$	$\hat{\theta}_z$	$\hat{\theta}_r$
• Outcomes within primary studies: normally-distributed, equal variances[1] • ESs: normal	Somewhat conservative ($\hat{\alpha} = .04$) even for $K \geq 50$ and $N_k \geq 100$	Minimal even for $K < 10$ and $N_k < 10$	Inflated ($\hat{\alpha} = .079$) unless $N_k > 100$	Good power for $K \geq 50$ (e.g., $1-\beta = .82$)	Excellent power ($1-\beta = .91$) even for $K < 10$ and $N_k < 30$	Moderate power ($1-\beta = .73$) unless $K \geq 50$
• Outcomes within primary studies: normally-distributed, unequal variances[2] • ESs: normal	Minimal (role of K and N_k for unequal variances unclear)	Limited or no data	Limited or no data	Low power (< .34)	No data	No data
• Outcomes within primary studies: non-normal, equal variances • ESs: normal	Minimal even for $K < 10$ and $N_k < 10$ (small number of non-normal primary study distributions studied)	Minimal even for $K < 10$ and $N_k < 10$ (small number of non-normal primary study distributions studied)	Limited or no data	Minimal (small number of non-normal primary study distributions studied)	Good power ($1-\beta = .81$) (small number of non-normal primary study distributions studied)	No data
• Outcomes within primary studies: non-normal, un equal variances • ESs: normal	Limited or no data	Limited or no data	Limited or no data	Limited or no data	No data	No data

Table 4 (continued)
Guidelines for Using the Q Test (α = .05)

	Effect on α			Effect on $1-\beta$		
Factor	$\hat{\theta}_g$	$\hat{\theta}_z$	$\hat{\theta}_r$	$\hat{\theta}_g$	$\hat{\theta}_z$	$\hat{\theta}_r$
• Observations in primary studies: normal, equal variances • ESs: non-normal	No data	No data	No data	No data	No data	No data
• Observations in primary studies: normal, unequal variances • ESs: non-normal	No data	No data	No data	No data	No data	No data
Measurement Error	Limited or no data	Minimal effect	Some inflation ($\hat{\alpha}$ = > .06) even when normality satisfied and K > 50 and N > 100	No data	Moderate power ($1-\beta$ = .61)	Moderate power ($1-\beta$ = .61)
Range Restriction	Limited or no data	Some effect	Inflating effect	No data	Moderate power ($1-\beta$ = .73) but role of K and N_k unclear	Good power ($1-\beta$ = .80) but may be partly due to inflated Type I error rate

Note.1.for $\hat{\theta}_z$and $\hat{\theta}_r$ this implies a bivariate normal distribution with equal variances; 2. for $\hat{\theta}_z$and $\hat{\theta}_r$this implies a bivariate normal distribution with unequal variances); $\hat{\theta}_g$=standardized effect size, $\hat{\theta}_z$= Fisher's r-to-z transformed correlation, $\hat{\theta}_r$= untransformed correlation. Limited or no data means that there were few (e.g., 3) or no cases.

Recommendations for Practice

Of course, our findings are limited by several factors including non-random sampling of studies in our meta-analysis, and coding choices of possible moderator variables. Still, we think our results suggest three recommendations for meta-analysts planning to use the *Q* test. First, the Fisher *r*-to-*z* transformation is preferred as an effect size. This effect size was associated with the best control of Type I error rates of *Q* and the greatest power for the largest number of conditions. The use of untransformed correlations as effect sizes is not recommended as these were consistently associated with inflated Type I error rates for *Q*.

Second, the *Q* test based on *r*-to-*z* effect sizes can generally be used with as few as 10 studies and primary study sample sizes of 30 or less and still show excellent power for detecting heterogeneity unless the data show a range restriction or measurement error. Third, meta-analysts need to pay careful attention to the possibility of range restriction and measurement error in primary study data, something that published meta-analyses do not always do. Coding variables that capture these characteristics in primary studies and including these variables in the meta-analysis is recommended.

These recommendations do not offer detailed guidance for exactly when the *Q* test can be used with confidence and when it should not be used, which is not surprising given the somewhat sparse state of the Monte Carlo literature for *Q*. However, the recommendations inform the routine use of this meta-analytic test in ways that should enhance the validity of test-based inferences. Moreover, our findings suggest that other meta-analytic tests or statistics such as I^2 (Huedo-Medina et al., 2006) for which a literature of computer simulation findings exists can usefully be summarized using a quantitative synthesis.

Author Notes

Yukiko Maeda is a Professor in the Department of Educational Studies in the College of Education at Purdue University.

Michael R. Harwell is a Professor in the Department of Educational Psychology in the College of Education and Human Development at the University of Minnesota.

Correspondence concerning this article should be addressed to Yukiko Maeda at ymaeda@purdue.edu.

References

Reference marked with an asterisk (*) indicates studies included in the meta-analysis.

*Aaron, L. T. (2003). *A comparative simulation of Type I error and power of four tests of homogeneity of effects for random- and fixed-effects models of meta-analysis*. (Doctoral dissertation. University of South Florida, Tampa). Retrieved from http://scholarcommons.usf.edu/etd/1319/

*Alexander, R. A., Scozzaro, M. J., & Borodkin, L. J. (1989). Statistical and empirical examination of the chi-square test for homogeneity of correlations in meta-analysis. *Psychological Bulletin, 106*(2), 329–331.

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* (2nd ed.; pp. 221-235). New York, NY: Sage.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis.* New York, NY: Wiley.

Box, G. E. P. (1954). Some theorems on quadratic forms applied in the study of analysis of variance problems I: Effects of inequality of variance in the one-way classification. *Annals of Mathematical Statistics, 25*(2), 290-302.

*Chang, L. (1993). *A power analysis of the test of homogeneity in effect-size meta-analysis.* (Unpublished Doctoral dissertation). Michigan State University, East Lansing.

*Cornwell, J. M. (1993). Monte Carlo comparisons of three tests for homogeneity of independent correlations. *Educational and Psychological Measurement, 53*(3), 605-618. doi: 10.1177/0013164493053003003

Cox, D. (1970). *The analysis of binary data.* London, U.K.: Methuen.

DeCoster, J. (2005). Meta-analysis. In K. Kempf-Leonard (Ed.), *The Encyclopedia of Social Measurement*, 683-688. San Diego, CA: Academic Press.

*Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6*(2), 161-180. doi: 10.1037/1082-989X.6.2.161

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* (2nd ed.; pp. 417-434). New York, NY: Russell Sage Foundation.

Hafdahl, A. R. (2010). Random-effects meta-analysis of correlations: Monte Carlo evaluation of mean estimators. *British Journal of Mathematical and Statistical Psychology*, *63*(1), 227-254. doi: 10.1348/000711009X431914

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, *17*(8), 841 – 856. doi: 10.1002/(SICI)1097-0258(19980430)17:8<841::AID-SIM781>3.0.CO;2-D

Harwell, M. (1997). An empirical study of Hedge's homogeneity test. *Psychological Methods, 2*(2), 219-231.

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. (1982a). Fitting categorical models to effect size from a series of experiments. *Journal of Educational Statistics, 7*(2), 119–137. doi: 10.3102/10769986007002119

*Hedges, L. V. (1982b). Estimation of effect size from a series of independent experiments. *Psychological Bulletin, 92*(2), 490–499. doi: 10.1037/0033-2909.92.2.490

Hedges, L. V., & Olkin, O. (1985). *Statistical methods for meta-analysis*. Orlando, FL: Academic Press.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: *Q* statistic or I^2 index?. *Psychological Methods, 11*(2), 193-206. doi: 10.1037/1082-989X.11.2.193

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying the 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

Hoogland, J. J., & Boomsa, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods & Research, 26*(3), 329-367. doi: 10.1177/0049124198026003003

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings* (2nd ed.). Thousand Oaks, CA: Sage.

*Kim, J. P. (2000, April). *An empirical study of the effect of pooling effect sizes on Hedges' homogeneity test*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Konstantopoulos, S., & Hedges, L. V. (2004). Meta-analysis. In D. Kaplan (Ed.), *Handbook of quantitative methodology for the social sciences* (pp. 281-297). New York, NY: Sage.

*Koslowsky, M., & Sagie, A. (1993). On the efficacy of credibility intervals as indicators of moderator effects in meta-analytic research. *Journal of Organizational Behavior*, *14*(7), 695-699. doi: 10.1002/job.4030140708

*Kulinskaya, E., Bollinger, M. B., & Bjørkestøl, K. (2009). *Testing for homogeneity in meta-analysis I. The one parameter case: Standardized mean difference*. Retrieved from http://arxiv.org/abs/0906.2999

Marascuilo, L. A., & McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Monterey, CA: Brooks/Cole.

McGrath, R. E., & Meyer, G. J. (2006). When effect sizes disagree: The case of *r* and *d*. *Psychological Methods, 11*(4), 386-401. doi: 10.1037/1082-989X.11.4.386

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychological Bulletin, 105*(1), 156-166. doi: 10.1037/0033-2909.105.1.156

*Morris, S. B. (2000). Distribution of the standard mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, *53*(1), 17-29.

Newman, I., Hall, R. J., & Fraas, J. (April, 2003). *Development of a regression model for estimating the effects of assumption violations on Type I error rates in the student's t test: Implications for practitioners.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Powell, D. A., & Schafer, W. D. (2001). The robustness of the likelihood ratio chi-square test for structural equation models: A meta-analysis. *Journal of Educational and Behavioral Statistics, 26*(1), 105-132. doi: 10.3102/10769986026001105

*Rasmussen, J. L., & Loher, B. T. (1988). Appropriate critical percentages for the Schmidt and Hunter meta-analysis procedure: Comparative evaluation of Type I error rate and power. *Journal of Applied Psychology*, *73*(4), 683-687. doi: 10.1037/0021-9010.73.4.683

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* (2nd ed.; pp. 295-316). New York, NY: Sage.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Ripley, B. D. (1987). *Stochastic simulation*. New York, NY: Wiley.

*Sackett, P. R., Harris, M. M., & Orr, J. M. (1986). On seeking moderator variables in the meta-analysis of correlation data: A Monte Carlo investigation of statistical power and resistance to Type I error. *Journal of Applied Psychology, 71*(2), 302 – 310. doi: 10.1037/0021-9010.71.2.302

*Sagie, A., & Koslowsky, M. (1993). Detecting moderators with met-analysis: An evaluation and comparison of techniques. *Personnel Psychology, 46*(3), 629 – 640. doi: 10.1111/j.1744-6570.1993.tb00888.x

*Sánchez-Meca, J., & Marín-Martínez, F. (1997). Homogeneity tests in meta-analysis: A Monte Carlo comparison of statistical power and Type I error. *Quality & Quantity, 31*(4), 385-399. doi: 10.1023/A:1004298118485

*Schulze, R. (2004). *Meta-analysis: A comparison of approaches*. Cambridge, MA: Hogrefe & Huber.

*Spector & P. E., & Levine, E. L. (1987). Meta-analysis for integrating study outcomes: A Monte Carlo study of its susceptibility to Type I and Type II errors. *Journal of Applied Psychology, 72*(1), 3-9. doi: 10.1037/0021-9010.72.1.3

Sutton, A. J., & Higgins, J. P. T. (2008). Recent developments in meta-analysis. *Statistics in Medicine, 27*(5), 625-650. doi: 10.1002/sim.2934

Takkochue, B., Cadarso-Suárez, C., & Spiegelman, D. (1999). Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis. *American Journal of Epidemiology, 150*(2), 206-215.

Viechtbauer, W. (2007). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology, 60*(1), 29-60. doi: 10.1348/000711005X64042