British Journal of Mathematical and Statistical Psychology (2007), 60, 29–60
© 2007 The British Psychological Society

The British Psychological Society

www.bpsjournals.co.uk

Hypothesis tests for population heterogeneity in meta-analysis

Wolfgang Viechtbauer*

University of Illinois at Urbana-Champaign, USA and University of Maastricht, The Netherlands

Choice of the appropriate model in meta-analysis is often treated as an empirical question which is answered by examining the amount of variability in the effect sizes. When all of the observed variability in the effect sizes can be accounted for based on sampling error alone, a set of effect sizes is said to be homogeneous and a fixed-effects model is typically adopted. Whether a set of effect sizes is homogeneous or not is usually tested with the so-called *Q* test. In this paper, a variety of alternative homogeneity tests – the likelihood ratio, Wald and score tests – are compared with the *Q* test in terms of their Type I error rate and power for four different effect size measures. Monte Carlo simulations show that the *Q* test kept the tightest control of the Type I error rate, although the results emphasize the importance of large sample sizes within the set of studies. The results also suggest under what conditions the power of the tests can be considered adequate.

1. Introduction

A fundamental issue in meta-analysis is the selection of the proper model underlying a set of effect sizes. Among the most common meta-analytic models adopted are the fixed-effects and the random-effects models (Hardy & Thompson, 1998; Hedges & Vevea, 1998; Hunter & Schmidt, 2000; Normand, 1999). These two models not only require a completely different conceptualization of the effect size estimates, but also can lead to very different conclusions about the presence of so-called *moderator variables*, that is, independent variables that influence the magnitude of the effect sizes.

Choice of the appropriate model is often treated as an empirical question which is answered by examining the amount of variability in the effect sizes. Each effect size estimate is subject to sampling error, and we can typically obtain an unbiased estimate of the amount of variance in a single effect size based on sampling error. This in turn allows us to estimate the *population heterogeneity*, meaning the amount of variability in the

* *Correspondence should be addressed to Wolfgang Viechtbauer, Department of Methodology and Statistics, University of Maastricht, P.O. Box 616, 6200 MD Maastricht, The Netherlands (e-mail: wolfgang.viechtbauer@stat.unimaas.nl).*

DOI:10.1348/000711005X64042

observed effect sizes over and beyond that which we would expect based on sampling error alone (Friedman, 2000; Thompson & Sharp, 1999; Viechtbauer, 2005). If the amount of population heterogeneity is sufficiently large, then we can no longer assume that the observed effect sizes are estimates of one and the same population effect. In other words, the population effect sizes cannot be assumed to be *homogeneous*. In that case, we can hypothesize that (a) an additional source of random variability besides sampling error is influencing the effect sizes, (b) moderator variables are introducing additional variability into the effect sizes, or (c) these two processes are operating in combination (Lipsey & Wilson, 2001).

Whether a set of effect sizes is homogeneous or not can be tested with a variety of homogeneity tests. The purpose of the present paper is to contrast such homogeneity tests in terms of their Type I error rate and power. The so-called Q test is the most commonly used test for examining the hypothesis of population homogeneity. However, it has been criticized for not having enough power to detect heterogeneity when sample sizes are small (Hunter & Schmidt, 2000; Sánchez-Meca & Marín-Martínez, 1997). Some alternatives to the Q test are the likelihood ratio, Wald test and score test. However, these other hypothesis tests have received little or no attention in the context of meta-analysis and it is generally unknown whether they are viable alternatives to the Q test. Moreover, the various tests can yield conflicting results and it is therefore of interest to determine whether one test should be preferred over the others.

The alternative tests can be based on either maximum likelihood (ML) or restricted maximum likelihood (REML) estimation, and both approaches were explored. Since an iterative algorithm must be employed to obtain the ML and REML estimates, the Type I error rate and power of the tests were compared numerically in a set of Monte Carlo simulations involving four different effect size measures: the unstandardized mean difference (UMD), the standardized mean difference (SMD), the correlation coefficient, and the correlation coefficient after applying Fisher's variance-stabilizing transformation.

The outline of this article is as follows. In Section 2, I will briefly define the fixed- and random-effects models in meta-analysis and show that these two models are distinguished by a single variance component that is equal to zero for the fixed-effects model and greater than zero for the random-effects model. Therefore, testing the homogeneity of effect sizes is a matter of testing whether this variance component is zero or not. I will outline in Section 3 the essential characteristics of the general linear mixed-effects model (GLMM) and then show in Section 4 that the meta-analytic fixed- and random-effects models are just special cases of the GLMM. Parameter estimation via ML and REML is discussed in this section as well. In Section 5, I introduce the various homogeneity tests and discuss some of their properties. Two examples are provided in Section 6 that illustrate the use of the various tests and demonstrate that the agreement between the tests is less than perfect. The results from the Monte Carlo simulations are provided in Section 7. A few general comments then conclude the paper.

2. Meta-analytic models

2.1. Fixed-effects model

Before discussing the homogeneity tests, it is useful to explicitly outline the models under consideration. Suppose we derive from a set of k studies k effect size estimates, $ES_1, \ldots, ES_k$, which describe the relationship between two variables that are of interest. The two variables might be measured on a continuous scale, which would

suggest the use of the correlation coefficient in its raw or variance-stabilized version as a natural effect size index (Rosenthal, 1994). Alternatively, if one variable indicates group membership and the other is some continuous outcome measure on which the groups are being compared, then the SMD is commonly chosen as the measure of effect size (Rosenthal, 1994). Finally, effect size measures for dichotomous dependent variables, often encountered in the medical field, include the risk difference, risk ratio and odds ratio (Fleiss, 1994). Regardless of the specific effect size index used for the k studies, ES_i describes the observed strength of the relationship between the two variables in the ith study. For the remainder of this paper, I will assume that each of the k studies provides a single independent effect size estimate.

The most basic meta-analytic model is the fixed-effects model given by

$$ES_i = \theta + \varepsilon_i, \tag{1}$$

where the effect size estimate of study i is decomposed into θ, the fixed population effect size for all k studies, and ε_i, a random error term by which estimate ES_i deviates from the true population effect size. The sampling error ε_i of the ith study is assumed to be normally distributed with mean 0 and variance $\sigma^2_{\varepsilon_i}$, from which it follows that

$$ES_i \sim N(\theta, \sigma^2_{\varepsilon_i}). \tag{2}$$

Note that, in contrast to the typical assumption of homoscedasticity in linear regression and analysis of variance models, we do not assume that the sampling variances are homogeneous for all k effect size estimates. The sampling variance of an effect size measure depends inversely on the sample size of the study from which it was derived (the within-study sample size) and possibly some other parameters. Therefore, holding everything else constant, larger within-study sample sizes yield smaller values of $\sigma^2_{\varepsilon_i}$ and consequently provide more precise estimates of θ. Because sample sizes are typically not homogeneous across studies, heteroscedasticity is to be expected in meta-analysis.

2.2. *Random-effects model*

Now suppose that the population effect sizes are heterogeneous for the set of studies as a result of random variability in the population effects. A common way to model the heterogeneity in the population effect sizes is to conceptualize the decomposition of the effect size estimates as a two-stage hierarchical process. First, the population effect size for the ith study is given by

$$\theta_i = \mu_\theta + \tau_i, \tag{3}$$

where τ_i is the amount by which θ_i differs from the average population effect μ_θ. In the second stage, the observed effect size is decomposed into the study-specific population effect and a sampling error component by which ES_i differs from θ_i:

$$ES_i = \theta_i + \varepsilon_i. \tag{4}$$

Combining these two equations yields the random-effects model, given by

$$ES_i = \mu_\theta + \tau_i + \varepsilon_i. \tag{5}$$

The typical assumptions for this model are as follows: (a) $\varepsilon_i \sim N(0, \sigma^2_{\varepsilon_i})$, (b) $\tau_i \overset{\text{iid}}{\sim} N(0, \sigma^2_\theta)$, (c) $\text{Cov}[\varepsilon_i, \varepsilon_{i'}] = 0$ for all $i \neq i'$, and (d) $\text{Cov}[\tau_i, \varepsilon_{i'}] = 0$ for all i and i'. Therefore, we assume independent and normally distributed sampling errors with heteroscedastic sampling variances and normally distributed variability among the population effect sizes that is independent of the sampling errors.

One way to conceptualize the population effect sizes is to think of them as realizations of a random variable with expectation μ_θ and variance σ^2_θ. Since ES_i is a linear combination of independent normally distributed random variables, it follows that

$$ES_i \sim N(\mu_\theta, \sigma^2_\theta + \sigma^2_{\varepsilon_i}). \tag{6}$$

In the random-effects model, the observed effect size from the ith study is an estimate of the average effect size in the population, as opposed to the fixed-effects model, where ES_i estimates a fixed population effect.

Note that the fixed-effects model is just a special case of the random-effects model where $\theta_i = \theta$ for $i = 1, \ldots, k$, or equivalently, $\sigma^2_\theta = 0$. In other words, in the absence of population heterogeneity, the random-effects model reduces to the fixed-effects model. Therefore, choice of the appropriate model is a question of whether $\sigma^2_\theta = 0$ or $\sigma^2_\theta > 0$.

3. General linear mixed-effects model

The fixed- and random-effects models discussed above are actually just special cases of the general linear mixed-effects model. GLMMs are of the general form

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z\gamma} + \boldsymbol{e},$$

where $\boldsymbol{y}$ is a $(k \times 1)$ vector of random variables, $\boldsymbol{X}$ is a $(k \times p)$ design matrix of known constants for the $(p \times 1)$ fixed-effects parameter vector $\boldsymbol{\beta}$, $\boldsymbol{Z}$ is the $(k \times q)$ design matrix for the $(q \times 1)$ random-effects vector $\boldsymbol{\gamma}$, and $\boldsymbol{e}$ is a $(k \times 1)$ vector of random error terms (Searle, Casella, & McCulloch, 1992). We assume $E[\boldsymbol{\gamma}] = \boldsymbol{0}$, $E[\boldsymbol{e}] = \boldsymbol{0}$, and $\mathrm{Cov}[\boldsymbol{\gamma}, \boldsymbol{e}] = \boldsymbol{0}$. Define $\boldsymbol{D}$ as the $(q \times q)$ covariance matrix of the random effects in $\boldsymbol{\gamma}$ and $\boldsymbol{R}$ as the $(k \times k)$ covariance matrix of $\boldsymbol{e}$. Then $\boldsymbol{V}$, the $(k \times k)$ covariance matrix of $\boldsymbol{y}$, is equal to $\boldsymbol{ZDZ}' + \boldsymbol{R}$. Typically, we impose some structure on the $k(k+1)/2$ parameters of $\boldsymbol{V}$ such that its elements are a function of m unobservable parameters, which we collectively denote by the vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_m)$, where m is typically substantially smaller than $k(k+1)/2$. The parameter space is given by $\{\boldsymbol{\beta}, \boldsymbol{\phi}\colon \boldsymbol{\beta} \in \mathbb{R}^p, \boldsymbol{\phi} \in \Omega\}$, where Ω is a subset of Euclidean m space such that $\boldsymbol{V}$ is non-singular. The parameters in $\boldsymbol{\phi}$ are usually estimated by ML or REML (Harville, 1977), which requires additional distributional assumptions. Typically, we assume that $\boldsymbol{\gamma} \sim N(\boldsymbol{0}, \boldsymbol{D})$ and $\boldsymbol{e} \sim N(\boldsymbol{0}, \boldsymbol{R})$, and consequently $\boldsymbol{y} \sim N(\boldsymbol{X\beta}, \boldsymbol{V})$.

4. Meta-analytic models in the GLMM framework

The meta-analytic random-effects model will now be put in the GLMM framework. The vector $\boldsymbol{y}$ consists of the k independent effect size estimates. The design matrix $\boldsymbol{X}$ is composed of a single column of 1s, corresponding to μ_θ, which is the only parameter in $\boldsymbol{\beta}$. Moreover, $\boldsymbol{Z}$ is the $(k \times k)$ identity matrix, $\boldsymbol{\gamma}$ is comprised of the τ_i-values at the population level, and $\boldsymbol{e}$ includes the random error terms, $\varepsilon_1, \ldots, \varepsilon_k$. Then $\boldsymbol{V}$ is diagonal with $v_i = (\sigma^2_\theta + \sigma^2_{\varepsilon_i})$ and $\boldsymbol{y} \sim \boldsymbol{N}(\mu_\theta \boldsymbol{1}, \boldsymbol{V})$, where $\boldsymbol{1}$ denotes a $(k \times 1)$ vector of 1s. The model can be written out explicitly in matrix form as follows:

$$\begin{bmatrix} ES_1 \\ ES_2 \\ \vdots \\ ES_k \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} [\mu_\theta] + \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ & & \ddots & \\ 0 & 0 & \dots & 1 \end{bmatrix} \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_k \end{bmatrix}.$$

A problem with this model is the large number of parameters. Specifically, there are $k + 1$ variance components in $\boldsymbol{\phi} = (\sigma^2_\theta, \sigma^2_{\varepsilon_1}, \dots, \sigma^2_{\varepsilon_k})$ and one fixed-effects parameter in $\boldsymbol{\beta}$ (i.e. μ_θ), but only k effect size estimates in $\boldsymbol{y}$. Therefore, estimating all parameters simultaneously, for example via ML estimation, becomes problematic due to model identifiability. This situation is related to the Neyman–Scott problem (Neyman & Scott, 1948), where a so-called incidental parameter is added to the model for each additional observation, leading to an inconsistent system of estimators.

To circumvent this difficulty, the common practice in meta-analysis is to estimate the parameters in two stages. First, estimates of $\sigma^2_{\varepsilon_i}$ are obtained algebraically based on the sampling theory underlying a particular effect size measure. Treating these estimates as known (i.e. ignoring their sampling variability) leaves only two parameters to be estimated, namely μ_θ and σ^2_θ, which yields an identifiable model. ML and REML estimation can then be applied in the usual manner, as will now be discussed.

4.1. Maximum likelihood estimation

If we treat the sampling variances as known, then the log-likelihood function of μ_θ and σ^2_θ is given by

$$\ln L(\mu_\theta, \sigma^2_\theta) = -\frac{1}{2}\sum_{i=1}^{k} \ln(\sigma^2_\theta + \sigma^2_{\varepsilon_i}) - \frac{1}{2}\sum_{i=1}^{k} \frac{(ES_i - \mu_\theta)^2}{\sigma^2_\theta + \sigma^2_{\varepsilon_i}}, \tag{7}$$

leaving out the additive constant. The likelihood equations are obtained by setting the partial derivatives with respect to μ_θ and σ^2_θ equal to zero and solving for the two parameters to be estimated. Doing so yields

$$\hat{\mu}^{(\mathrm{ML})}_\theta = \frac{\sum_{i=1}^{k} w_i ES_i}{\sum_{i=1}^{k} w_i} \tag{8}$$

and

$$\hat{\sigma}^{2(\mathrm{ML})}_\theta = \frac{\sum_{i=1}^{k} w_i^2 \left[\left(ES_i - \hat{\mu}^{(\mathrm{ML})}_\theta \right)^2 - \sigma^2_{\varepsilon_i} \right]}{\sum_{i=1}^{k} w_i^2}, \tag{9}$$

with $w_i = 1/(\sigma^2_\theta + \sigma^2_{\varepsilon_i})$.

Since $\hat{\mu}^{(\mathrm{ML})}_\theta$ depends on $\hat{\sigma}^{2(\mathrm{ML})}_\theta$ and vice versa, an iterative approach must be used to obtain solutions to these equations. Typically, one starts out with an initial guess for $\hat{\sigma}^{2(\mathrm{ML})}_\theta$ and then iterates between (8) and (9) until convergence (Brockwell & Gordon, 2001; Erez, Bloom, & Wells, 1996; National Research Council, 1992). Usually, only a few iterations are necessary for the solution to stabilize. It can be shown that this approach is

identical to applying the Fisher scoring algorithm to this estimation problem (Viechtbauer, 2004).

4.2. Restricted maximum likelihood estimation

Viechtbauer (2005) showed that the maximum likelihood estimator (MLE) of σ^2_θ is negatively biased. In fact, it is generally known that MLEs of variance components in the GLMM are negatively biased (Corbeil & Searle, 1976a; Patterson & Thompson, 1971, 1974). REML estimation (Corbeil & Searle, 1976b; Harville, 1977; Patterson & Thompson, 1971, 1974) is known to reduce, or in this case essentially removes, the bias in the variance component estimate (Viechtbauer, 2005). The restricted log-likelihood function for the random-effects model is given by

$$\ln L_{\mathrm{R}}(\sigma^2_\theta) = -\frac{1}{2}\sum_{i=1}^{k}\ln(\sigma^2_\theta + \sigma^2_{\varepsilon_i}) - \frac{1}{2}\ln\sum_{i=1}^{k}\frac{1}{\sigma^2_\theta + \sigma^2_{\varepsilon_i}} - \frac{1}{2}\sum_{i=1}^{k}\frac{(ES_i - \hat{\mu}^{(\mathrm{ML})}_\theta)^2}{\sigma^2_\theta + \sigma^2_{\varepsilon_i}}, \quad (10)$$

leaving out additive constants. Setting the derivative with respect to σ^2_θ equal to zero and solving leads to the REML estimator of σ^2_θ, which is given by

$$\hat{\sigma}^{2(\mathrm{REML})}_\theta = \frac{\sum_{i=1}^{k} w_1^2\left[\left(ES_i - \hat{\mu}^{(\mathrm{ML})}_\theta\right)^2 - \sigma^2_{\varepsilon_i}\right]}{\sum_{i=1}^{k} w_i^2} + \frac{1}{\sum_{i=1}^{k} w_i}, \quad (11)$$

with w_i defined as before. The same iterative scheme as described for the MLE can be employed to obtain a solution for (11).

An approximation to the REML estimator frequently found in the literature (e.g. Berkey, Hoaglin, Mosteller, & Colditz, 1995; Normand, 1999; Thompson & Sharp, 1999) is given by

$$\hat{\sigma}^{2(\mathrm{REML})}_\theta \approx \frac{\sum_{i=1}^{k} w_1^2\left[k(k-1)^{-1}\left(ES_i - \hat{\mu}^{(\mathrm{ML})}_\theta\right)^2 - \sigma^2_{\varepsilon_i}\right]}{\sum_{i=1}^{k} w_i^2}. \quad (12)$$

It appears that (12) originated with C. N. Morris (1983), who suggested it as an 'approximate' REML estimator (p. 53). While (12) and the exact REML estimator are equal to each other when the sampling variances are homogeneous – that is $\sigma^2_{\varepsilon_i} = \sigma^2_\varepsilon, i = 1, \ldots, k$ (Viechtbauer, 2004) — this is rare in practice and therefore (11) should be preferred.

4.3. Other population heterogeneity estimators

Several other estimators for σ^2_θ have been suggested in the literature. Viechtbauer (2005) examined and contrasted a total of five different estimators for σ^2_θ analytically and via Monte Carlo simulations. The five estimators examined included one suggested by Hunter and Schmidt (1990) from the validity generalization literature, one proposed by Hedges (1983, 1989), an estimator by DerSimonian and Laird (1986), the MLE, and the REML estimator. The MLE was found to be negatively biased as expected, while the REML estimator was essentially bias-free. On the other hand, the MLE was more efficient and had lower mean-squared error (MSE), in particular when k was small. The Hunter and Schmidt (HS) estimator was found to have similar properties to the MLE, while the

DerSimonian and Laird (DL) estimator performed similarly in terms of bias and MSE to the REML estimator. Hedges's (HE) estimator was bias-free, but had the largest MSE of all estimators considered. The class of (approximately) unbiased estimators was therefore composed of the REML, DL and HE estimators. However, since the sampling variability of the HE estimator generally exceeded that of the REML estimator and since the DL estimator was not always well defined asymptotically, Viechtbauer (2005) recommended the REML estimator for general use.

5. Homogeneity tests

As was mentioned earlier, the random-effects model reduces to the fixed-effects model when $\sigma_\theta^2 = 0$. Therefore, when σ_θ^2 is estimated to be zero, then the fixed-effects model is automatically obtained (negative estimates of σ_θ^2 are usually set equal to zero and lead to the same conclusion). Estimating the amount of population heterogeneity in the effect sizes is one possible approach for model selection. However, even when $\sigma_\theta^2 = 0$, estimates of σ_θ^2 can be greater than zero simply due to sampling fluctuations. Therefore, one can either employ substantive considerations to judge whether a certain amount of population heterogeneity is 'large' or employ one of the following hypothesis tests.

5.1. Q test

The Q test is the most frequently applied test in meta-analysis to determine whether the population effect sizes are homogeneous. Some of the earliest references to the Q test in the meta-analytic context can be found in Hedges (1982a, 1982b, 1983), Rosenthal and Rubin (1982), and Viana (1980). Earlier references to the test, foreshadowing its use in meta-analysis, can be found in Cochran (1937, 1954) and Rao (1973).

When the null hypothesis H_0: $\theta_1 = \ldots = \theta_k$ (i.e. H_0: $\sigma_\theta^2 = 0$) is true and $\overline{ES}$ is given by (8) with $w_i = 1/\sigma_{\varepsilon_i}^2$, then

$$Q = \sum_{i=1}^{k} w_i (ES_i - \overline{ES})^2 \tag{13}$$

is distributed as chi-squared with $k - 1$ degrees of freedom. As the variability in the observed effect sizes starts to exceed the amount of variability due to sampling error alone, Q will increase accordingly. Therefore, rejection of H_0 depends on whether the observed test statistic exceeds the $100(1 - \alpha)$th percentile of a chi-squared random variable at a chosen Type I error rate given by α.

If the value of Q is calculated with $w_i = 1/\hat{\sigma}_{\varepsilon_i}^2$, where $\hat{\sigma}_{\varepsilon_i}^2$ is some consistent estimator of the sampling variance, and/or the ES_i -values are only normally distributed for large sample sizes, then the distribution of Q is asymptotically chi-squared. It should be emphasized that the asymptotic behaviour of the Q statistic then depends only on the within-study sample sizes being large and not on the value of k. In other words, the distribution of Q under H_0 is exactly chi-squared when the sample sizes of all k studies become large, even when k itself is small.

A substantial number of simulation studies have been carried out to investigate the Type I error rate and power of the Q test (Alexander, Scozzaro, & Borodkin, 1989; Callender & Osburn, 1988; Field, 2001; Hardy & Thompson, 1998; Harwell, 1997; Hedges, 1982a, 1982b; Johnson, Mullen, & Salas, 1995; Koslowsky & Sagie, 1993; Morris, 2000; Rasmussen & Loher, 1988; Sackett, Harris, & Orr, 1986; Sagie & Koslowsky, 1993;

Sánchez-Meca & Marín-Martínez, 1997; Schmidt & Hunter, 1999; Spector & Levine, 1987). Also, Hedges and Pigott (2001) derived equations for calculating the approximate power of the Q test. Based on this collection of research, some general conclusions are warranted.

The studies suggest that the Q test generally keeps the Type I error rate close to the nominal α-value when assumptions underlying the effect size measures are not severely violated and within-study sample sizes are not too small. On the other hand, the Q test lacks sufficient power to detect heterogeneity when the within-study sample sizes and/or the number of studies are small. However, power of the test seems adequate (here defined as exceeding about .80) when based on at least 40 effect size estimates and the sample sizes exceed at least 40 observations (more precisely, a sample size of 40 observations per experimental group when using the SMD and 40 observations in total when using the correlation coefficient as the effect size measure). Nevertheless, if the amount of heterogeneity in the population effect sizes is small, then it is unlikely to be detected. In those cases, sample sizes exceeding 100 observations within each study and/or a larger number of effect size estimates would be needed.

5.2. Likelihood ratio test

As shown earlier, the meta-analytic models considered here are just special cases of the GLMM. Likelihood ratio (LR) tests (Hartley & Rao, 1967; Verbeke & Molenberghs, 1997, 2003) are frequently used in the GLMM framework to test the significance of variance components. This immediately raises the possibility of using LR tests in meta-analysis to examine the null hypothesis that the amount of population heterogeneity is zero. The test is carried out as follows. Let $\hat{\theta}^{(\mathrm{ML})}$ be the value of (8) for $w_i = 1/\sigma^2_{\varepsilon_i}$, which is the MLE of θ under the fixed-effects model. Next, we obtain the ML estimates $\hat{\mu}^{(\mathrm{ML})}_\theta$ and $\hat{\sigma}^{2(\mathrm{ML})}_\theta$ iteratively as described earlier. Then under H_0,

$$LR = -2\left(\ln L(\hat{\theta}^{(\mathrm{ML})}, 0) - \ln L\left(\hat{\mu}^{(\mathrm{ML})}_\theta, \hat{\sigma}^{2(\mathrm{ML})}_\theta\right)\right) \tag{14}$$

is asymptotically distributed as a 50:50 mixture of a degenerate random variable with all of its probability mass concentrated at 0 and a chi-square random variable with one degree of freedom. This follows from results obtained by Stram and Lee (1994), which in turn are based on Self and Liang (1987). Therefore, when $P(\chi^2_1 > LR) < 2\alpha$, where χ^2_1 is a chi-squared random variable with one degree of freedom, we reject the null hypothesis. Alternatively, when using REML estimation, the test statistic is given by

$$LR_{\mathrm{R}} = -2\left(\ln L_{\mathrm{R}}(0) - \ln L_{\mathrm{R}}\left(\hat{\sigma}^{2(\mathrm{REML})}_\theta\right)\right) \tag{15}$$

and we reject H_0 when $P(\chi^2_1 > LR_{\mathrm{R}}) < 2\alpha$.

As defined above, these two test statistics approach the given mixture distribution when k becomes large. However, we typically do not know the values of $\sigma^2_{\varepsilon_i}$ exactly and instead must rely on consistent estimates thereof. Therefore, $\hat{\theta}^{(\mathrm{ML})}$, $\hat{\mu}^{(\mathrm{ML})}_\theta$, $\hat{\sigma}^{2(\mathrm{ML})}_\theta$ and $\hat{\sigma}^{2(\mathrm{REML})}_\theta$ are only truly ML/REML estimates when the within-study sample sizes are large. Consequently, the correct asymptotic behaviour of the LR and LR_{R} statistics requires not only k to become large, but also large sample sizes within each of the k studies.

The properties of the LR tests for examining the homogeneity of effect sizes have not been studied extensively. Hardy and Thompson (1996) briefly mention the test statistic LR, but do not provide any results regarding its statsitical properties. However,

because σ_θ^2 falls on the boundary of the parameter space under the null hypothesis, the usual likelihood principles (Hartley & Rao, 1967) do not apply (Self & Liang, 1987; Stram & Lee, 1994) and the correct distribution under H_0 is the aforementioned mixture distribution.

LR testing was also briefly mentioned by DerSimonian and Laird (1986) and Hedges and Olkin (1985), but not further examined because of the computational difficulties involved in obtaining ML estimates. The only study examining the Type I error rate and power of the LR test in the meta-analytic context was conducted by Takkouche, Cadarso-Suárez, and Spiegelman (1999). However, their paper did not actually describe how to obtain the ML estimates and neither REML estimation nor the LR_R statistic was mentioned. Also, the paper focused specifically on the odds ratio and relative risk as effect size measures, which are less frequently used in the behavioural sciences than mean differences and correlations. Nevertheless, Takkouche *et al.* found the *LR* statistic to be somewhat conservative in controlling the Type I error rate and its power was slightly lower than that of the *Q* test. However, without further study, it is unclear whether this is also true for the LR_R statistic and for other effect size measures.

5.3. Wald test

Under certain regularity conditions, estimators based on the likelihood principle can be shown to be consistent, asymptotically fully efficient, and asymptotically normally distributed (Lehmann, 1999). In those cases, the inverse of the Fisher information evaluated at the parameter estimate provides a consistent estimate of the asymptotic sampling variance of the MLE. Dividing the MLE by the estimated standard error yields the Wald statistic, which can be compared against the critical values of a standard normal distribution to test whether the parameter is significantly different from zero.

The asymptotic sampling variances of $\hat{\sigma}_\theta^{2(\mathrm{ML})}$ and $\hat{\sigma}_\theta^{2(\mathrm{REML})}$ are given by

$$\mathrm{Var}_\infty\left[\hat{\sigma}_\theta^{2(\mathrm{ML})}\right] = 2\left(\sum_{i=1}^{k} w_i^2\right)^{-1} \tag{16}$$

and

$$\mathrm{Var}_\infty\left[\hat{\sigma}_\theta^{2(\mathrm{REML})}\right] = 2\left(\sum_{i=1}^{k} w_i^2 - 2\frac{\sum_{i=1}^{k} w_i^3}{\sum_{i=1}^{k} w_i} + \frac{\left(\sum_{i=1}^{k} w_i^2\right)^2}{\left(\sum_{i=1}^{k} w_i\right)^2}\right)^{-1}, \tag{17}$$

respectively. Estimates of the sampling variances are obtained by evaluating the equations with the corresponding estimates of σ_θ^2, namely, by setting w_i to $1/(\hat{\sigma}_\theta^{2(\mathrm{ML})} + \hat{\sigma}_{\varepsilon_i}^2)$ and $1/(\hat{\sigma}_\theta^{2(\mathrm{REML})} + \hat{\sigma}_{\varepsilon_i}^2)$, respectively. The corresponding Wald statistics are then given by

$$z = \frac{\hat{\sigma}_\theta^{2(\mathrm{ML})}}{\sqrt{\mathrm{Var}_\infty\left[\hat{\sigma}_\theta^{2(\mathrm{ML})}\right]}} \tag{18}$$

and

$$z_R = \frac{\hat{\sigma}_\theta^{2(\mathrm{REML})}}{\sqrt{\mathrm{Var}_\infty\left[\hat{\sigma}_\theta^{2(\mathrm{REML})}\right]}}. \tag{19}$$

Let $z_{1-\alpha}$ be the $100(1-\alpha)$th percentile of the standard normal distribution. Then we reject H_0 if $z > z_{1-\alpha}$ when using the ML-based test and $z_R > z_{1-\alpha}$ when using its

REML-based counterpart. Note that these are one-sided tests since the alternative hypothesis, $H_A : \sigma_\theta^2 > 0$, is also one-sided.

However, testing whether a variance component is equal to zero implies that the parameter falls on the boundary of the parameter space under the null hypothesis. Standard likelihood principles are no longer valid in this case (Self & Liang, 1987; Stram & Lee, 1994; Verbeke & Molenberghs, 1997, 2003). Therefore, the Wald tests are not expected to control the Type I error rate adequately. However, it is unknown to what extent these tests can still be used as approximations or as a rough first check for testing the homogeneity of effect sizes in the meta-analytic context. This is of particular concern as Wald tests (or corresponding confidence intervals) are still used occasionally for testing the homogeneity of effect sizes (Wang & Bushman, 1999).

5.4. Score test

Rao's score test (Lehmann, 1999) can also be adopted as another alternative for testing the homogeneity of effect sizes. The score function is defined as the first derivative of the log-likelihood evaluated under H_0. Dividing the score function by the square root of the Fisher information yields the score statistic, which can be compared against the critical values of a standard normal distribution.

Based on the log-likelihood given in (7), we obtain the score statistic

$$s = \frac{\sqrt{\frac{1}{2}\sum_{i=1}^{k} w_i^2 \left[(ES_i - \mu_0)^2 - \sigma_{\varepsilon_i}^2 \right]}}{\sqrt{\sum_{i=1}^{k} w_i^2}}, \tag{20}$$

where $w_i = 1/\sigma_{\varepsilon_i}^2$. Note that (20) contains the unknown parameter μ_θ, which we must replace with some sample estimate to obtain a usable test statistic. Recall that $\hat{\theta}^{(\mathrm{ML})}$ was earlier defined as the value of (8) evaluated with $w_i = 1/\sigma_{\varepsilon_i}^2$. Specifically, $\hat{\theta}^{(\mathrm{ML})}$ is the MLE of the population effect size θ under the fixed-effects model. Since the score function is evaluated under the null hypothesis, an asymptotically equivalent test is obtained by substituting $\hat{\theta}^{(\mathrm{ML})}$ for μ_θ in (20). Alternatively, starting with the restricted log-likelihood function given by (10) yields the test statistic

$$s_R = \frac{\sqrt{\frac{1}{2}\sum_{i=1}^{k} w_i^2 \left[\left(ES_i - \hat{\theta}^{(\mathrm{ML})}\right)^2 - \sigma_{\varepsilon_i}^2 + \left(\sum_{i=1}^{k} w_i\right)^{-1} \right]}}{\sqrt{\sum_{i=1}^{k} w_i^2 - 2\left(\sum_{i=1}^{k} w_i^3\right)/\sum_{i=1}^{k} w_i + \left(\sum_{i=1}^{k} w_i^2\right)^2 / \left(\sum_{i=1}^{k} w_i\right)^2}}, \tag{21}$$

which automatically involves $\hat{\theta}^{(\mathrm{ML})}$ without requiring any further substitutions. In practice, we must set $w_i = 1/\hat{\sigma}_{\varepsilon_i}^2$ and replace $\sigma_{\varepsilon_i}^2$ with $\hat{\sigma}_{\varepsilon_i}^2$ in (20) and (21). Again, one would reject the null hypothesis when s or s_R exceeds $z_{1-\alpha}$, depending on which of the two tests is being used. Note that these are one-sided score tests (Verbeke & Molenberghs, 2003).

One advantage of the score test over the LR and Wald tests is that it does not require explicit estimation of the parameter being tested. This is particularly useful when the parameter estimate does not have a closed-form solution and must be found numerically, as in the case considered here when using either ML or REML estimation. Although some non-iterative estimators of σ_θ^2 exist, the score test avoids this additional computational step altogether. No reference to the score test for examining the homogeneity of effect sizes in the meta-analytic context could be found in the literature. Therefore, it is unknown at this point whether the score test is a viable alternative to the other tests.

6. Examples

Two examples will illustrate the use of the different homogeneity tests and demonstrate that they can provide conflicting conclusions about the presence of population heterogeneity. The first data set, given in Table 1, provides the results for $k = 10$ studies that examined the effectiveness of open versus traditional education programmes on student creativity (Hedges & Olkin, 1985, p. 25). The table lists the effect size estimate (ES_i), the estimated sampling variance ($\hat{\sigma}^2_{\varepsilon_i}$), and the inverse sampling variance weight ($w_i = 1/\hat{\sigma}^2_{\varepsilon_i}$) for each study.[1]

Table 1. Results for 10 studies of the effectiveness of open versus traditional education on student creativity

Study	Effect size (ES_i)	Variance ($\hat{\sigma}^2_{\varepsilon_i}$)	Weights ($w_i = 1/\hat{\sigma}^2_{\varepsilon_i}$)
1	−0.581	0.023	43.478
2	0.530	0.052	19.231
3	0.771	0.060	16.667
4	1.031	0.115	8.696
5	0.553	0.095	10.526
6	0.295	0.203	4.926
7	0.078	0.200	5.000
8	0.573	0.210	4.762
9	−0.176	0.051	19.608
10	−0.232	0.040	25.000

Note. The ES_i are unbiased estimates of the SMD effect size measure. The data were obtained from Hedges and Olkin (1985, p. 25).

The weighted average of the effect size estimates (Equation (8) with $w_i = 1/\hat{\sigma}^2_{\varepsilon_i}$) is equal to 0.050. The value of Q is then easily computed with (13) and is found to be 46.03. The 95th percentile of a chi-squared random variable with $k - 1 = 9$ degrees of freedom is equal to 16.92 and therefore the Q test indicates the presence of heterogeneity over and beyond that which we would expect based on sampling error alone.

To apply the LR tests, we must first obtain the ML and REML estimates of σ^2_θ. Using the iterative scheme described earlier, we find $\hat{\sigma}^{2(\mathrm{ML})}_\theta = 0.197$ and $\hat{\sigma}^{2(\mathrm{REML})}_\theta = 0.223$. Knowing this, we set $w_i = 1/(1\hat{\sigma}^{2(\mathrm{ML})}_\theta + \hat{\sigma}^2_{\varepsilon_i})$ and then find $\hat{\mu}^{(\mathrm{ML})}_\theta = 0.240$ based on (8). Finally, the values of the LR tests are obtained from (14) and (15) and are equal to 24.42 and 25.98, both of which are significant when compared to the mixture distribution (both p-values are smaller than .001).

The asymptotic sampling variances of the ML and REML estimates of σ^2_θ are given by (16) and (17) and turn out to be 0.016 and 0.021, respectively. The Wald statistics are then easily computed with (18) and (19) and are equal to 1.57 and 1.52. Compared to 1.65, the 95th percentile of a standard normal random variable, we conclude that the effect sizes are homogeneous.

[1] *The two examples discussed in the present section use the SMD (to be discussed in more detail below) as the effect size measure. The* ES_i *are the unbiased effect size estimates.*

Finally, the score statistics are computed with (20) and (21) where $w_i = 1/\hat{\sigma}^2_{\varepsilon_i}$ and μ_θ in (20) is replaced with 0.050, the weighted average of the effect size estimates under the fixed-effects model. We find $s = 11.49$ and $s_R = 13.53$, both of which are significant when compared against the 95th percentile of a standard normal random variable. Therefore, the score tests indicate the presence of population heterogeneity.

In this particular example, the Q, LR, and score tests all indicate heterogeneous effect sizes, while the two Wald tests suggest the opposite. A different pattern of results is obtained for the data in Table 2, which provides results for $k = 18$ studies comparing open versus traditional education using student self-concept as the outcome variable (Hedges & Olkin, 1985, p. 25). Here, $Q = 23.46$, which is not significant when compared against the 95th percentile of a chi-square random variable with 17 degrees of freedom. The two Wald tests ($z = 0.88$ and $z_R = 0.94$) and the two LR tests ($LR = 1.71$, $p = .10$ and $LR_R = 2.14$, $p = .07$) also indicate the absence of heterogeneity in the effect sizes. However, the score statistics are equal to $s = 1.73$ and $s_R = 2.03$, which would lead to the rejection of the null hypothesis. In this particular example, the two score tests indicate heterogeneity, while using the Q, LR, and Wald tests would lead to the opposite conclusion.

Table 2. Results for 18 studies of the effectiveness of open versus traditional education on student self-concept

Study	Effect size (ES_i)	Variance ($\hat{\sigma}^2_{\varepsilon_i}$)	Weights ($w_i = 1/\hat{\sigma}^2_{\varepsilon_i}$)
1	.100	.016	62.500
2	−.162	.015	66.667
3	−.090	.050	20.000
4	−.049	.050	20.000
5	−.046	.032	31.250
6	−.010	.052	19.231
7	−.431	.036	27.778
8	−.261	.024	41.667
9	.134	.034	29.412
10	.019	.033	30.303
11	.175	.031	32.258
12	.056	.034	29.412
13	.045	.039	25.641
14	.103	.167	5.988
15	.121	.134	7.463
16	−.482	.096	10.417
17	.290	.016	62.500
18	.342	.035	28.571

Note. The ES_i are unbiased estimates of the SMD effect size measure. The data were obtained from Hedges and Olkin (1985, p. 25).

7. Monte Carlo simulations

The examples in the previous section demonstrate that the results from the various homogeneity tests can lead to conflicting conclusions about the presence of heterogeneity in the effect sizes. However, without further analysis of the tests, it is unclear whether we should attribute more confidence to any particular test result.

Because the ML and REML estimates must be obtained numerically, it would be difficult to compare the properties of the various tests analytically. Instead, Monte Carlo simulations were conducted to assess the Type I error rate and power of these tests. Simulations were carried out with four different effect size measures: the UMD, the SMD, the raw correlation coefficient, and the variance-stabilized correlation coefficient after applying the Fisher transformation. The influence of four factors on the Type I error rate and power of the tests was examined: the number of effect sizes (k), the average effect size in the population (μ_θ), the amount of sampling error in the effect size estimates ($\sigma^2_{\varepsilon_i}$) and the amount of population heterogeneity (σ^2_θ).

Five different values of k were chosen, namely 5, 10, 20, 40 and 80, representing values typically seen in practice. For example, in 24 meta-analyses conducted in three research domains (medicine and behavioural medicine, social and clinical psychology, and organizational psychology), k ranged between 5 and 76 (Rosenthal & DiMatteo, 2001).

For the SMD and the correlation coefficient (and its variance-stabilized version), values of μ_θ were chosen according to Cohen's (1988) conventional definitions of small, medium, and large effect sizes, which also represent typical effect size values encountered in practice. The inclusion of various values of μ_θ for these effect size measures is important for the following reason. As discussed below, the shape of the distribution for these effect size measures is related to the location parameter θ_i. Since θ_i is a function of μ_θ, this relationship introduces a possible dependence between the performance of the homogeneity tests and μ_θ. On the other hand, the distribution of the UMD is exactly normal and does not depend (in its shape) on the location parameter. Therefore, different values of μ_θ should not have any influence on the performance of the homogeneity tests. This conjecture was checked by including very disparate values of μ_θ in the simulations for this effect size measure.

The amount of sampling error in the effect size estimates ($\sigma^2_{\varepsilon_i}$) was manipulated by adjusting the within-study sample sizes (as mentioned previously, $\sigma^2_{\varepsilon_i}$ is inversely related to the sample size). Within-study sample sizes between 20 and 640 observations were chosen to cover a very wide range of possible cases.

Finally, the values of σ^2_θ were chosen as follows. Naturally, $\sigma^2_\theta = 0$ was included in the simulations to investigate the performance of the tests when no population heterogeneity is present. The remaining values of σ^2_θ were chosen to cover a similar range $\sigma^2_{\varepsilon_i}$. Therefore, we obtain representative results for the condition where we have no population heterogeneity up to the case where we have a large amount of heterogeneity that is approximately equal to the amount of sampling variability in an effect size estimate from a very small study.

In each iteration of the simulations, k values of θ_i were generated from $N(\mu_\theta, \sigma^2_\theta)$. Next, k values of ES_i and $\hat{\sigma}^2_{\varepsilon_i}$ were generated from the appropriate distributions, as detailed in the methods sections below. The various test statistics were then computed and tested for significance with $\alpha = .05$. Any trial in which the ML or REML estimators did not converge was skipped and replaced by an additional trial. Overall, this occurred in about 0.04% of the trials and therefore should not affect the results substantially. The Type I error rate of a test was estimated by the proportion of iterations in which the true value of σ^2_θ was set to zero, but $H_0 : \sigma^2_\theta = 0$ was rejected. On the other hand, the proportion of iterations rejecting H_0 when $\sigma^2_\theta > 0$ indicates the empirical power of a test.

7.1. Unstandardized mean difference

7.1.1. Methods

Let X_{ij}^{C} and X_{ij}^{E} be the jth observations from a control and an experimental group in the ith study. Assume that $X_{ij}^{C} \sim N(\mu_i^{C}, \sigma_i^2)$ and $X_{ij}^{E} \sim N(\mu_i^{E}, \sigma_i^2)$. For the ith study, the unstandardized mean difference is defined as $\theta_i = \mu_i^{E} - \mu_i^{C}$. Given n_i^{C} and n_i^{E} observations from the control and experimental group, respectively, $ES_i = \bar{X}_i^{E} - \bar{X}_i^{C}$ provides an unbiased estimate of θ_i. The distribution of ES_i is $N(\mu_i^{E} - \mu_i^{C}, \sigma_i^2(1/n_i^{E} + 1/n_i^{C}))$ and the sampling variance of ES_i can be estimated without bias by $\hat{\sigma}_{\varepsilon_i}^2 = s_i^2(1/n_i^{E} + 1/n_i^{C})$, where s_i^2 is the usual pooled within-group variance.

The UMD is a useful effect size measure when all studies use a commensurable measurement scale to assess group differences (Bond, Wiitala, & Richard, 2003; Lipsey & Wilson, 2001). Moreover, the UMD provides a useful benchmark for the homogeneity tests, because all assumptions of the model (except known values of $\sigma_{\varepsilon_i}^2$) are satisfied. In particular, the effect size measure is exactly normally distributed, it is unbiased, and its sampling variance is independent of θ_i. Some or all of these properties only hold asymptotically for many other effect size measures that are commonly used. Therefore, the UMD allows us to examine how well the tests perform under what might be considered ideal conditions.

To make the number of simulated conditions more manageable, it was assumed that $n_i = n_i^{C} = n_i^{E}$ and σ_i^2 was set to 10.[2] Moreover, μ_i^{C} was set to zero and μ_i^{E} was sampled from $N(\mu_\theta, \sigma_\theta^2)$ to generate heterogeneous values of θ_i. The levels of the factors were as follows: $\mu_\theta = (0, 1, 10, 100)$, $\sigma_\theta^2 = (0, 0.125, 0.25, 0.5, 1)$ and $\bar{n}_i = (20, 40, 80, 160, 320, 640)$, and consequently, $\sigma_{\varepsilon_i}^2 = (1, 0.5, 0.25, 0.125, 0.0625, 0.3125)$. To simulate heterogeneous sample sizes (and therefore heterogeneous sampling variances), the values of n_i were sampled from a normal distribution with mean $\bar{n}_i$ and standard deviation $\bar{n}_i/3$. Including the various values for k, this yields a $5 \times 4 \times 5 \times 6$ factorial design with a total of 600 conditions. For each condition, 10,000 iterations were carried out. However, to get more accurate estimates of the Type I error rate, 100,000 iterations were used for the conditions where $\sigma_\theta^2 = 0$.

7.1.2. Results

Not surprisingly, the Type I error rate and power of the homogeneity tests did not depend on μ_θ. All subsequent results were therefore averaged over this factor for this effect size measure. The Type I error rates for the various homogeneity tests are shown in Figure 1 as a function of k and $\bar{n}_i$. The results for the two Wald tests were so similar (they never differed by more than 1%) that only those based on the z statistic were plotted (i.e. the results for the Wald test based on the ML estimate). The horizontal dashed line indicates the nominal $\alpha = .05$ value. In general, smaller values of $\bar{n}_i$ were associated with higher Type I error rates, but once $\bar{n}_i$ exceeded 80 observations, further increases in sample size resulted in little or no change in the behaviour of the tests. On the other hand, increasing k when $\bar{n}_i$ is small resulted in increasingly inflated Type I error rates for all of the tests.

The Type I error rate of the Q test approached the nominal α-value as the average within-study sample size increased. As expected, this was true regardless of the value of

[2] *For $n_i = n_i^{C} = n_i^{E}$, the sampling variance of ES_i is given by $\sigma_{\varepsilon i}^2 = 2\sigma_i^2/n_i$. Therefore, changes in $\sigma_{\varepsilon i}^2$ can be induced by manipulating n_i or σ_i^2. For simplicity, change in $\sigma_{\varepsilon i}^2$ was induced by setting $\sigma_i^2 = 10$ and manipulating n_i as detailed in the text.*

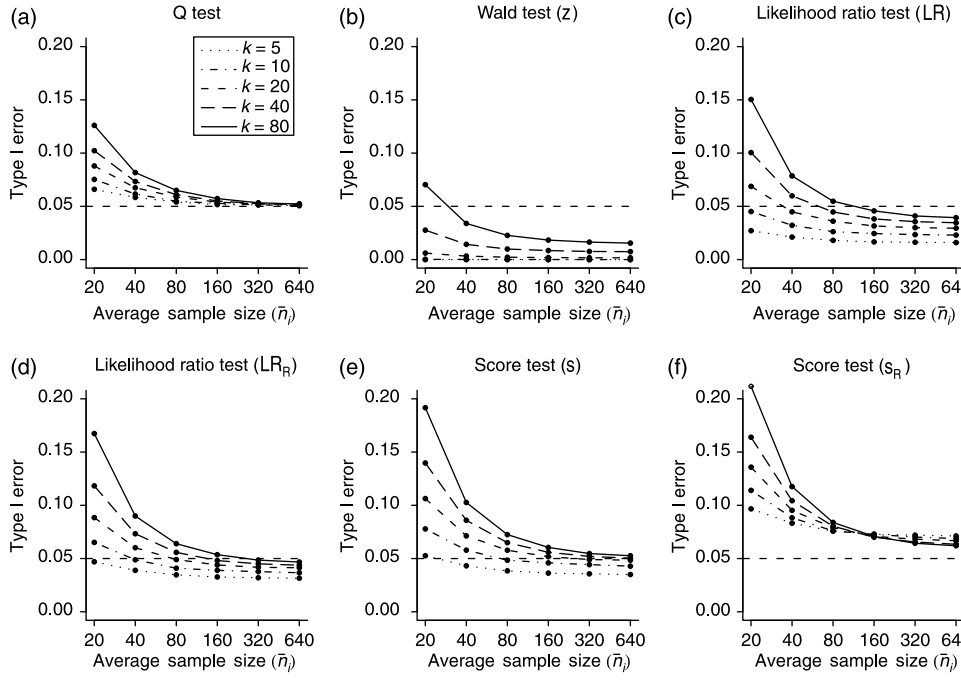

Figure 1. Type I error rates of the homogeneity tests for the unstandardized mean difference.

k. On the other hand, both Wald tests were overly conservative (except for $k = 80$ and $\bar{n}_i = 20$) and appeared to converge only very slowly to nominal behaviour as k increased. Within the conditions studied, the Wald tests never actually reached the nominal α level.[3]

The results for the LR tests were as expected. As hypothesized earlier, the Type I error rate converged to the nominal α-value only when both $\bar{n}_i$ and k increased. Larger values of $\bar{n}_i$ ensure that the assumption of known $\sigma^2_{\varepsilon_i}$ is approximately satisfied, which in turn ensures convergence of the LR statistics to the mixture distribution as k increases.[4] The ML-based LR test was slightly more conservative than its REML-based counterpart.

The last two homogeneity tests examined were the score statistics s and s_R. The Type I error rate of s converged to the nominal α-value with increasing $\bar{n}_i$ as long as k was slightly larger than 10. The s_R statistic, on the other hand, showed somewhat inflated Type I error rates even for large $\bar{n}_i$ and k and appeared to converge to a Type I error rate a few percentage points above the nominal value.

Figure 2 shows plots of the observed rejection rates as a function of σ^2_θ for nine different combinations of k and $\bar{n}_i$. When interpreting Figure 2, one must bear in mind that differences among the rejection rates are not only a function of how sensitive the

[3] *Additional simulations were run with larger values of k and $\bar{n}_i$ to examine whether the Type I error rate of the Wald tests would converge to .05. The behaviour of the tests approached the nominal α level only when a very large number of effect sizes (e.g. $k \geq 500$) was paired with large sample sizes (e.g., $\bar{n}_i \geq 640$). However, meta-analyses of this magnitude are unrealistic in practice and, therefore, the behaviour of the Wald tests is generally too conservative.*

[4] *The validity of this conjecture was further supported by running additional simulations where $\hat{\sigma}^2_{\varepsilon_i}$ was set to $\sigma^2_{\varepsilon_i}$, that is simulating the case where the sampling variances are exactly known even when the within-study sample sizes are small. In this case, the Type I error rate of the tests no longer depended on $\bar{n}_i$ and was approximately equal to that displayed in Figure 1 for $\bar{n}_i = 640$.*

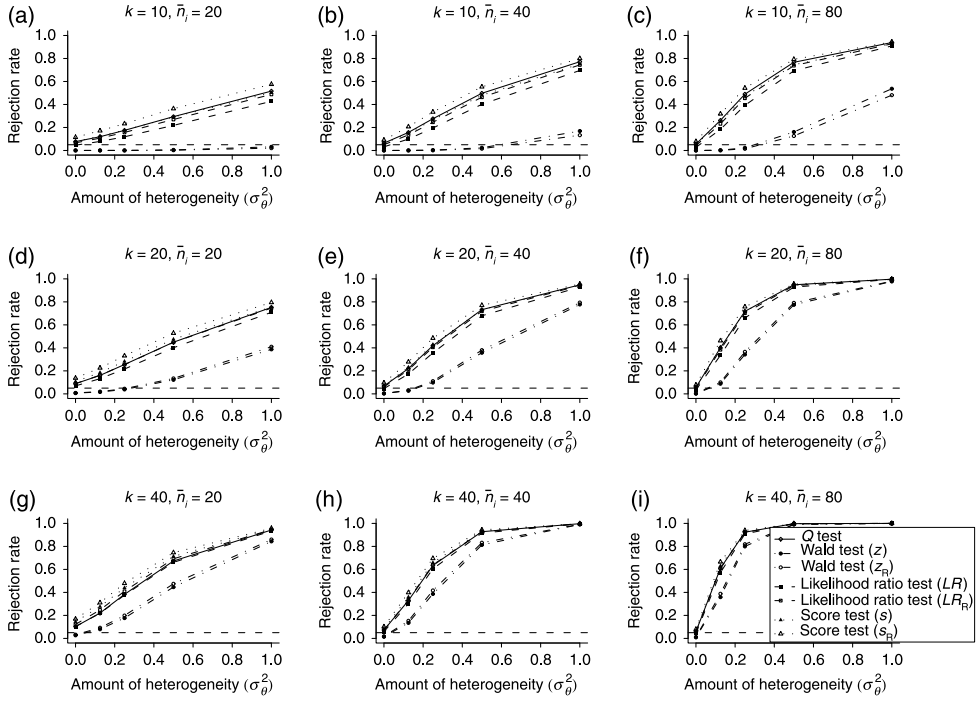

Figure 2. Rejection rates of the homogeneity tests for the unstandardized mean difference.

tests are in detecting heterogeneity, but also a function of how well the tests control the Type I error rate. For example, the Type I error rates of the two Wald tests were generally too conservative, which in turn reduces the probability that these tests will detect heterogeneity when it is in fact present. By adjusting the critical values of the tests such that the Type I error rates are exactly nominal, one could determine whether the tests differ strictly in their sensitivity to the presence of heterogeneity. However, such adjustments are not a feasible option in practice. Therefore, Figure 2 indicates how well the tests detect heterogeneity, bearing in mind that the rejection rates are also influenced by the actual Type I error rates of the tests.

Under these considerations, we note that increasing k, $\bar{n}_i$, and/or σ^2_θ resulted in higher probabilities of rejecting the null hypothesis and that the various homogeneity tests did not differ greatly with respect to their ability to detect heterogeneity, except for the two Wald tests, whose rejection rates were substantially lower than those of the other tests. The slight differences between the Q, LR, and score tests appear to be attributable to the fact that some of the tests did not control the Type I error rate adequately under certain conditions.

7.2. Standardized mean difference

7.2.1. Methods

When $X^{\mathrm{C}}_{ij} \sim N(\mu^{\mathrm{C}}_i, \sigma^2_i)$ and $X^{\mathrm{E}}_{ij} \sim N(\mu^{\mathrm{E}}_i, \sigma^2_i)$ as for the UMD and the measurement scale is not commensurable across studies, then the standardized mean difference is often chosen as an effect size measure. Now, $\theta_i = (\mu^{\mathrm{E}}_i - \mu^{\mathrm{C}}_i)/\sigma_i$, which can be estimated unbiasedly by $ES_i = c(m_i)(\bar{X}^{\mathrm{E}}_i = \bar{X}^{\mathrm{C}}_i)/s_i$, where

$$c(m_i) = \frac{\Gamma(m_i/2)}{(m_i/2)^{1/2}\Gamma((m_i - 1)/2)}$$

and $m_i = n_i^{\mathrm{E}} + n_i^{\mathrm{C}} - 2$ (Hedges, 1981). An unbiased estimate of $\sigma^2_{\varepsilon_i}$ is given by

$$\hat{\sigma}^2_{\varepsilon_i} = \frac{1}{\tilde{n}_i} + \left(1 - \frac{m_i - 2}{m_i[c(m_i)]^2}\right) ES_i^2,$$

where $\tilde{n}_i = (n_i^{\mathrm{E}} n_i^{\mathrm{C}})/(n_i^{\mathrm{E}} + n_i^{\mathrm{C}})$ (Hedges, 1983). The distribution of ES_i is asymptotically normal and is closely related to a non-central t-distribution. In fact, $c(m_i)^{-1}(\tilde{n}_i)^{1/2} ES_i$ is distributed as non-central t with m_i degrees of freedom and non-centrality parameter $\theta_i(\tilde{n}_i)^{1/2}$. Therefore, the distribution of ES_i is symmetric for $\theta_i = 0$ (and quickly approaches normality as m_i increases), while larger values of $|\theta_i|$ result in an increasingly skewed distribution of ES_i. The exact sampling variance of ES_i is equal to

$$\sigma^2_{\varepsilon_i} = \frac{[c(m_i)]^2 m_i[1 + \tilde{n}_i\theta_i^2]}{(m_i - 2)\tilde{n}_i} - \theta_i^2.$$

Note that $\sigma^2_{\varepsilon_i}$ depends on θ_i, which in turn is determined by τ_i. This violates the assumption that $\mathrm{Cov}[\tau_i, \varepsilon_i] = 0$.

Again, it was assumed that $n_i = n_i^{\mathrm{C}} = n_i^{\mathrm{E}}$. The levels of the various factors were as follows: $\mu_\theta = (0, 0.2, 0.5, 0.8)$, $\sigma^2_\theta = (0, 0.01, 0.25, 0.05, 0.1)$ and $\bar{n}_i = (20, 40, 80, 160, 320, 640)$, and consequently $\sigma^2_{\varepsilon_i}$ was roughly between 0.101 and 0.110 for $n_i = 20$ and around 0.003 for $n_i = 640$ (the exact value of $\sigma^2_{\varepsilon_i}$ depends on θ_i). We again obtain a $5 \times 4 \times 5 \times 6$ factorial design with 600 conditions in total. For each condition, 10,000 iterations were carried out, while 100,000 iterations were used for the $\sigma^2_\theta = 0$ condition.

7.2.2. Results

The Type I error rate of the homogeneity tests again did not depend on μ_θ, therefore results could be averaged over this factor. This finding was somewhat surprising, because of the aforementioned dependence between θ_i and the distribution of ES_i. It appears that, within the range of μ_θ-values studied, this dependence is not large enough to influence the performance of the homogeneity tests (in fact, an additional set of simulations revealed essentially unchanged Type I error rates even when μ_θ was set to 1.5, 2 or 3).

Figure 3 shows the Type I error rates of the homogeneity tests as a function of the average sample size and number of effect sizes. The results for the REML-based Wald test were again so similar to those for its ML counterpart that only the latter were plotted. All tests showed behaviour that indicated convergence to the nominal Type I error rate as $\bar{n}_i$ and k increased. The tests approached the nominal α level from below, except for the score test based on the s_{R} statistic, which overshot the .05 level slightly for larger values of $\bar{n}_i$. Generally, average sample size had relatively little influence on Type I error rates once $\bar{n}_i \geq 80$. The Q test controlled the Type I error rate quite well after that point, regardless of k. The convergence of the Wald tests, on the other hand, was very slow and did not reach nominal levels within the conditions studied. Of the two likelihood ratio tests, the REML-based one yielded results slightly closer to the nominal α level, but both tests were still a little bit too conservative even when $k = 80$. The score statistic s controlled the Type I error rate adequately as long as $k \geq 40$ and $\bar{n}_i \geq 160$, while the score test based on s_{R} was slightly too liberal when $\bar{n}_i > 80$ even when $k = 80$.

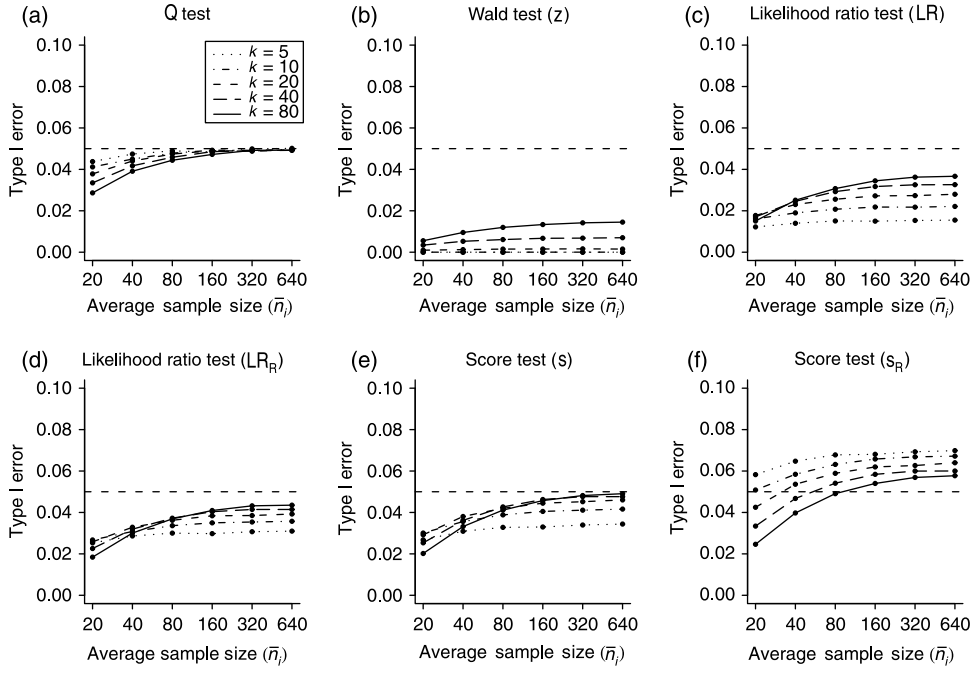

Figure 3. Type I error rates of the homogeneity tests for the standardized mean difference.

Figure 4 shows plots of the observed rejection rates as a function of σ_θ^2 for nine different combinations of k and $\bar{n}_i$. Keeping the considerations with respect to the relationship between Type I error rates and power in mind, the results were very similar to those found for the UMD: (a) larger values of k, $\bar{n}_i$, and/or σ_θ^2 were associated with higher probabilities of detecting heterogeneity when it was present, (b) the Wald tests had substantially lower rejection rates than the other tests, and (c) the rejection rates of the Q, LR, and score tests were very similar to each other.

7.3. Raw correlation coefficient

7.3.1. Methods

Let X_{ij} and Y_{ij} denote a pair of observations of the jth individual on two random variables in the ith study. Assume that X_{ij} and Y_{ij} are distributed bivariate normal with means μ_i^X, μ_i^Y, standard deviations σ_i^X, σ_i^Y, and correlation ρ_i. Now the effect size is defined simply as $\theta_i = \rho_i$. Given $j = 1, \ldots n_i$ pairs of scores from this bivariate distribution, we can estimate θ_i with the sample product-moment correlation coefficient r_i. Hotelling (1953) showed that r_i is a negatively biased estimator of its population value. An exactly unbiased estimator of θ_i was derived by Olkin and Pratt (1958) and is given by

$$r_i^{\mathrm{U}} = r_i F\left(\frac{1}{2}, \frac{1}{2}, \frac{n_i - 2}{2}, 1 - r_i^2\right), \tag{22}$$

where

$$F(\alpha, \beta, \gamma, \chi) = \sum_{j=0}^{\infty} \frac{\Gamma(\alpha + j)\Gamma(\beta + j)\Gamma(\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma + j)} \frac{\chi_j}{j!} \tag{23}$$

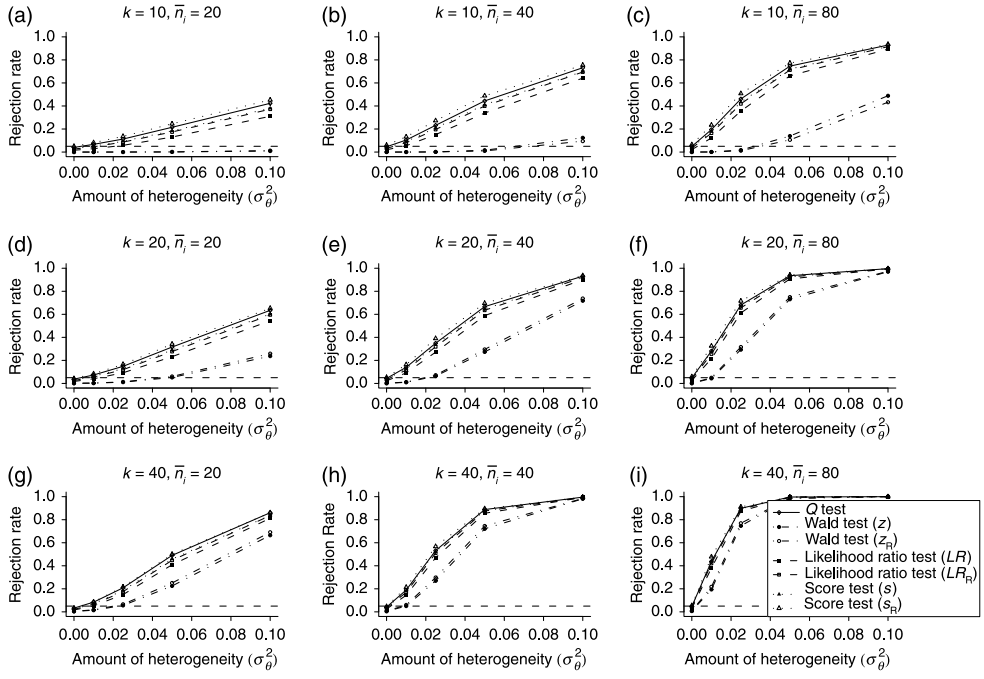

Figure 4. Rejection rates of the homogeneity tests for the standardized mean difference.

denotes the well-known hypergeometric function. The exactly unbiased estimator is closely approximated by

$$ES_i = r_i + \frac{r_i(1 - r_i^2)}{2(n_i - 4)}. \tag{24}$$

Note that one will often find $n_i - 3$ instead of $n_i - 4$ written in the denominator of the second term in (24). Olkin and Pratt used N_i to denote the sample size and n_i for the degrees of freedom, which are $N_i - 1$ in the case where all the parameters are unknown. Since that is the usual case, dividing by $n_i - 4$ yields a more accurate approximation than when dividing by $n_i - 3$.

The exact sampling variance of ES_i is unknown. However, since $ES_i \to r_i$ as $n_i \to \infty$, the large-sample approximation of the sampling variance of ES_i is the same as that of the regular correlation coefficient (Olkin & Pratt, 1958), namely,

$$\sigma^2_{\varepsilon_i} = \frac{(1 - \theta_i^2)^2}{n_i - 1}. \tag{25}$$

Replacing θ_i by either r_i or its (approximately) unbiased estimate ES_i yields a consistent estimate for the sampling variance, albeit a biased one. An approximately unbiased estimate of the sampling variance of ES_i is given by

$$\hat{\sigma}^2_{\varepsilon_i} = (ES_i)^2 - 1 + \left(\frac{n_i - 3}{n_i - 2}\right) \times \left((1 - r_i^2) + \frac{2(1 - r_i^2)^2}{n_i} + \frac{8(1 - r_i^2)^3}{n_i(n_i + 2)} + \frac{48(1 - r_i^2)^4}{n_i(n_i + 2)(n_i + 4)}\right) \tag{26}$$

(Hedges, 1989, with correction of his Equation 19, p. 474).

The distribution of ES_i is normal only when n_i is large. The skew in the distribution of ES_i depends on the extent to which $|\theta_i|$ diverges from 0. Finally, we note that the sampling variance of ES_i depends on the parameter θ_i. Therefore, τ_i and ε_i are not independent for this effect size measure.

For the simulations, k values of θ_i were first sampled from $N(\mu_\theta, \sigma^2_\theta)$. Values of θ_i outside the admissible range of -1 to 1 were truncated to $-.99$ and .99. Next, k sets of n_i random variables were sampled from bivariate standard normal distributions with $\theta_i = \rho_i$. Within each set, ES_i and $\hat{\sigma}^2_{\varepsilon_i}$ were computed using (24) and (26), respectively. The resulting k values of ES_i were then tested for homogeneity. The levels of the factors were: $\mu_\theta = (0,\ 0.1,\ 0.3,\ 0.5)$, $\sigma^2_\theta = (0, 0.005, 0.01, 0.02, 0.04)$, and $\bar{n}_i = (20, 40, 80, 160, 320, 640)$, and consequently, $\sigma^2_{\varepsilon_i}$ was roughly between 0.03 and 0.05 for $\bar{n}_i = 20$ and between 0.001 and 0.002 for $\bar{n}_i = 640$. This again yields a $5 \times 4 \times 5 \times 6$ factorial design. Because the data generating process is computationally more intensive for this effect size measure, the number of iterations was reduced to 1,000, with 100,000 iterations run for the $\sigma^2_\theta = 0$ conditions.

7.3.2. Results

The Type I error rate of the homogeneity tests depended on μ_θ when $\bar{n}_i \leq 40$, with smaller values of μ_θ being associated with more inflated Type I error rates, especially when k was large. However, once the average within-study sample size exceeded 40 observations, the behaviour of the homogeneity tests was essentially unaffected by changes in μ_θ. Figures 5 and 6 illustrate this by showing the Type I error rate of the tests as a function of k and $\bar{n}_i$ for $\mu_\theta = 0$ and $\mu_\theta = 0.5$, respectively. The two Wald tests again differed so little in their behaviour that only the results for the ML-based test were plotted.

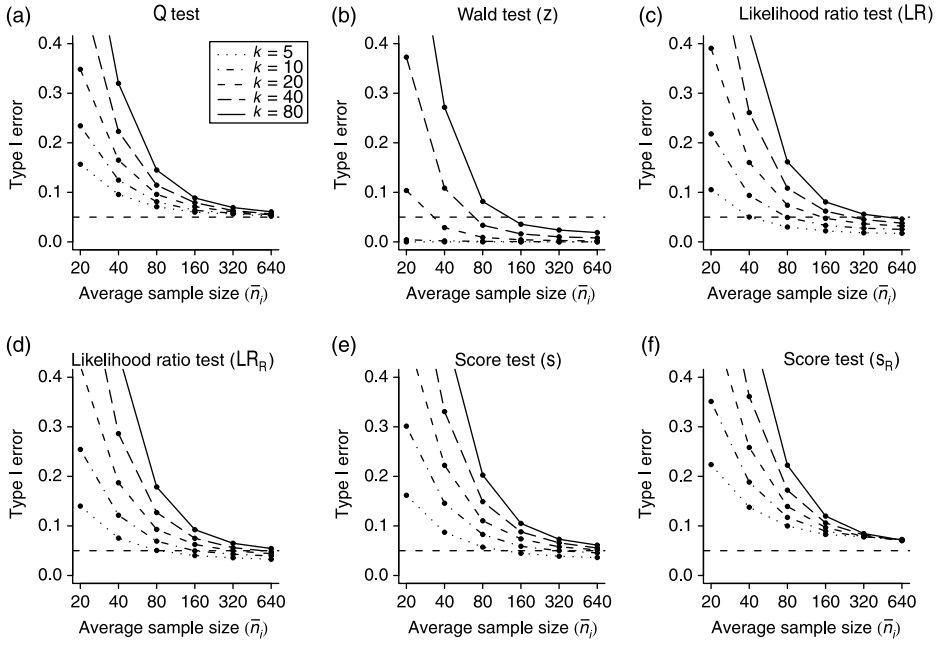

Figure 5. Type I error rates of the homogeneity tests for the correlation coefficient when $\mu_\theta = 0$.

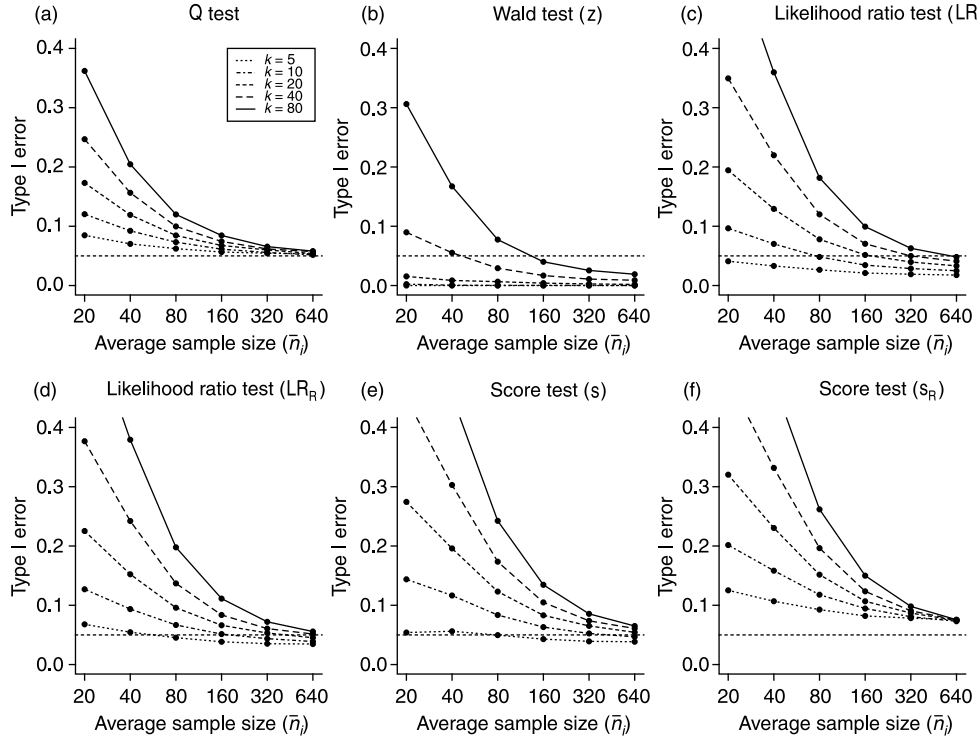

Figure 6. Type I error rates of the homogeneity tests for the correlation coefficient when $\mu_\theta = 0.5$.

The most noticeable finding in this set of simulations were the extremely inflated Type I error rates for small sample sizes, especially when k is large. In fact, increasing k resulted in progressively higher probabilities of rejecting the null hypothesis. The severity of this finding was so surprising that another set of simulations was run, this time using Equation (25) to estimate $\sigma^2_{\varepsilon_i}$ (with θ_i replaced by ES_i). However, the inflation in Type I error rate rates for small $\bar{n}_i$ was the same or even more extreme in this case.

As $\bar{n}_i$ increases, the Type I error rate of the Q test converges to nominal levels regardless of k. The two LR tests and the score test based on the s statistic controlled the Type I error rate only when both $\bar{n}_i$ and k were large. On the other hand, the score test based on s_R did not reach the nominal α level within the conditions studied and instead remained slightly too liberal. Finally, the Wald tests quickly became conservative with increasing $\bar{n}_i$ and began to converge very slowly to $\alpha = .05$ as k increased. Nevertheless, neither of the two Wald tests actually reached the nominal Type I error rate even when $k = 80$ and $\bar{n}_i = 640$.

Figure 7 shows plots of the observed rejection rates as a function of σ^2_θ for nine different combinations of k and $\bar{n}_i$ when $\mu_\theta = 0$. As before, increasing k, $\bar{n}_i$, and/or σ^2_θ results in higher probabilities of detecting heterogeneity when it is in fact present. The tests did not differ greatly in their behaviour, except for the two Wald tests which again revealed substantially lower rejection rates.

7.4. Correlation coefficient with Fisher transformation

7.4.1. Methods

For large values of $|\rho_i|$, the distribution of the raw correlation coefficient approaches the normal distribution only relatively slowly as the sample size increases. Therefore, several

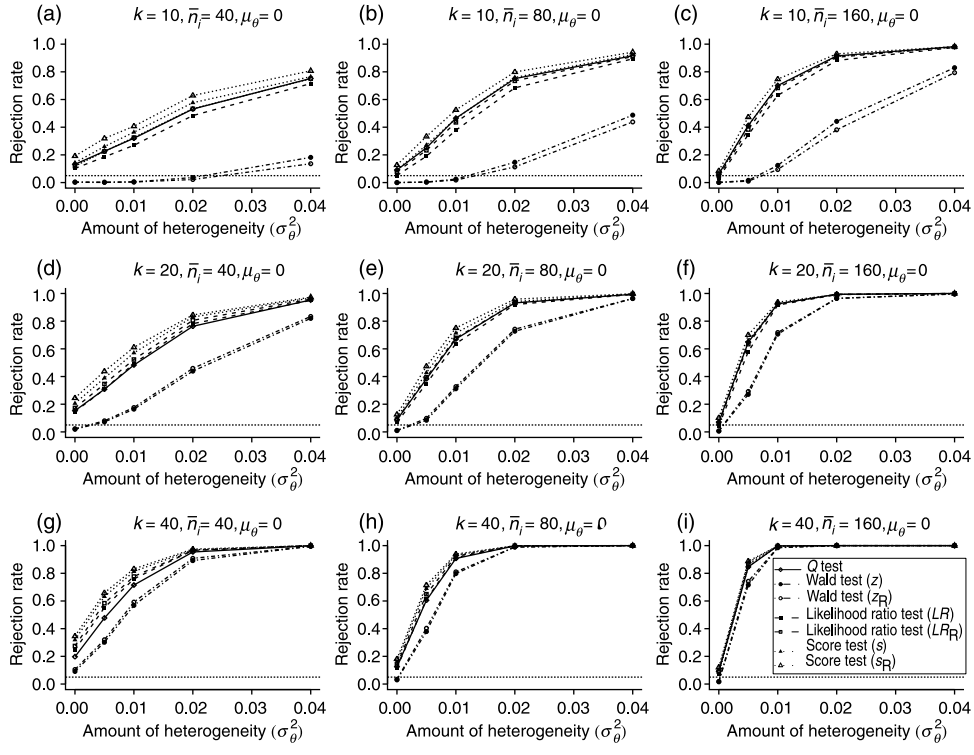

Figure 7. Rejection rates of the homogeneity tests for the correlation coefficient when $\mu_\theta = 0$.

researchers (Hedges & Olkin, 1985; Lipsey & Wilson, 2001; Rosenthal, 1991) have recommended the use of Fisher's variance-stabilizing transformation (Fisher, 1915) before using correlation coefficients as part of a meta-analysis. Specifically, the observed product-moment correlation coefficients are transformed into corresponding z_{r_i}-values with the equation

$$ES_i = z_{r_i} = \frac{1}{2}\ln\left[\frac{1+r_i}{1-r_i}\right]. \tag{27}$$

The asymptotic variance of ES_i is approximately given by $\hat{\sigma}^2_{\varepsilon_i} = 1/(n_i - 3)$. Therefore, as a result of the transformation, the asymptotic sampling variance of the effect size measure no longer depends on ρ_i. In other words, $\sigma^2_{\varepsilon_i}$ is approximately independent of θ_i. Moreover, the distribution of the transformed correlation coefficient is much closer to that of a normal distribution (Hotelling, 1953), although (27) is only asymptotically exactly normally distributed.

Simulations using transformed correlation coefficients as the effect size measure were carried out in the exact same way as described for the raw correlation coefficients, with the only difference being that the r_i-values were first transformed with (27) before applying the homogeneity tests.

7.4.2. Results

Since the Type I error rate of the homogeneity tests no longer depended on μ_θ, all subsequent results were averaged over this factor. Figure 8 shows the Type I error rates

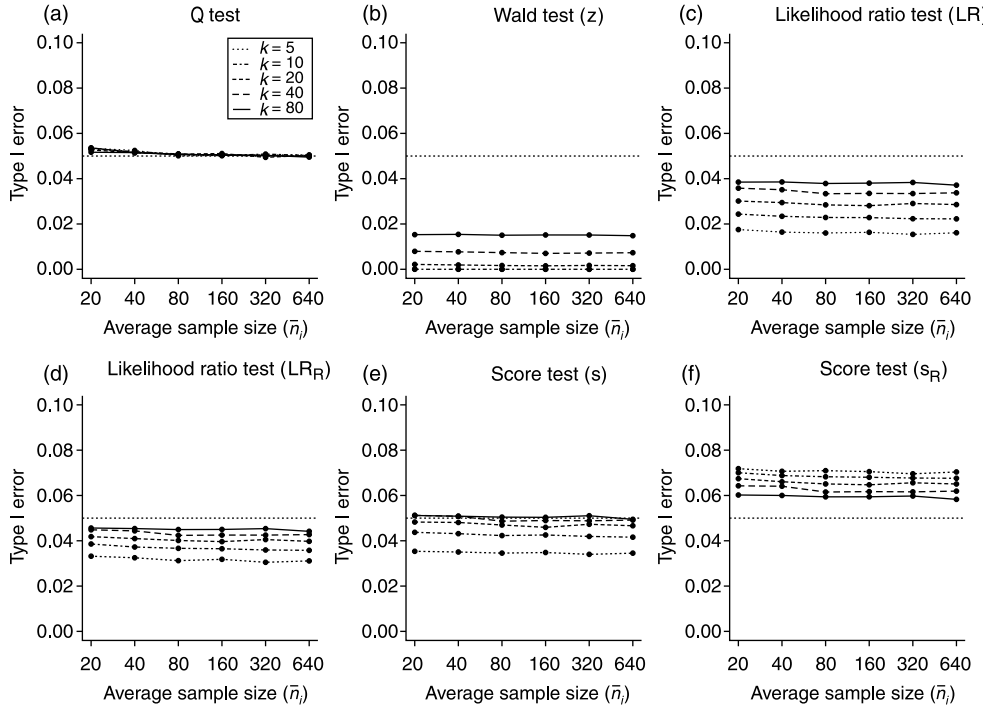

Figure 8. Type I error rates of the homogeneity tests for the correlation coefficient after using Fisher's variance-stabilizing transformation.

for the various homogeneity tests as a function of k and $\bar{n}_i$ (the results for z and z_R were again so similar that only results for the former were plotted). Most notably, the Type I error rate was now essentially unaffected by changes in $\bar{n}_i$.

The Q test controlled the Type I error rate in all of the conditions. On the other hand, the Wald, likelihood ratio, and score tests converged towards the nominal α level only as k increased. Not all of these tests actually reached .05 within the conditions studied, but the trend towards the nominal α level was apparent from the results. Finally, the tests tended to be somewhat conservative for smaller values of k except for the s_R test, which was slightly too liberal in rejecting the null hypothesis.

Figure 9 shows that the probability of rejecting the null hypothesis increased with k, $\bar{n}_i$ and/or σ_θ^2 for all of the tests. Any differences among the rejection rates can be attributed to the fact that some of the tests did not control the Type I error rate adequately when k was small. The exceptions again were the two Wald tests, whose rejection rates were substantially lower than those of the other tests.

7.5. Some general conclusions about the simulations

One particular finding with respect to the Type I error rate of the tests was especially disconcerting. When using raw correlation coefficients as the effect size measure and the average sample size within studies was low, increases in k resulted in progressively higher probabilities of rejecting the null hypothesis when in fact H_0 was true. This finding is consistent with the results of Sánchez-Meca and Marín-Martínez (1997), who also reported inflated Type I error rates when the Q test was applied to raw correlation coefficients under similar conditions.

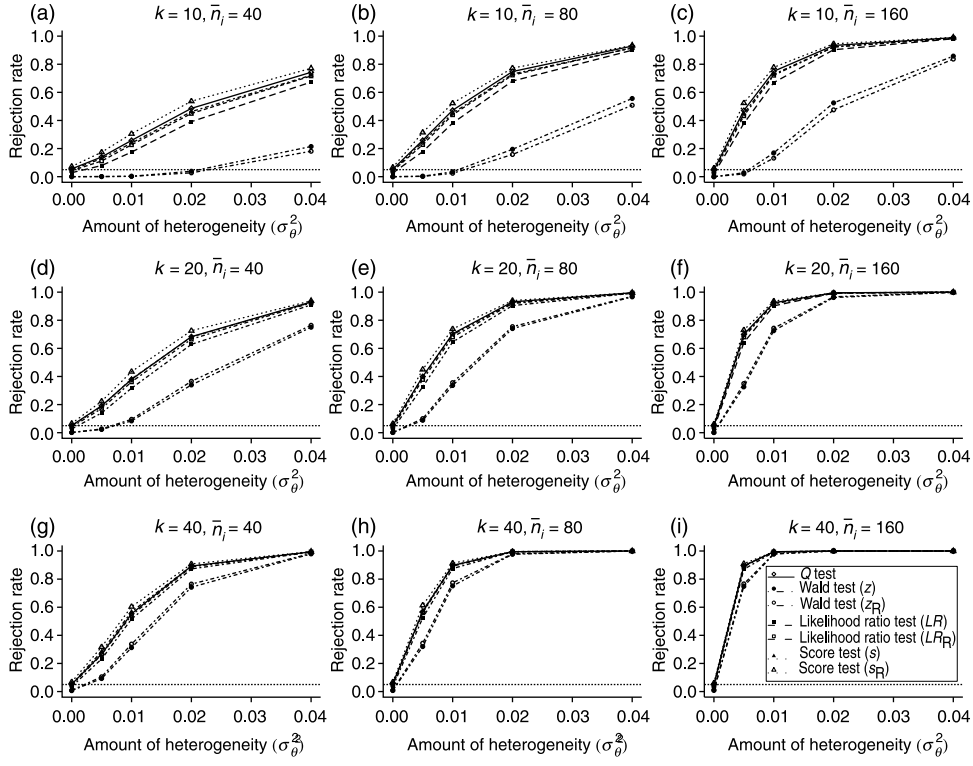

Figure 9. Rejection rates of the homogeneity tests for the correlation coefficient after using Fisher's variance-stabilizing transformation.

While the Type I error rate was actually adequately controlled for certain combinations of $\bar{n}_i$ and k (e.g. the REML-based LR test kept the Type I error rate very close to $\alpha = .05$ when $\bar{n}_i = 80$ and $k = 5$, as can be seen in Figures 5 and 6), the inflation in Type I error rates was so extreme for most conditions that the homogeneity tests are practically meaningless for meta-analyses based on a large number of correlation coefficients derived from studies with small sample sizes. In those cases, one is almost guaranteed to reject the null hypothesis and conclude that the effect sizes are heterogeneous, whether this is true or not.

However, for larger values of $\bar{n}_i$, the tests at least showed behaviour that indicated convergence towards the nominal α level as k increased. Moreover, an obvious remedy for the inflated Type I error rates is the use of the Fisher transformation. The Type I error rate of the homogeneity tests using transformed correlation coefficients was essentially independent of $\bar{n}_i$ and generally well controlled, especially for large k (see Figure 8).

The problem with testing the homogeneity of raw correlation coefficients was previously noted by Hunter and Schmidt (1990) in connection with the Q test. However, instead of suggesting the use of transformed correlation coefficients, Hunter and Schmidt (1990, 1994) proposed a modification to the way we estimate $\sigma^2_{\varepsilon_i}$, namely by using (25), but setting $\theta_i = \bar{r}$, where

$$\bar{r} = \frac{\sum_{i=1}^{k} n_i r_i}{\sum_{i=1}^{k} n_i}. \tag{28}$$

Since H_0 implies $\theta_i = \theta$ for $i = 1, \ldots, k$, it seems reasonable to set θ_i in (25) equal to a more efficient estimate of θ, such as the sample size weighted average of the observed correlation coefficients. Some additional simulations indicated that using this modified estimate of $\sigma^2_{\varepsilon_i}$ does reduce the severe inflation in Type I error rates observed for cases where $\bar{n}_i$ is small, but does not remove it completely, especially when μ_θ is large. Alexander *et al.* (1989) and Field (2001) reported similar findings with respect to the Q test. Therefore, when testing the heterogeneity of correlation coefficients, the use of the Fisher transformation is highly recommended.

The difference in behaviour of the tests for UMDs and for transformed correlation coefficients is also worth noting. Even after applying the Fisher transformation to a set of correlation coefficients, the distribution of z_{r_i} is only exactly normal for large sample sizes. Moreover, the sampling variance of the transformed values is only approximated by $\hat{\sigma}^2_{\varepsilon_i} = 1/(n_i - 3)$, and this approximation again relies on large samples. On the other hand, the UMD effect size satisfies all of the assumptions underlying the tests except known $\sigma^2_{\varepsilon_i}$. Nevertheless, the behaviour of the homogeneity tests for transformed correlation coefficients could be considered closer to optimal, especially for cases where $\bar{n}_i$ is small. The crucial difference lies in the way $\hat{\sigma}^2_{\varepsilon_i}$ is calculated for these two effect size measures. Calculating $\hat{\sigma}^2_{\varepsilon_i}$ for the UMD requires estimating an additional parameter, namely σ^2_i. Due to sampling fluctuations in estimates of σ^2_i (which decrease as $\bar{n}_i$ increases), we are again dependent on large sample sizes. On the other hand, in the case of transformed correlation coefficients, the equation for $\hat{\sigma}^2_{\varepsilon_i}$ does not involve additional unknown parameters, thereby avoiding this additional source of variability. Therefore, while the distribution of z_{r_i} is only asymptotically normal with sampling variance $1/(n_i - 3)$, the approximation appears to be accurate enough for the homogeneity tests even when n_i is quite small.

In general, the Type I error rate was controlled most adequately when using the Q test, especially when avoiding the use of raw correlation coefficients as the effect size measure. As expected, under the null hypothesis, the Q test follows a chi-squared distribution with $k - 1$ degrees of freedom for sufficiently large $\bar{n}_i$-values, even when k is small. Moreover, the Q test enjoys the advantage of being the easiest test to carry out. It does not require an estimate of σ^2_θ and avoids cumbersome computational equations such as those used in the score tests. Finally, the other homogeneity tests did not yield rejection rates over and above those observed for the Q test. Any apparent differences in the rejection rates appear to be attributable to the fact that some of the tests did not control the Type I error rate adequately under certain conditions. For example, the score test based on the s_R statistic has slightly higher rejection rates when k and $\bar{n}_i$ are small, but those are also the conditions where the score test rejected the null hypothesis too often.

The use of the Wald tests should be discouraged. The Wald tests not only were overly conservative in their Type I error rates, but also showed substantially slower gains in their rejection rates as k, $\bar{n}_i$ and/or σ^2_θ increased when compared to the other homogeneity tests. In fact, in some conditions (e.g. $k = 10$ and $\bar{n}_i = 20$), the probability of rejecting the null hypothesis when using the Wald tests was almost unaffected by increases in σ^2_θ (see Figures 2 and 4). Therefore, the fixed-effects model will be adopted too often when relying on the results of the Wald tests, which in turn leads meta-analysts to attribute unwarranted precision to their estimate of the overall effect size and/or to ignore the presence of potential moderators.

On the other hand, the Wald tests have the advantage of being easily converted to confidence intervals for σ^2_θ. If we let $z_{1-\alpha/2}$ denote the 100(1 − α/2)th percentile of a standard normal distribution, then a 100(1 − α)% confidence interval for σ^2_θ is obtained by adding and subtracting $z_{1-\alpha/2}$ times the square root of either (16) or (17) from the ML or REML estimate of σ^2_θ (see Wang & Bushman, 1999, p. 294, for an example of such a confidence interval). In contrast, it is not possible to base confidence intervals directly on the Q or score tests as they do not require explicit estimation of σ^2_θ. Also, the LR tests cannot be simply rearranged to yield confidence intervals and instead require the use of further iterative methods (Hardy & Thompson, 1996). However, while Wald-type confidence intervals for σ^2_θ are relatively easy to obtain, the results in the present paper indicate that such confidence intervals tend to be too wide on average and therefore should be avoided as well.

As a final remark about the two Wald tests, it is interesting to note that their performance was essentially indistinguishable. As mentioned earlier, the MLE of σ^2_θ is negatively biased, while the REML estimator is approximately bias-free (Viechtbauer, 2005). Therefore, ML estimates tend to be too small on average, which would suggest that $z < z_R$. On the other hand, it can be shown that the asymptotic sampling variance of the MLE (16) falls below that of the REML estimator (17) in finite samples (Viechtbauer, 2004). This finding would suggest that $z > z_R$. However, it appears that the negative bias in the MLE is offset by its higher efficiency, resulting in two approximately equivalent tests.

It is not overly surprising that the LR and score tests did not control the Type I error rate quite as well as the Q test. By construction, these tests assume normally distributed effect size estimates and known sampling variances. Large within-study sample sizes ensure that these assumptions are approximately met. However, even when $\bar{n}_i$ is very large, we also need a sufficient number of effect size estimates due to the asymptotic nature of these tests. Therefore, the asymptotic convergence of these tests requires both large k and $\bar{n}_i$.

The present results are consistent with those obtained by Takkouche *et al.* (1999), who also found the *LR* statistic to be somewhat conservative in its Type I error rate when the odds ratio was used as the effect size measure. As Figures 1(c), 3(c) and 8(c) show, this finding held in the present study for the SMD and the transformed correlation coefficient and also for the UMD as long as small within-study sample sizes were not paired with large k.

The current study also indicates that the performance of the LR test can be slightly improved by basing the test on REML estimation. Specifically, the Type I error rate of the LR_R statistic tended to be closer to nominal than when using the *LR* statistic. Takkouche *et al.* (1999) did not examine the REML-based LR test, but a similar finding was reported by Morrell (1998) when testing the significance of variance components in the context of repeated-measures designs.

To put the observed rejection rates in perspective, it is useful to note that population heterogeneity can be a result of either random population effects or the presence of moderators (or a combination of both) and that the rejection rates of the homogeneity tests should be roughly the same regardless of the true source of the heterogeneity as long as the amount of heterogeneity is the same. We can define the amount of heterogeneity in a particular set of effect sizes as

$$\sigma_\theta^2 = \frac{\sum_{i=1}^{k}(\theta_i - \bar{\theta})^2}{k}. \tag{29}$$

Now take, for example, the case where the set of studies can be split into two groups based on some dichotomous moderator variable. For instance, random assignment might have been used in only some of the studies, or two different types of outcome measures employed in the studies. Both of these are typical moderator variables used in meta-analysis that might influence the size of the true effect. Moreover, assume that the population effect size does depend on the level of this moderator variable and that, within each group, the effect sizes are homogeneous. The appropriate model is then given by

$$ES_i = \theta_i + \varepsilon_i, \tag{30}$$

with $\theta_i = \theta_1$ for $i = 1, \ldots, k_1$ and $\theta_i = \theta_2$ for $i = k_1 + 1, \ldots, k$. When the two groups are of equal size and k is even, then (29) is equal to $((\theta_1 - \theta_2)/2)^2$, or in words, the squared mean difference between the two population effect sizes. From this it follows that $|\theta_1 - \theta_2| = 2\sqrt{\sigma_\theta^2}$. Note that we have to attach a slightly different interpretation to the amount of heterogeneity in this scenario. Specifically, σ_θ^2 no longer describes the variance of random population effects, but rather characterizes the degree of departure from homogeneity as a result of fixed differences in the population effect sizes.

The probability of the homogeneity tests detecting the presence of such a moderator can be determined from the results of the present study. For example, Figure 4(e) shows the rejection rates of the various tests as a function of σ_θ^2 for 20 SMDs derived from studies with an average of 40 observations per group. To obtain a rejection rate around .80 with the Q, LR, and score tests, σ_θ^2 must exceed roughly 0.075. This is equivalent to $|\theta_1 - \theta_2| > .55$ under the dichotomous moderator case. A difference of 0.55 is quite substantial for SMDs, when considering that .2, .5, and .8 have conventionally been described as small, medium, and large effect sizes (Cohen, 1988). For raw correlation coefficients, we see based on Figure 7(d) that σ_θ^2 would have to exceed about 0.025 to obtain power values equal to .80 for $k = 20$ and $\bar{n}_i = 40$ (leaving aside the issue that the Type I error rate is somewhat inflated for the Q, LR, and Wald tests). This translates into $|\theta_1 - \theta_2| > 0.32$, which is also a substantial difference considering the conventional definition of .1, .3 and .5 as small, medium, and large effects for this effect size measure (Cohen, 1988). Obviously, for larger values of k and/or $\bar{n}_i$, smaller values of $|\theta_1 - \theta_2|$ are needed in order to obtain adequate rejection rates.

In general, we can estimate the value of σ_θ^2 or $|\theta_1 - \theta_2|$ needed in order to achieve a certain rejection rate based on the simulation results. Table 3 indicates the minimum value of $|\theta_1 - \theta_2|$ that would lead to rejection of the null hypothesis in 80% of the cases for the Q test as a function of k and $\bar{n}_i$ (the entries are approximations based on interpolation).[5] The results for the LR and score tests were very similar to those of the Q test. On the other hand, $|\theta_1 - \theta_2|$ would have to be substantially larger for the Wald tests.

[5] *As discussed earlier, the Type I error rate was not always controlled adequately for some conditions, in particular when $\bar{n}_i$ was small. This explains why the entries for r_i are smaller than those for z_{r_i} in some cases.*

Table 3. Minimum value of $|\theta_1 - \theta_2|$ needed to obtain a rejection rate of at least .80 with the *Q* test[a]

		k				
Effect size	$\bar{n}_i$	5	10	20	40	80
UMD	20				1.71	1.31
	40			1.61	1.25	0.96
	80		1.54	1.16	0.90	0.69
	160	1.68	1.09	0.83	0.66	0.63
	320	1.19	0.78	0.65	0.63	0.63
	640	0.84	0.65	0.63	0.63	0.63
SMD	20				0.60	0.51
	40			0.55	0.42	0.34
	80		0.51	0.38	0.30	0.26
	160	0.54	0.36	0.28	0.20	0.18
	320	0.38	0.26	0.19	0.18	0.18
	640	0.28	0.19	0.18	0.18	0.18
r_i[b]	20			0.359	0.228	0.104
	40			0.317	0.235	0.171
	80		0.332	0.240	0.182	0.139
	160	0.359	0.247	0.176	0.134	0.126
	320	0.271	0.176	0.132	0.126	0.126
	640	0.185	0.134	0.126	0.126	0.126
z_{r_i}	20				0.373	0.307
	40			0.343	0.265	0.214
	80		0.317	0.240	0.185	0.146
	160	0.347	0.224	0.169	0.134	0.126
	320	0.247	0.161	0.131	0.126	0.126
	640	0.173	0.130	0.126	0.126	0.126

Note. UMD = unstandardized mean difference; SMD = standardized mean difference; r_i = raw correlation coefficient; z_{r_i} = correlation coefficient with Fisher transformation.

[a]Empty cells indicate cases where the value of $|\theta_1 - \theta_2|$ needed to obtain rejection rates of at least .80 was larger than those included in the simulations.

[b]For $\mu_\theta = 0$.

8. Conclusion

During the data-analytic step in meta-analysis, the researcher is faced with the difficult problem of having to determine the appropriate model underlying a set of effect size estimates. This process is usually of an exploratory nature and typically commences with an examination of the homogeneity of the effect sizes. Since effect size estimates are subject to sampling error, the goal is to determine whether any additional variability over and beyond that expected solely due to sampling error is present in the data.

Two approaches are commonly used for this purpose. In the first approach, we estimate the amount of population heterogeneity with one of the heterogeneity estimators that have been described in the literature (Viechtbauer, 2005). If the heterogeneity estimate is found to be greater than zero, we reject the fixed-effects model and can either adopt a random-effects model, search for moderators, or use a

combination of these two explanations to account for the heterogeneity (e.g. Berkey *et al.*, 1995; Thompson & Sharp, 1999; van Houwelingen, Arends, & Stijnen, 2002).

Alternatively, in the second approach, we apply a homogeneity test and only adopt a random-effects model or search for moderator variables when the hypothesis of homogeneity is rejected. However, this approach has been criticized by the National Research Council (1992) on the basis that 'the current practice of assuming a fixed effects model . . . [unless] a significance test of the nonhomogeneity of information sources rejects the hypothesis of homogeneity, is inefficient and can lead to understatement of uncertainty about the underlying effect of interest' (p. 186). This assertion was shown to have some merit in particular by the results of Brockwell and Gordon (2001). The alternative seems to be the first approach, which rejects the homogeneity assumption whenever *any* amount of heterogeneity is present in the data.

Heterogeneity estimates greater than zero can indicate either the presence of heterogeneous population effect sizes, the presence of moderators, or a combination of both (Lipsey & Wilson, 2001). However, a fourth explanation for non-zero estimates of σ_θ^2 are random sampling fluctuations. This is what Hunter and Schmidt (1990) called *second-order sampling error*, meaning sampling variability in meta-analytic estimates of the population effect size (i.e. μ_θ or θ) and population heterogeneity (i.e. σ_θ^2). Second-order sampling error can lead to estimates of σ_θ^2 being greater than zero even when the population effect sizes are actually homogeneous.

Consequently, if estimates of σ_θ^2 greater than zero send researchers on an exploratory search for potential moderators, then the danger of committing Type I errors during this process is increased. A homogeneity tests therefore serves a similar purpose as the overall *F*-test in ANOVA or regression, protecting researchers from conducting a large number of exploratory hypothesis tests that can lead to Type I errors or, in the case of meta-analysis, to the discovery of spurious moderators.

However, homogeneity tests are only useful if they provide results that can be trusted. Two criteria by which we may judge the quality of a test are adequate control of the Type I error rate and sufficient sensitivity to detect departures from the null hypothesis under realistic conditions.[6] The present results demonstrate that the Q test adequately controls the Type I error rate for meta-analyses based on studies with at least moderately large sample sizes. Whether the power to detect heterogeneity is sufficient in any particular meta-analysis depends on the number of effect sizes, the sample sizes within the studies, and the actual amount of heterogeneity one would like to detect. Therefore, it is an oversimplification to simply claim that the heterogeneity tests have insufficient power. The present results are useful in pointing out some of the specific conditions under which the results of the homogeneity tests are reliable (i.e. where the Type I error rate is controlled and power is sufficiently large). These results can help meta-analysts decide with how much conviction one can accept the results from a homogeneity test.

[6] *Robustness to violations of the assumption underlying the test would be a third criterion, which is beyond the scope of the present paper, but see Harwell (1997) for some results relevant to this issue.*

References

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*, 329–331.

Berkey, C. S., Hoaglin, D. C., Mosteller, F., & Colditz, G. A. (1995). A random-effects regression model for meta-analysis. *Statistics in Medicine*, *14*, 395–411.

Bond, C. F., Jr, Wiitala, W. L., & Richard, F. D. (2003). Meta-analysis of raw mean differences. *Psychological Methods*, *8*, 406–418.

Brockwell, S. E., & Gordon, I. R. (2001). A comparison of statistical methods for meta-analysis. *Statistics in Medicine*, *20*, 825–840.

Callender, J. C., & Osburn, H. E. (1988). Unbiased estimation of sampling variance of correlations. *Journal of Applied Psychology*, *73*, 312–315.

Cochran, W. G. (1937). Problems arising in the analysis of a series of similar experiments. *Supplement to the Journal of the Royal Statistical Society*, *4*, 102–118.

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

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Corbeil, R. R., & Searle, S. R. (1976a). A comparison of variance component estimation. *Biometrics*, *32*, 779–791.

Corbeil, R. R., & Searle, S. R. (1976b). Restricted maximum likelihood (REML) estimation of variance components in the mixed model. *Technometrics*, *18*, 31–38.

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

Erez, A., Bloom, M. C., & Wells, M. T. (1996). Using random rather than fixed effects models in meta-analysis: Implications for situational specificity and validity generalization. *Personnel Psychology*, *49*, 275–306.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods*, *6*, 161–180.

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*, 507–521.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245–260). New York: Russell Sage Foundation.

Friedman, L. (2000). Estimators of random effects variance components in meta-analysis. *Journal of Educational and Behavioral Statistics*, *25*, 1–12.

Hardy, R. J., & Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, *15*, 619–629.

Hardy, R. J., & Thompson, S. G. (1998). Detecting and describing heterogeneity in meta-analysis. *Statistics in Medicine*, *17*, 841–856.

Hartley, H. O., & Rao, J. N. K. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, *54*, 93–108.

Harville, D. A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, *72*, 320–338.

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

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.

Hedges, L. V. (1982a). Fitting categorical models to effect size from a series of experiments. *Journal of Educational Statistics*, *7*, 119–137.

Hedges, L. V. (1982b). Estimation of effect size from a series of independent experiments. *Psychological Bulletin*, *92*, 490–499.

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*, 388–395.
Hedges, L. V. (1989). An unbiased correction for sampling error in validity generalization studies. *Journal of Applied Psychology*, *74*, 469–477.
Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.
Hedges, L. V., & Pigott, T. D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods*, *6*, 203–217.
Hedges, L. V., & Vevea, J. L. (1998). Fixed-versus random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.
Hotelling, H. (1953). New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society, Series B*, *15*, 193–232.
Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Newbury Park, CA: Sage.
Hunter, J. E., & Schmidt, F. L. (1994). Estimation of sampling error variance in the meta-analysis of correlations: Use of average correlation in the homogeneous case. *Journal of Applied Psychology*, *79*, 171–177.
Hunter, J. E., & Schmidt, F. L. (2000). Fixed effects vs. random effects meta-analysis models: Implications for cumulative research knowledge. *International Journal of Selection and Assessment*, *8*, 275–292.
Johnson, B. T., Mullen, B., & Salas, E. (1995). Comparison of three major meta-analytic approaches. *Journal of Applied Psychology*, *80*, 94–106.
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*, 695–699.
Lehmann, E. L. (1999). *Elements of large-sample theory*. New York: Springer-Verlag.
Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.
Morrell, C. H. (1998). Likelihood ratio testing of variance components in the linear mixed-effects model using restricted maximum likelihood. *Biometrics*, *54*, 1560–1568.
Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and practice (with discussion). *Journal of the American Statistical Association*, *78*, 47–65.
Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, *53*, 17–29.
National Research Council (1992). *Combining information: Statistical issues and opportunities for research*. Washington, DC: National Academic Press.
Neyman, J., & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica*, *16*, 1–32.
Normand, S. T. (1999). Meta-analysis: Formulating, evaluating, combining, and reporting. *Statistics in Medicine*, *18*, 321–359.
Olkin, I., & Pratt, J. W. (1958). Unbiased estimation of certain correlation coefficients. *Annals of Mathematical Statistics*, *29*, 201–211.
Patterson, H. D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*, 545–554.
Patterson, H. D., & Thompson, R. (1974). Maximum likelihood estimation of components of variance. *Proceedings of the 8th International Biometrics Conference*, 197–207.
Rao, C. R. (1973). *Linear statistical inference and its application*. New York: McGraw-Hill.
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*, 683–687.
Rosenthal, R. (1991). *Meta-analytic procedures for social research*. Newbury Park, CA: Sage.
Rosenthal, R. (1994). Parametric measures of effect size. In H. M. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.
Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review in Psychology*, *52*, 59–82.

Rosenthal, R., & Rubin, D. B. (1982). Comparing effect sizes of independent studies. *Psychological Bulletin*, *92*, 500–504.

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

Sagie, A., & Koslowsky, M. (1993). Detecting moderators with meta-analysis: An evaluation and comparison of techniques. *Personnel Psychology*, *46*, 629–640.

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*, 385–399.

Schmidt, F. L., & Hunter, J. E. (1999). Comparison of three meta-analysis methods revisited: An analysis of Johnson, Mullen, and Salas (1995). *Journal of Applied Psychology*, *84*, 144–148.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York: Wiley.

Self, S. G., & Liang, K. Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, *82*, 605–610.

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*, 3–9.

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

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

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine*, *18*, 2693–2708.

van Houwelingen, H. C., Arends, L. R., & Stijnen, T. (2002). Advanced methods in meta-analysis: Multivariate approach and meta-regression. *Statistics in Medicine*, *21*, 589–624.

Verbeke, G., & Molenberghs, G. (1997). *Linear mixed models in practice: A SAS-oriented approach*. New York: Springer-Verlag.

Verbeke, G., & Molenberghs, G. (2003). The use of score tests for inference on variance components. *Biometrics*, *59*, 254–262.

Viana, M. A. G. (1980). Statistical methods for summarizing independent correlational results. *Journal of Educational Statistics*, *5*, 83–104.

Viechtbauer, W. (2004). *Choosing between the fixed-, random-, and mixed-effects model in meta-analysis: An analysis of existing and new model selection methods*. Unpublished doctoral dissertation, University of Illinois, Urbana-Champaign.

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

Wang, M. C., & Bushman, B. J. (1999). *Integrating results through meta-analytic review using SAS software*. Cary, NC: SAS Institute.

Received 22 April 2004; revised version received 4 March 2005