A Q statistic with constant weights for assessing heterogeneity in meta-analysis[†]

Elena Kulinskaya*[1] | David C. Hoaglin[2] | Ilyas Bakbergenuly[1] | Joseph Newman[1]

[1]School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK
[2]Department of Population and Quantitative Health Sciences, University of Massachusetts Medical School, Worcester, Massachusetts 01605, USA

Correspondence
*Elena Kulinskaya,University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK. Email: e.kulinskaya@uea.ac.uk

Summary

The conventional Q statistic, using estimated inverse-variance (IV) weights, underlies a variety of problems in random-effects meta-analysis. In previous work on standardized mean difference and log-odds-ratio, we found superior performance with an estimator of the overall effect whose weights use only group-level sample sizes. The Q statistic with those weights has the form proposed by DerSimonian and Kacker. The distribution of this Q and the Q with IV weights must generally be approximated. We investigate approximations for those distributions, as a basis for testing and estimating the between-study variance (τ^2). A simulation study, with mean difference as the effect measure, provides a framework for assessing accuracy of the approximations, level and power of the tests, and bias in estimating τ^2. Two examples illustrate estimation of τ^2 and the overall mean difference. Use of Q with sample-size-based weights and its exact distribution (available for mean difference and evaluated by Farebrother's algorithm) provides precise levels even for very small and unbalanced sample sizes. The corresponding estimator of τ^2 is almost unbiased for 10 or more small studies.This performance compares favorably with the extremely liberal behavior of the standard tests of heterogeneity and the largely biased estimators based on inverse-variance weights.

KEYWORDS:
inverse-variance weights, effective sample size weights, random effects, mean difference, exact distribution

1 | INTRODUCTION

In meta-analysis, many shortcomings in assessing heterogeneity and estimating an overall effect arise from using weights based on estimated variances without accounting for sampling variation. Our studies of methods for random-effects meta-analysis of standardized mean difference[1] and log-odds-ratio[2] included an estimator of the overall effect that combines the studies' estimates with weights based only on their groups' sample sizes. That estimator, SSW, outperformed estimators that use (estimated) inverse-variance-based (IV) weights. Those weights use estimates of the between-study variance (τ^2) derived from the popular Q statistic discussed by Cochran[3], which uses inverse-variance weights and which we refer to as Q_{IV}. Thus, parallel to SSW, we investigate an alternative, Q_{SW}, in which the studies' weights are their effective sample sizes. This Q_{SW} is an instance of the generalized Q statistic Q_F introduced by DerSimonian and Kacker[4], in which the weights are fixed positive constants.

[†]This is an example for title footnote.

We consider the following random-effects model (REM): For Study i ($i = 1, \ldots, K$), with sample size $n_i = n_{iT} + n_{iC}$, the estimate of the effect is $\hat{\theta}_i \sim G(\theta_i, v_i^2)$, where the effect-measure-specific distribution G has mean θ_i and variance v_i^2, and $\theta_i \sim N(\theta, \tau^2)$. Thus, the $\hat{\theta}_i$ are unbiased estimators of the true conditional effects θ_i, and the $v_i^2 = \mathrm{Var}(\hat{\theta}_i|\theta_i)$ are the true conditional variances.

The general Q statistic is a weighted sum of squared deviations of the estimated effects $\hat{\theta}_i$ from their weighted mean $\bar{\theta}_w = \sum w_i \hat{\theta}_i / \sum w_i$:

$$Q = \sum w_i (\hat{\theta}_i - \bar{\theta}_w)^2. \tag{1}$$

In Cochran[3] w_i is the reciprocal of the *estimated* variance of $\hat{\theta}_i$, resulting in Q_{IV}. In meta-analysis those w_i come from the fixed-effect model. In what follows, we discuss approximations to the distribution of Q_F and estimation of τ^2 when the w_i are arbitrary positive constants. Because it is most tractable, but still instructive, we focus on a single measure of effect, the mean difference (MD). In this favorable situation, the cumulative distribution function of Q_F can be evaluated by the algorithm of Farebrother[5]. We also consider approximations that match the first two or the first three moments of Q_F. In simulations and examples, we concentrate on Q_{SW}. For comparison we also include some of the popular inverse-variance-based methods of estimating τ^2, approximating the distribution of Q_{IV}, and testing for the presence of heterogeneity. A simulation study provides a framework for assessing accuracy of the approximations, level and power of the tests based on Q_{SW} and Q_{IV}, and bias in estimating τ^2.

2 | EXPECTED VALUE OF Q_F AND ESTIMATION OF τ^2

Define $W = \sum w_i$, $q_i = w_i/W$, and $\Theta_i = \hat{\theta}_i - \theta$. In this notation, and expanding $\bar{\theta}_w$, Equation (1) can be written as

$$Q = W \left[\sum q_i(1 - q_i)\Theta_i^2 - \sum_{i \neq j} q_i q_j \Theta_i \Theta_j \right]. \tag{2}$$

Under the above REM, and assuming that the w_i are arbitrary fixed constants, it is straightforward to obtain the first moment of Q_F as

$$\mathrm{E}(Q_F) = W \sum q_i(1 - q_i)\mathrm{Var}(\Theta_i) = W \sum q_i(1 - q_i)(\mathrm{E}(v_i^2) + \tau^2). \tag{3}$$

This expression is similar to Equation (4) in DerSimonian and Kacker[4]. Rearranging the terms gives the moment-based estimator of τ^2

$$\hat{\tau}_M^2 = \max\Big(\frac{Q_F/W - \sum q_i(1 - q_i)\hat{\mathrm{E}}(v_i^2)}{\sum q_i(1 - q_i)}, \, 0 \Big). \tag{4}$$

This equation is similar to Equation (6) inDerSimonian and Kacker[4]; they use the within-study (i.e., conditional) estimate s_i^2 instead of $\hat{\mathrm{E}}(v_i^2)$, an important distinction because v_i^2 is a random variable whose distribution depends on that of θ_i.

3 | APPROXIMATIONS TO THE DISTRIBUTION OF Q_F

For approximations to the distribution of Q_F, we draw on results for quadratic forms, which generalize the sums of squares that arise in analysis of variance. The Q statistic, Equation (2), can be expressed as a quadratic form in the random variables Θ_i. Appendix A.1 gives the details and discusses approaches for evaluating and approximating distributions of quadratic forms in normal variables. Conveniently, the variablesΘ_i for the mean difference (MD) are normal.

Two approaches are most suitable, especially for obtaining upper-tail probabilities, $P(Q_F > x)$. One matches moments of Q_F, either the first two or the first three moments; Appendix A.2 gives the details. The other uses an algorithm developed by Farebrother[5].

4 | SIMULATION STUDY FOR MEAN DIFFERENCE

For MD as the effect measure, we use simulation of the distribution of Q with constant effective-sample-size weights (SW) $\tilde{n}_i = n_{iC} n_{iT}/(n_{iC} + n_{iT})$ to study three approximations: the Farebrother approximation (F SW), implemented in the R package *CompQuadForm*[6]; the two-moment Welch-Satterthwaite approximation (M2 SW); and the three-moment chi-square approximation (M3 SW) by Solomon and Stephens[7]. Details of these two moment-based approximations are given in Appendix A.2. We also study the bias of the moment estimator $\hat{\tau}_M^2$ in Equation (4), denoted by SDL, for this choice of constant weights.

For comparison, we also simulate Q with IV weights, and study three approximations to its distribution: the standard chi-square approximation, the approximation based on the Welch test to the null distribution of $Q_{+}IV$, and the "exact" distribution of Biggerstaff and Jackson[8] (BJ) when $\tau^2 > 0$. To compare the bias of SDL with that of estimators of τ^2 that use the IV weights, we also consider DerSimonian and Laird[9] (DL), Mandel and Paule[10] (MP), REML, and a corrected DL estimator[1] (CDL), which uses an improved non-null first moment of Q_{IV}. Table 1 lists abbreviations for all methods used in our simulations.

We varied five parameters: the number of studies K, the total (average) sample size of each study n (or $\bar{n}$), the proportion of observations in the Control arm f, the between-study variance τ^2, and the within-study variance σ_T^2 (keeping $\sigma_C^2 = 1$). We set the overall true MD $\mu = 0$ because the estimators of τ^2 do not involve μ and the estimators of μ are equivariant.

We generate the within-study sample variances s_{ij}^2 $(j = T,\ C)$ from chi-square distributions $\sigma_{ij}^2\chi_{n_{ij}-1}^2/(n_{ij}-1)$ and the estimated mean differences y_i from a normal distribution with mean 0 and variance $\sigma_{iT}^2/n_{iT} + \sigma_{iC}^2/n_{iC} + \tau^2$. We obtain the estimated within-study variances as $\hat{v}_i^2 = s_{iT}^2/n_{iT} + s_{iC}^2/n_{iC}$. As would be required in practice, all approximations use these $\hat{v}_i^2$, even though the $\sigma_{iT}^2/n_{iT} + \sigma_{iC}^2/n_{iC}$ are available in the simulation.

All simulations use the same numbers of studies $K = 5,\ 10,\ 30$ and, for each combination of parameters, the same vector of total sample sizes $n = (n_1, \ldots, n_K)$ and the same proportions of observations in the Control arm $f_i = .5,\ .75$ for all i. The sample sizes in the Treatment and Control arms are $n_{iT} = \lceil (1 - f_i)n_i \rceil$ and $n_{iC} = n_i - n_{iT}$, $i = 1, \ldots, K$. The values of f reflect two situations for the two arms of each study: approximately equal (1:1) and quite unbalanced (1:3).

We study equal and unequal study sizes. For equal study sizes n_i is as small as 20, and for unequal study sizes average sample size $\bar{n}$ is as small as 13 (individual n_i are as small as 4), in order to examine how the methods perform for the extremely small sample sizes that arise in some areas of application. In choosing unequal study sizes, we follow a suggestion of Sánchez-Meca and Marín-Martínez[11]. Table 2 gives the details.

TABLE 1 *Abbreviations*

Weights	IV F SSW	Inverse-variance weights $w_i = 1/v_i^2$ arbitrary constant weights effective sample size weights $\tilde{n} = n_C n_T/n$
Approximations to distribution of Q_{SW} and Q_{IV}	F SW M2 SW M3 SW BJ	Farebrother approximation two-moment approximation three-moment approximation Biggerstaff and Jackson[8]
Estimators of τ^2	DL CDL SDL REML MP	DerSimonian-Laird[9] Corrected DerSimonian-Laird[1] new moment estimator based on Q_{SW} restricted maximum-likelihood estimator Mandel-Paule[10]

TABLE 2 *Data patterns in the simulations*

Parameter	Equal study sizes	Unequal study sizes
K (number of studies)	5, 10, 30	5, 10, 30
n or $\bar{n}$ (average size of individual study — total of the two arms) For $K = 10$ and $K = 30$, the same set of unequal study sizes is used twice or six times, respectively.	20, 40, 100, 250	13 (4, 6, 7, 8, 40), 15 (6, 8, 9, 10, 42), 30 (12, 16, 18, 20, 84), 60 (24, 32, 36, 40, 168)
f (proportion in the control arm)	1/2, 3/4	1/2, 3/4
μ	0	0
σ_C^2, σ_T^2 (within-study variances)	(1,1), (1,2)	(1,1), (1,2)
τ^2 (variance of random effects)	0(0.1)1	0(0.1)1

We use a total of 10,000 replications for each combination of parameters. Thus, the simulation standard error for an empirical p-value $\hat{p}$ under the null is roughly $\sqrt{1/(12 * 10,000)} = 0.0029$. The simulations were programmed in R version 3.6.2 using the University of East Anglia 140-computer-node High Performance Computing (HPC) Cluster, providing a total of 2560 CPU cores, including parallel processing and large memory resources. For each configuration, we divided the 10,000 replications into 10 parallel sets of 1000.

5 | RESULTS

For each configuration of parameters in the simulation study and for each approximation, we calculated, for each generated value of Q, the probability of a larger Q: $\tilde{p} = 1 - \hat{F}(Q)$ ($\hat{F}$ denotes the distribution function of the approximation). We recorded empirical p-values $\hat{p} = \#(\tilde{p} < p)/10000$ at $p = .001, .0025, .005, .01, .025, .05, .1, .25, .5$ and the complementary values $.75, \ldots, .99$. The values of τ^2 included both null ($\tau^2 = 0$) and non-null ($\tau^2 > 0$) values (Table 2). The approximations to the non-null distribution of Q were based on the value of τ^2 used in the simulation. These data provide the basis for P–P plots (versus the true null distribution) for three approximations to the distribution of Q with effective-sample-size weights (F SW, M2 SW, and M3 SW) and two approximations to the distribution of Q with IV weights (chi-square/BJ and Welch) and for estimating their null levels, non-null empirical tail areas, and (roughly) their power. We also estimate the bias of five point estimators of τ^2 (SDL, DL, REML, MP, and CDL). In the Figures 1-5, we tried to present configurations that illustrate the differences in methods very clearly. The full results are presented, graphically, in Appendix B of Kulinskaya et al.[12].

In some instances M3 SW produced anomalous results or no results at all (because numerical problems kept us from obtaining estimates of its parameters).

5.1 | P–P plots

To compare an approximation for a distribution function of Q against the theoretical distribution function, with no heterogeneity ($\tau^2 = 0$), we use probability–probability (P–P) plots[13]. Evaluating two distribution functions, F_1 and F_2, at x yields $p_1 = F_1(x)$ and $p_2 = F_2(x)$. One varies x, either continuously or at selected values, and plots the points $(p_1(x), p_2(x))$ to produce the usual P–P plot of F_2 versus F_1. If $F_2 = F_1$, the points lie on the line from (0, 0) to (1, 1). If smaller x are more likely under F_2, the points will lie above the line, and conversely. (Working with upper tail areas reverses these interpretations.) If F_2 is similar to F_1, the points will lie close to the line, and departures will show areas of difference. To make these more visible, we flatten the plot by subtracting the line; that is, we plot $p_2 - p_1$ versus p_1.

The simulations offer a shortcut that does not require evaluating the true distribution function of Q (which is unknown for IV weights). If F is the distribution of the random variable X, $F(X)$ has the uniform distribution on [0, 1], and so does $1 - F(X)$. Thus, for the values of p listed above, we plot $\hat{p} - p$ versus p.

Our P–P plots (illustrated by Figure 1) show no differences between the M3 and M2 approximations for Q with constant weights. Very minor differences between the Farebrother and the moment approximations are visible, mainly at very small sample sizes. Other comparisons show three distinct patterns.

The chi-square approximation has strikingly higher empirical tail areas than the true distribution of Q with IV weights over the whole domain. This pattern is especially noticeable for $K = 30$ and small unequal sample sizes, though it persists for equal sample sizes as large as 100. It indicates that the approximating chi-square distribution produces values that are systematically too large,

The Welch test provides a much better fit that is especially good for balanced sample sizes, equal variances, and small K. When sample sizes are small and vary among studies or are unbalanced between arms, however, its fit is worse. It produces values of Q that are systematically too small when $K = 5$; produces more small values and, to a lesser extent, more large values when $K = 10$; and produces more large values and, to a lesser extent, more small values when $K = 30$.

The three approximations to Q with constant weights provide reasonably good fits, which appear to be similar to the fit of the Welch test to Q with IV weights.

5.2 | Empirical levels when $\tau^2 = 0$

To better visualize the quality of the approximations as the basis for a test for heterogeneity at the .05 level, we plot their empirical levels under the null $\tau^2 = 0$ versus sample size. Figure 2 presents typical results for a range of sample sizes at the .05 level.

For equal variances, the empirical levels depend on the sample size. The chi-square test is very liberal up to $n = 100$, especially for unbalanced arms, and the problem becomes worse as K increases. The Welch test is considerably better than the chi-square test, but is still noticeably liberal when the arms are unbalanced. Tests based on Q with constant weights are generally less liberal, though they may have level up to .07 for $n = 20$, for unbalanced arms and small K. The M3 approximation breaks down and results in very liberal levels for unequal sample sizes and unbalanced arms and large K. The Farebrother and M2 approximations perform better for larger K, and overall are the best choice. They also hold the level well at smaller nominal levels. The Welch test is rather unstable for very low levels such as $\alpha = .001$ (which corresponds, in our simulations, to just 10 occurrences in 10,000 replications), but improves from $\alpha = .005$.

5.3 | Empirical levels when $\tau^2 > 0$

To understand how the approximations behave as τ^2 increases, we plot the empirical p-values ($\hat{p}$) vs τ^2 for the nominal levels .05 and .01 (Figure 3). For unequal sample sizes, the Farebrother and the 3-moment approximations differ slightly at the .01 level, but those differences disappear at the .05 level and for equal sample sizes. When $K = 30$, M3 sometimes fails; and when it does not, it breaks down for small and large values of τ^2. The 2-moment approximation is almost indistinguishable from the Farebrother approximation.

Overall, the Farebrother approximation performs superbly across all τ^2 values. This is as it should be, as it is practically an exact distribution in the case of MD. The M2 approximation is reasonably good at the .05 level. The BJ approximation is much too liberal, especially at smaller values of τ^2 and for larger K. It is considerably more liberal for very small sample sizes such as $\bar{n} = 13$; but it improves when sample sizes increase, and it is reasonable by $n = 100$ or $\bar{n} = 60$.

For larger values of n and $\bar{n}$ (not shown in Figure 3), the traces approach α as n or $\bar{n}$ increases (they are farther away from α when $\bar{n} < 30$).

5.4 | Power of tests for heterogeneity

"Power" is a reasonable term as a heading, but not as an accurate description for most of the results. Although discussions of simulation results in meta-analysis do not always do this, comparisons of power among tests that are intended to have a specified level (i.e., rate of Type I error) are not valid unless the tests' estimated levels are equal or nearly so. This complication is evident in Figure 4, which depicts the power of tests of heterogeneity at the .05 level for $n = 20$ and equal and unequal sample sizes.

The chi-square test appears to be more powerful, and the Welch test slightly less powerful, than the tests based on Q with constant weights. These differences are much smaller when $n = 40$ (not shown) and disappear when n is larger. But even for $n = 20$, these appearances are misleading. For $n = 20$, Figure 2 shows that for balanced arms, the level of the chi-square test is .08 for $K = 5$, .1 for $K = 10$, and considerably higher than .1 for $K = 30$. For unbalanced arms, the level of the chi-square test substantially exceeds .1 for all K. This behavior is a consequence of using an incorrect null distribution. Thus, our results do not show that the chi-square test has higher power, and its power may actually be lower. It is not clear how to modify the chi-square test so that it has the correct level in a broad range of situations.

The Welch test has levels similar to those of the tests based on Q with constant weights when $K = 5$ or 10. But for $K = 30$ and $f = .75$, its level is approximately .09. This may mean that it does have somewhat lower power.

When $n = 40$, the traces rise more steeply, and when $\bar{n} < 30$, they spread out and rise less steeply. When $n \geq 100$ (or $\bar{n} \geq 60$ for unequal sample sizes), visible differences among the traces for the tests disappear. Given higher levels of the chi-square test, this means that its power is the same or even lower than that of the tests based on Q with constant weights.

5.5 | Bias in estimation of τ^2

Here we compare the SDL estimator of τ^2 with the well-known estimators DL, MP, and REML and the recently suggested CDL. Figure 5 depicts the biases of the five estimators for small sample sizes.

All five estimators have positive bias at $\tau^2 = 0$, because of truncation at zero. The bias across all values of τ^2 is quite substantial, and it increases for unequal variances and/or sample sizes. Among the standard estimators, DL has the most bias and MP the least. SDL and CDL generally have similar bias, considerably less than the standard estimators. The relation of their bias to K when $\bar{n} = 13$ is interesting, but atypical. As K increases, the trace for SDL flattens toward 0, demonstrating no bias at all for larger values of τ^2, whereas the trace for CDL rises toward the other three. The traces flatten and approach 0 as $\bar{n}$ increases to 15 and 30. When $n \geq 100$ or $\bar{n} \geq 60$, the differences among the five estimators of τ^2 are quite small.

6 | EXAMPLES

6.1 | Exercise training in people with heart failure

The systematic review of Rees et al.[14] studied results of short-term trials of exercise training in people with mild to moderate heart failure. Exercise capacity was assessed by the maximal oxygen uptake, VO_2max; an increase from baseline to follow-up indicates improvement with exercise. However, for the pooled analysis, the authors reversed the sign of the mean change in VO_2max for both the intervention and control groups, so the beneficial effect is negative. We consider the results from Comparison 2.1.7, for the $K = 15$ studies with mean age above 55 years. Figure 6 shows the data and forest plot. The sample sizes in these trials are rather small, varying from 7 to 48 per arm; the average sample size is 18.4 in the treatment arm and 17.6 in the control arm. The trials are mostly balanced, with only one trial having a 2:1 allocation ratio, and they have similar variances in the two arms.

The review used a DL-based analysis and found significant heterogeneity ($p = .03$), $I^2 = 45.73\%$, $\hat{\tau}^2 = 0.79$, and a significant effect of exercise, with a mean difference in VO_2max of -1.77 (-2.50, -1.03).

Table 3 brings together meta-analyses of these data by seven methods. When testing heterogeneity, the standard chi-square test gives p-value .027, and the Welch test gives .030, indicating significant heterogeneity. These differ substantially from the p-values for Q_{SW} with constant weights, where all three approximations give .43 or .44. This agrees with our simulation results, illustrating how liberal the standard heterogeneity tests can be in the case of small sample sizes and medium to large K.

Comparing the estimated τ^2 values, the DL method provides an estimate of 0.791. CDL is very similar at 0.783, the REML estimate is lower at 0.652, and the MP estimate is considerably lower at 0.255. Unsurprisingly, the SDL estimate is very close to zero, at 0.009. Because the standard estimators are all positively biased in this setting, we consider the SDL estimate to be the closest to the true value of τ^2.

These differences in the estimated heterogeneity variance have no substantial impact on the estimated overall effect of exercise on VO_2max. Table 3 includes IV estimates of Δ with 95% confidence intervals. Because of IV weighting, the smaller τ^2 values result in stronger effects of exercise. SDL results in the most pronounced effect, -2.14 (-2.20, -1.68). However, we do not recommend IV weights for pooling effects, and instead advocate effective-sample-size-based methods[1]. These weights are denoted by SSW in Table 3, and the corresponding confidence intervals are based on t_{K-1} critical values. Ironically, for these data the result, -1.78 (-2.37, -1.18), is very close to the original estimate reported in Rees et al.[14].

The differences between SDL and other estimators of τ^2 are rather striking. However, they have a simple explanation. The largest study, by Bellardinelli et al. (1999), the first on the forest plot in Figure 6, is a low outlier with $\hat{\Delta}_1 = -3.20$, and its inverse-variance weight, when $\tau^2 = 0$, is 39.3%. This study is the major contributor to the high value of $Q_{IV} = 25.79$ and the only reason for the seemingly high heterogeneity. The SSW weight of this study is less than half as large, at 17.5%, and the test based on Q_{SW} does not find heterogeneity in the data. Setting this study aside decreases the Q_{IV} statistic to 7.41 on 14 d.f.; the p-values for all Q tests are very similar, at .88 for all IV tests and at .87 for the tests based on Q_F; and all estimators of τ^2 agree on $\hat{\tau}^2 = 0$.

6.2 | Drugs for prevention of exercise-induced asthma

The systematic review of Spooner et al.[15] compared several types of drugs for prevention of exercise-induced asthma attacks in asthma sufferers. We consider Comparison 6.2.2, which compared inhaling a single dose of mast cell stabilizer (MCS) prior to strenuous exercise with a single dose of short-acting beta-agonists (SABA). The measure of effect was the maximum percentage decrease in pulmonary function (PFT). This meta-analysis pooled results from seven high-quality clinical trials involving a total of 187 patients. Figure 7 shows the data and forest plot. The sample sizes in these perfectly balanced trials vary from 8 to 20 per arm; the average sample size is 13.4 in each arm; the variances mostly differ in the two arms, but without any clear pattern. The review used a DL-based analysis and found that heterogeneity was not significant, $\hat{\tau}^2_{DL} = 0.65$ and $I^2 = 2.14\%$, and that SABA provided significantly lower PFT, $\hat{\Delta} = 6.32$ (2.47,10.18).

Method	$\hat{\tau}^2$	$\hat{\Delta}$	Lower	Upper
SDL SSW t	0.0088	−1.7761	−2.3703	−1.1820
SDL IV	0.0088	−2.1404	−2.6024	−1.6785
DL IV	0.7907	−1.7656	−2.4968	−1.0345
REML IV	0.6524	−1.7784	−2.4779	−1.0790
MP IV	0.2554	−1.8696	−2.4550	−1.2842
CDL SSW t	0.7826	−1.7761	−2.6039	−0.9484
CDL IV	0.7826	−1.7663	−2.4957	−1.0369

TABLE 3 Meta-analyses of the Rees (2004) data on exercise-related changes in VO_2max in people with mild to moderate heart failure.

Method	$\hat{\tau}^2$	$\hat{\Delta}$	Lower	Upper
SDL SSW t	0.0000	9.3002	3.1817	15.4187
SDL IV	0.0000	6.1874	2.4232	9.9516
DL IV	0.6684	6.3223	2.4673	10.1774
REML IV	9.8200	7.3904	2.6364	12.1444
MP IV	0.3386	6.2574	2.4464	10.0684
CDL SSW t	0.6576	9.3002	3.1332	15.4672
CDL IV	0.6576	6.3203	2.4666	10.1740

TABLE 4 Meta-analyses of the Spooner et al. (2003) data on drugs for prevention of exercise-induced asthma attacks.

Table 4 shows the results of meta-analyses of these data by seven methods. Heterogeneity is not significant by any method: the p-values are .409 for the chi-square and Welch tests and .799 to .812 for all three approximations to the distribution of Q_{SW}. However, the estimated values of τ^2 vary widely: 0 for SDL, 0.34 for MP, 0.66 and 0.67 for CDL and DL, and 9.82 for REML. These results agree with the positive biases in estimation of τ^2 at zero in our simulation results, though the result for REML is quite aberrant. Its value is not so extreme, at 5.72, but the maximum-likelihood estimator of τ^2 behaves similarly. The presence of a study with a noticeably lower $\hat{\Delta}_i$ whose estimated variance is substantially lower strains the assumption that $\Delta_i \sim N(\Delta, \tau^2)$.

These differences in the estimated values of τ^2 are reflected in the width of confidence intervals for the pooled effect, but even more so, in the width of prediction intervals[16]. As SDL is zero, the prediction interval is not different from the confidence interval, MP IV has a 95% prediction interval of (1.04, 11.47), DL IV a somewhat wider prediction interval of (0.85, 11.80), and REML IV a much wider interval of (−2.80, 17.58). Thus, REML IV analysis does not find SABA drugs to be more beneficial than MCS. This conclusion does not change if REML IV is used in combination with the Harting-Knapp-Sidik-Johnson variance[17,18], as recommended in Partlett and Riley[19], resulting in a slightly tighter prediction interval of (−2.09, 16.87).

In the forest plot (Figure 7), the study by Vazquez 1984 has a considerably lower mean than the other studies; and, because of its lower variance, its weight varies from 33.3% in the REML IV analysis to 45.5% in the DL IV analysis, in comparison to 13% in SSW. As a result, all the IV-weighted methods yield substantially lower estimates of the pooled effect (6.19 to 7.39) than SSW (9.30). Once more, for these data, the sample-size-based weights provide more robust and more sensible inference than the IV-weighted methods.

7 | DISCUSSION

As a way of avoiding the shortcomings associated with the customary Q, which uses inverse-variance weights based on estimated variances, we are involved in studying a version of Q in which the weights are fixed constants. Such weights simplify derivation of higher moments of Q and facilitate approximation of its distribution.

In a simulation study we compared the properties of the test for heterogeneity for MD based on a Q statistic that uses constant sample-size-based weights, Q_{SW}, with its IV-weights-based counterparts. From Q_{SW} we also derived an estimator (SDL) of the heterogeneity variance τ^2; the simulation yielded estimates of its bias and comparisons with the bias of several other estimators.

A large number of small studies is the worst-case scenario for the statistical properties of meta-analysis[1]. This situation may not be very widespread in medical meta-analyses, but it is very common in the social sciences and in ecology[20,21]. Thus, our simulations included additional small sample sizes.

Overall, the proposed test for heterogeneity for MD, combined with its exact distribution as obtained by the Farebrother algorithm[5] or, alternatively, with the two-moment approximation, provides very precise control of the significance level, even when sample sizes are small and unbalanced, in contrast to the extremely liberal behavior of the standard tests, especially for a large number of studies. (These results suggest that the null distribution of Q_{IV} is more difficult to approximate than the null distribution of Q_F.) Similarly, the proposed SDL estimator is almost unbiased for $K \geq 10$, even in the case of extremely small sample sizes, and we recommend its exclusive use in practice.

Further, because it uses an incorrect null distribution for Q_{IV}, the chi-square test generally has level much greater than .05, so our simulations could give only substantially inflated estimates of its power. An important conclusion of our work is that the power of the popular Q test is even lower than generally believed. As another consequence of the incorrect null distribution, we avoid I^2 and related measures of heterogeneity.

Our meta-analyses of the data from Rees et al.[14] demonstrated just how liberal the standard tests for heterogeneity are. However, the substantial differences among the estimates of τ^2 produced only modest differences among the estimates and confidence intervals for the overall effect. On the other hand, the example illustrated how easily a single discrepant study could distort the IV-weighted estimates of τ^2.

In a second example none of the methods found significant heterogeneity. The SDL estimate was $\hat{\tau}^2 = 0$, whereas the IV-weighted methods produced substantial positive estimates, consistent with the biases that we found in our simulations. In this instance the SSW estimate of the overall effect was noticeably higher than the IV-weighted estimates.

It is enlightening to observe that, for the non-null distribution of Q_{IV}, the approximation of Biggerstaff and Jackson[8] (using Farebrother's algorithm) is no better than the standard chi-square approximation to the null distribution. The problem here evidently lies with the IV weights.

We found that, even though both moment approximations performed well overall, the three-moment approximation sometimes fails, and it breaks down in the case of very small and unbalanced sample sizes and a large number of studies. Therefore, for MD we recommend the Farebrother[5] approximation to the distribution of Q with constant weights.

In further work we intend to develop tests for heterogeneity in other effect measures based on Q with constant weights. Even though we derived general expressions for moments of Q, application of these expressions to such effect measures as SMD and the log-odds-ratio involves a lot of tedious algebra. The moment approximations are less precise than the exact distribution or the approximation by Farebrother[5] for the case of normal variables in the quadratic form, but they are much faster and may be a better option when the distribution is only asymptotically normal.

HIGHLIGHTS

What is already known?

- The conventional Q statistic in meta-analysis underlies the usual test for heterogeneity, but that test produces p-values that are too high for small to medium sample sizes.

- The use of inverse-variance weights based on estimated variances makes it very difficult to approximate the distribution of Q, which varies depending on an effect measure.

- Related moment-based estimators of the heterogeneity variance (τ^2), such as the DerSimonian-Laird estimator, have considerable bias.

What is new?

- We introduce a new Q statistic with constant weights based on studies' effective sample sizes. Its null distribution is calculated exactly by the Farebrother algorithm; alternatively, a two-moment approximation can be used. Both provide very precise control of the significance level, even when sample sizes are small and unbalanced.

- The new Q statistic yields a new estimator of the heterogeneity variance. This estimator, SDL, is almost unbiased for 10 or more studies, even with extremely small sample sizes.

Potential impact for RSM readers outside the authors' field

- The usual chi-square test of heterogeneity generally has level much greater than .05, and its power is even lower than generally believed, because it uses an incorrect null distribution for Q.
- Our new Q statistic, with constant weights, results in a very precise test, and the related new estimate of τ^2 is almost unbiased. We recommend its exclusive use in practice.

FUNDING

The work by E. Kulinskaya was supported by the Economic and Social Research Council [grant number ES/L011859/1].

DATA AVAILABILITY STATEMENT

Our full simulation results are available as an e-print (Bakbergenuly et al.[12]). R procedures that implement the Q test for heterogeneity in meta-analysis of MD, the SDL and the CDL estimators of τ^2 are available in the file `fixedQmethodsForMD.r`

References

1. Bakbergenuly I, Hoaglin DC, Kulinskaya E. Estimation in meta-analyses of mean difference and standardized mean difference. *Statistics in Medicine* 2020; 39(2): 171–191.
2. Bakbergenuly I, Hoaglin DC, Kulinskaya E. Methods for estimating between-study variance and overall effect in meta-analyses of odds-ratios. *Research Synthesis Methods* 2020; 11: 426–442. doi: 10.1002/jrsm.1404
3. Cochran WG. The combination of estimates from different experiments. *Biometrics* 1954; 10(1): 101–129.
4. DerSimonian R, Kacker R. Random-effects model for meta-analysis of clinical trials: an update. *Contemporary Clinical Trials* 2007; 28(2): 105–114.
5. Farebrother RW. Algorithm AS 204: The distribution of a positive linear combination of χ^2 random variables. *Journal of the Royal Statistical Society, Series C* 1984; 33(3): 332–339.
6. Duchesne P, Lafaye De Micheaux P. Computing the distribution of quadratic forms: Further comparisons between the Liu–Tang–Zhang approximation and exact methods. *Computational Statistics & Data Analysis* 2010; 54(4): 858–862.
7. Solomon H, Stephens MA. Distribution of a sum of weighted chi-square variables. *Journal of the American Statistical Association* 1977; 72(360): 881–885.
8. Biggerstaff BJ, Jackson D. The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. *Statistics in Medicine* 2008; 27(29): 6093–6110.
9. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Cinical Trials* 1986; 7(3): 177–188.
10. Mandel J, Paule RC. Interlaboratory evaluation of a material with unequal numbers of replicates. *Analytical Chemistry* 1970; 42(11): 1194–1197.
11. Sánchez-Meca J, Marín-Martínez F. Testing the significance of a common risk difference in meta-analysis. *Computational Statistics & Data Analysis* 2000; 33(3): 299–313.
12. Kulinskaya E, Hoaglin DC, Newman J, Bakbergenuly I. Simulations for a Q statistic with constant weights to assess heterogeneity in meta-analysis of mean difference. *eprint arXiv:2010.11009v1 [stat.ME]* 2020.
13. Wilk MB, Gnanadesikan R. Probability plotting methods for the analysis of data. *Biometrika* 1968; 53(1): 1–17.

14. Rees K, Taylor RRS, Singh S, Coats AJS, Ebrahim S. Exercise based rehabilitation for heart failure. *Cochrane Database of Systematic Reviews* 2004(3). doi: 10.1002/14651858.CD003331.pub2

15. Spooner C, Spooner G, Rowe B. Mast-cell stabilising agents to prevent exercise-induced bronchoconstriction. *Cochrane Database of Systematic Reviews* 2003(4). doi: 10.1002/14651858.CD002307

16. Higgins JPT, Thompson SG, Spiegelhalter DJ. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 2009; 172(1): 137-159. doi: https://doi.org/10.1111/j.1467-985X.2008.00552.x

17. Hartung J, Knapp G. A refined method for the meta-analysis of controlled clinical trials with binary outcome. *Statistics in Medicine* 2001; 20(24): 3875–3889.

18. Sidik K, Jonkman JN. A simple confidence interval for meta-analysis. *Statistics in Medicine* 2002; 21(21): 3153–3159.

19. Partlett C, Riley RD. Random effects meta-analysis: Coverage performance of 95% confidence and prediction intervals following REML estimation. *Statistics in Medicine* 2017; 36(2): 301-317. doi: https://doi.org/10.1002/sim.7140

20. Sánchez-Meca J, Marín-Martínez F. Meta-analysis in psychological research.. *International Journal of Psychological Research* 2010; 3(1): 150–162.

21. Hamman EA, Pappalardo P, Bence JR, Peacor SD, Osenberg CW. Bias in meta-analyses using Hedges' d. *Ecosphere* 2018; 9(9): e02419. doi: 10.1002/ecs2.2419

22. Welch BL. The significance of the difference between two means when the population variances are unequal. *Biometrika* 1938; 29(3/4): 350–362.

23. Satterthwaite FE. Synthesis of variance. *Psychometrika* 1941; 6(5): 309–316.

24. Imhof JP. Computing the distribution of quadratic forms in normal variables. *Biometrika* 1961; 48(3/4): 419–426.

25. Pearson ES. Note on an approximation to the distribution of non-central χ^2. *Biometrika* 1959; 46(3/4): 364.

26. Liu H, Tang Y, Zhang HH. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics & Data Analysis* 2009; 53(4): 853–856.

27. Yuan KH, Bentler PM. Two simple approximations to the distributions of quadratic forms. *British Journal of Mathematical and Statistical Psychology* 2010; 63(2): 273–291.

28. Sheil J, O'Muircheartaigh I. Algorithm AS 106. The distribution of non-negative quadratic forms in normal variables. *Applied Statistics* 1977; 26: 92–98.

29. Kuonen D. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* 1999; 86(4): 929–935.

30. Zghoul AA. APPROXIMATING THE DISTRIBUTION OF QUADRATIC FORMS USING ORTHOGONAL POLYNOMIALS. *Qatar University Science Journal* 1999; 18: 15–25.

31. Ha HT, Provost SB. AN ACCURATE APPROXIMATION TO THE DISTRIBUTION OF A LINEAR COMBINATION OF NON-CENTRAL CHI-SQUARE RANDOM VARIABLES. *REVSTAT – Statistical Journal* 2013; 11(3): 231–254.

32. Bodenham D, Adams N. A comparison of efficient approximations for a weighted sum of chi-squared random variables. *Statistics and Computing* 2016; 26: 917–928. doi: 10.1007/s11222-015-9583-4

33. Chen T, Lumley T. Numerical evaluation of methods approximating the distribution of a large quadratic form in normal variables. *Computational Statistics & Data Analysis* 2019; 139: 75–81. doi: 10.1016/j.csda.2019.05.002

34. Götze F, Tikhomirov A. Asymptotic distribution of quadratic forms and applications. *Journal of Theoretical Probability* 2002; 15(2): 426–475.

35. Kulinskaya E, Dollinger MB, Knight E, Gao H. A Welch-type test for homogeneity of contrasts under heteroscedasticity with application to meta-analysis. *Statistics in Medicine* 2004; 23(23): 3655–3670. doi: 10.1002/sim.1929

36. Welch BL. On the comparison of several mean values: an alternative approach. *Biometrika* 1951; 38(3/4): 330–336.

37. Kulinskaya E, Dollinger MB, Bjørkestøl K. Testing for homogeneity in meta-analysis I. The one-parameter case: standardized mean difference. *Biometrics* 2011; 67(1): 203–212.

38. Kulinskaya E, Dollinger MB, Bjørkestøl K. On the moments of Cochran's Q statistic under the null hypothesis, with application to the meta-analysis of risk difference. *Research Synthesis Methods* 2011; 2(4): 254–270.

39. Kulinskaya E, Dollinger MB. An accurate test for homogeneity of odds ratios based on Cochran's Q-statistic. *BMC Medical Research Methodology* 2015; 15: 49.

40. Soetaert K, Hindmarsh AC, Eisenstat S, Moler C, Dongarra J, Saad Y. Nonlinear Root Finding, Equilibrium and Steady-State Analysis of Ordinary Differential Equations. https://cran.r-project.org/web/packages/rootSolve/rootSolve.pdf; 2020. Version 1.8.2.1.

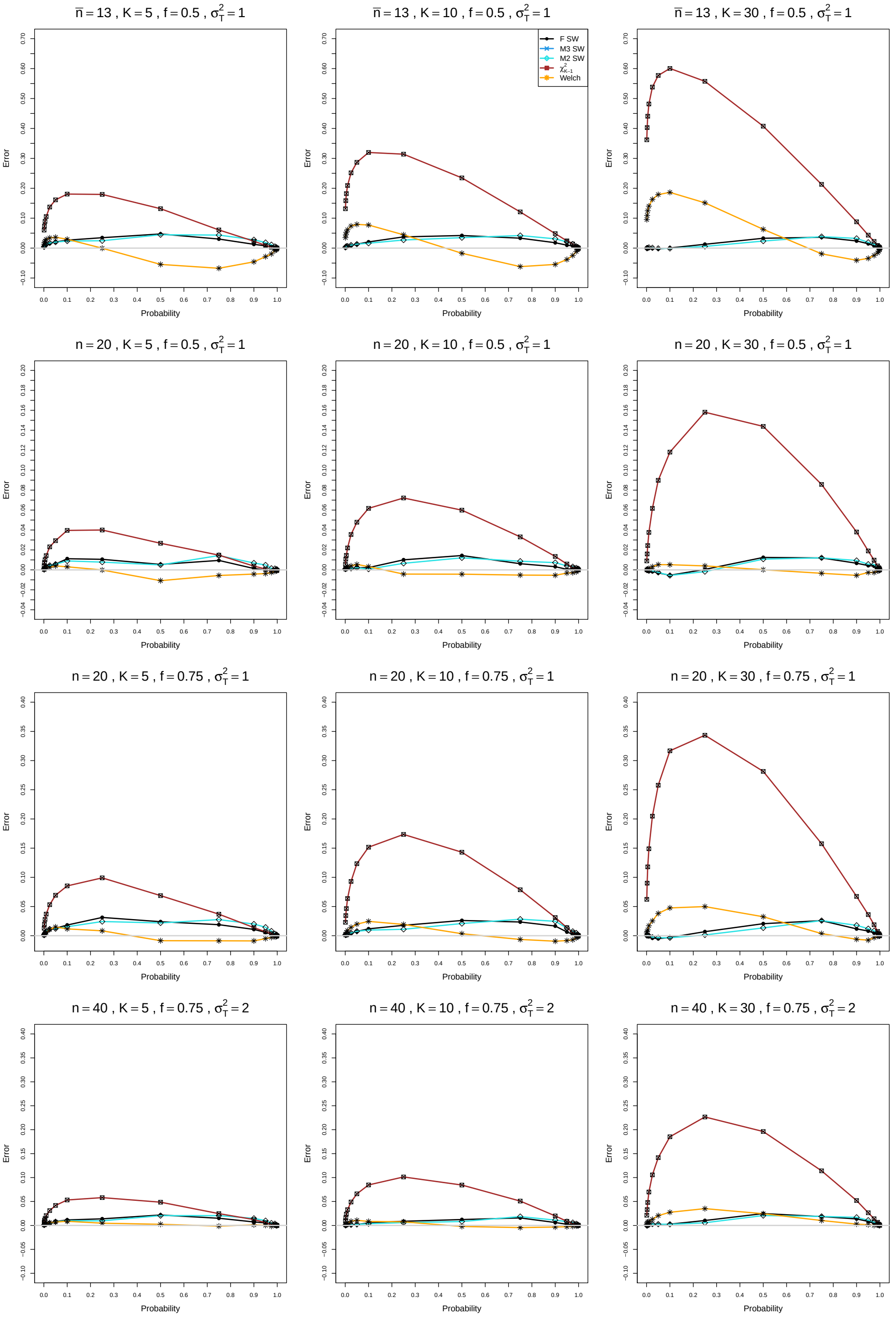


FIGURE 1 P–P plots of the Farebrother, M2, and M3 approximations to the distribution of Q with sample-size-based weights, and of the chi-square and Welch approximations to the distribution of Q with IV-based weights. First row: unequal sample sizes, $\bar{n} = 13$, $\sigma_C^2 = \sigma_T^2 = 1$, $f = .5$; second and subsequent rows: equal sample sizes, $\sigma_C^2 = 1$. Second row: $n = 20$, $\sigma_T^2 = 1$, $f = .5$; third row: $n = 20$, $\sigma_T^2 = 1$, $f = .75$; fourth row: $n = 40$, $\sigma_T^2 = 2$, $f = .75$. (The scale on the vertical axis varies among the rows.)

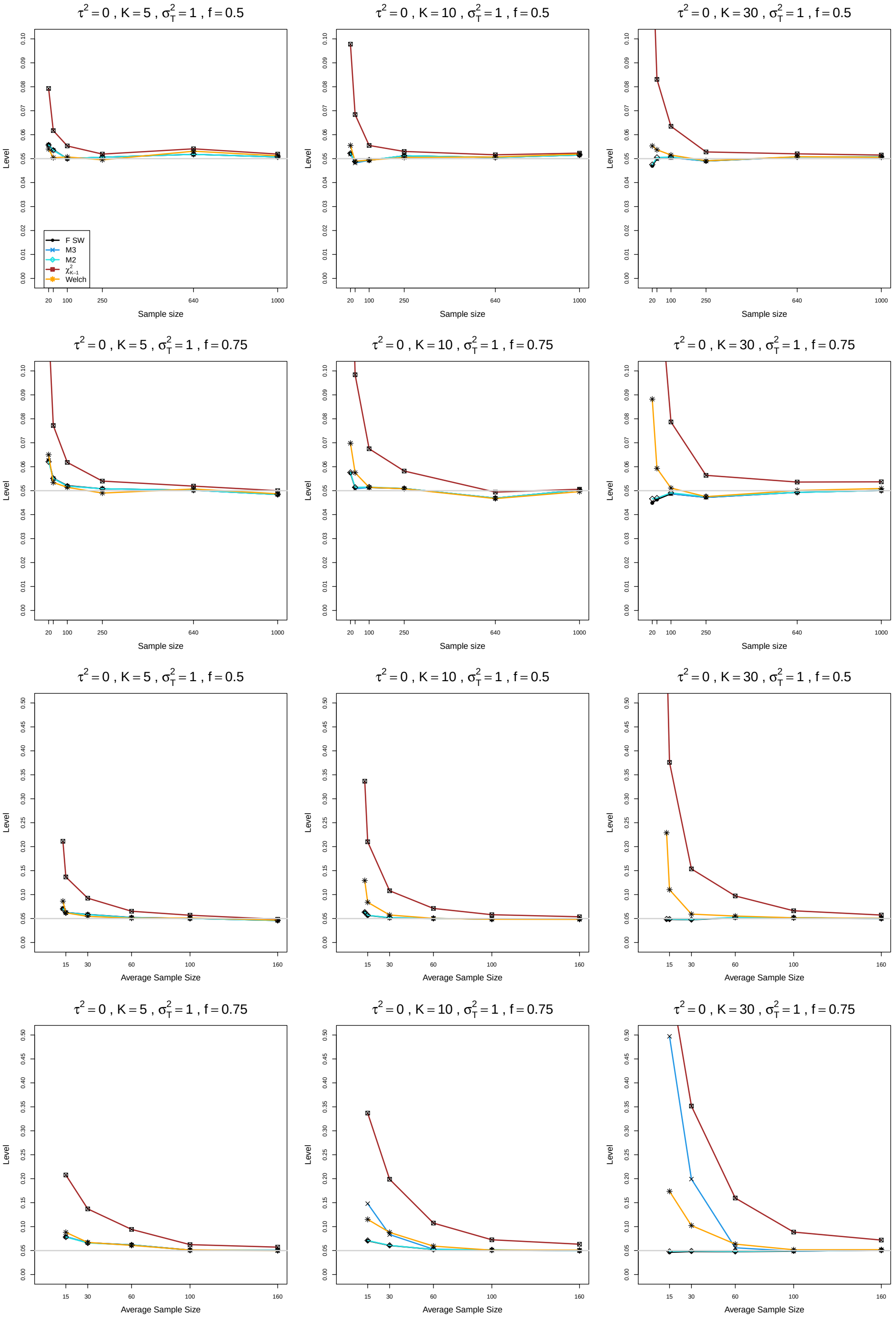


FIGURE 2 Empirical levels of approximations to the distribution of Q with IV or sample-size-based weights at nominal .05 level vs sample size n. In all plots, $\tau^2 = 0$ and $\sigma_C^2 = \sigma_T^2 = 1$. Top two rows: equal sample sizes, $f = .5$ and $f = .75$. Bottom two rows: unequal sample sizes, $f = .5$ and $f = .75$. (The vertical scale in the top two rows differs from that in the bottom two rows.)

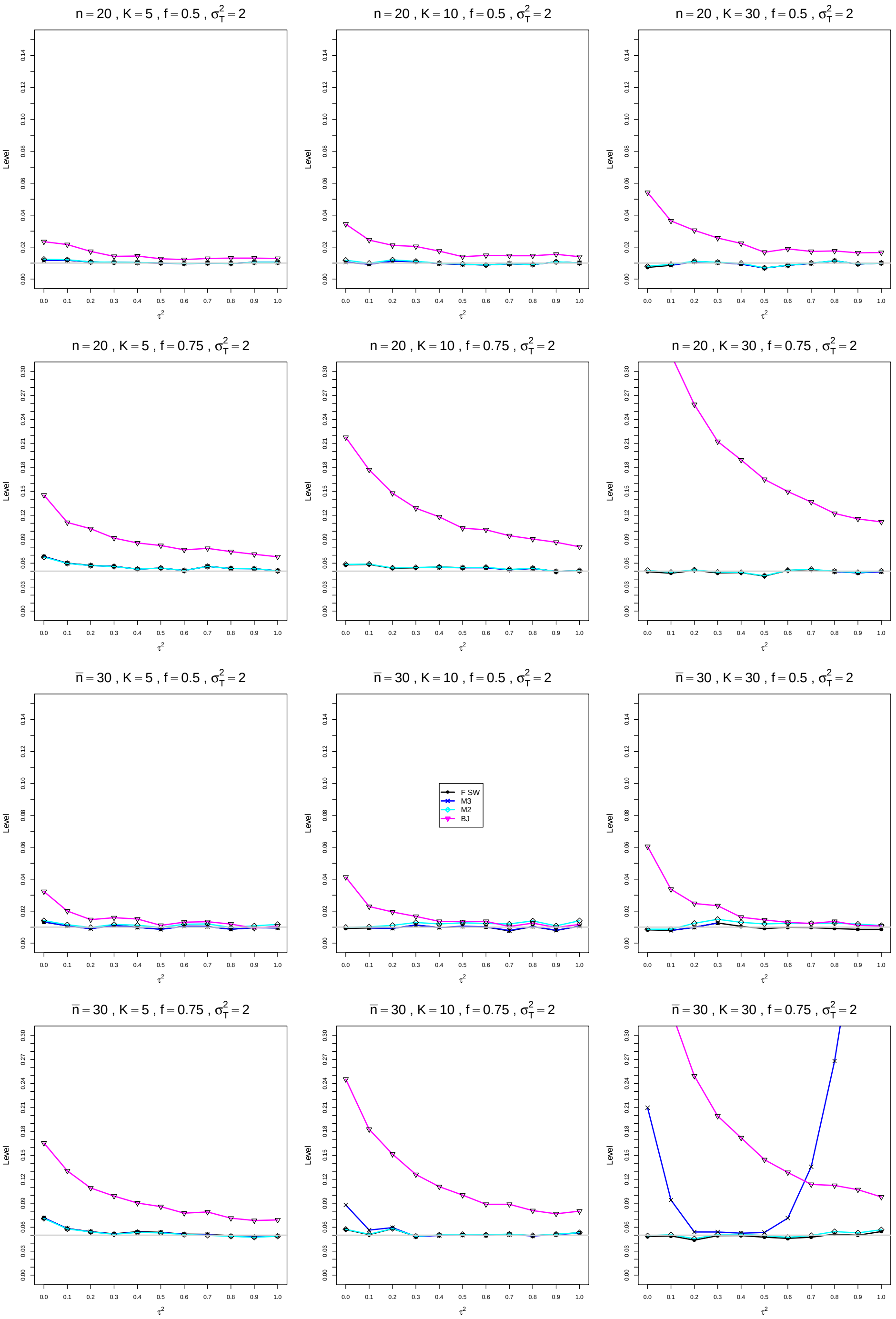


FIGURE 3 Empirical p-values of approximations to the distribution of Q with IV or sample-size-based weights at the nominal .01 and .05 levels vs between-study variance τ^2. In all plots, $\sigma_C^2 = 1$ and $\sigma_T^2 = 2$. Top two rows: equal sample sizes $n = 20$, $f = .5$ and $f = .75$. Bottom two rows: unequal sample sizes, $f = .5$ and $f = .75$. First and third rows: .01. Second and fourth rows: .05.

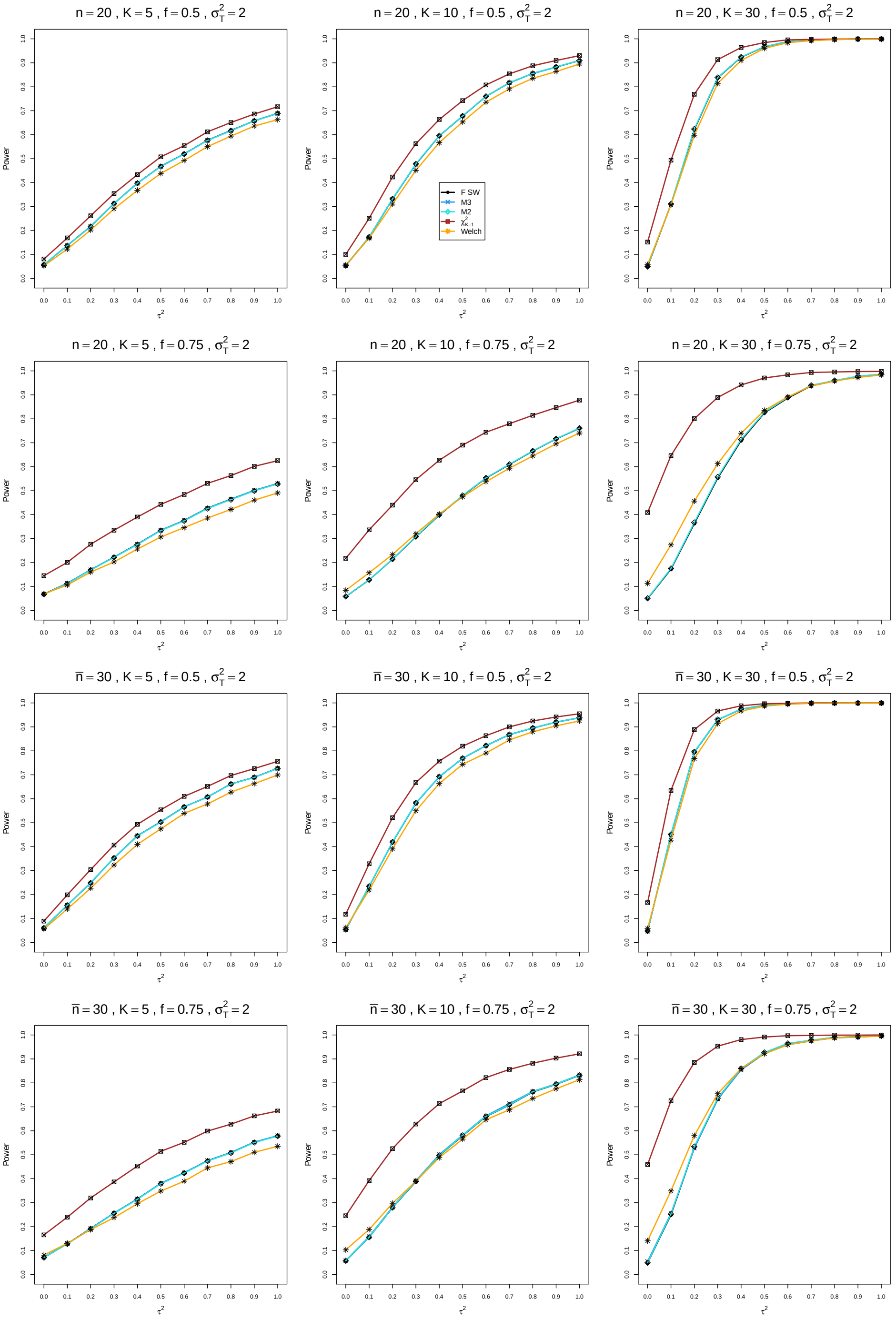


FIGURE 4 Power of tests of heterogeneity at .05 level for equal ($n = 20$) and unequal ($\bar{n} = 30$) sample sizes. In all plots, $\sigma_C^2 = \sigma_T^2 = 1$. Top two rows: equal sample sizes, $f = .5$ and $f = .75$. Bottom two rows: unequal sample sizes, $f = .5$ and $f = .75$.

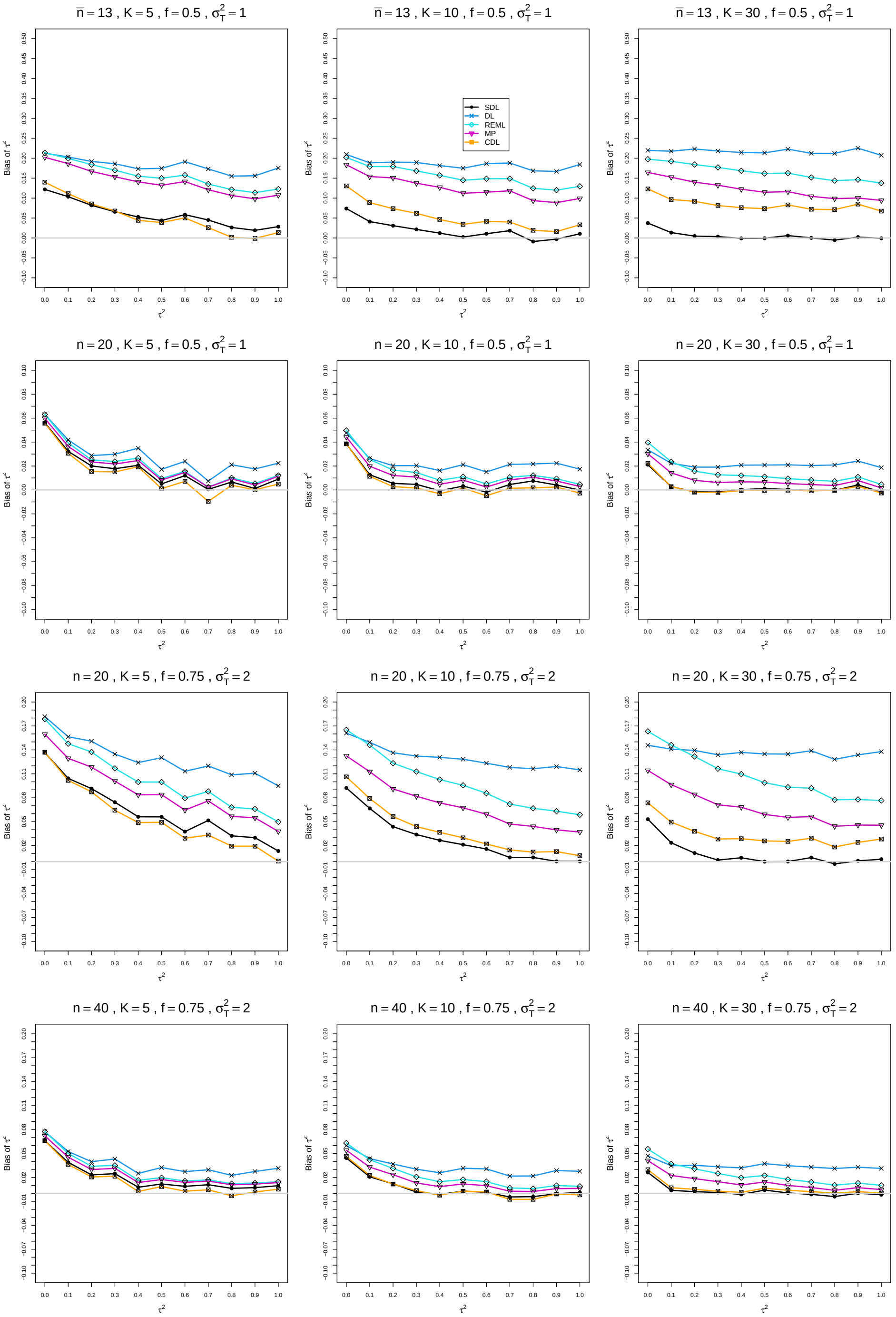
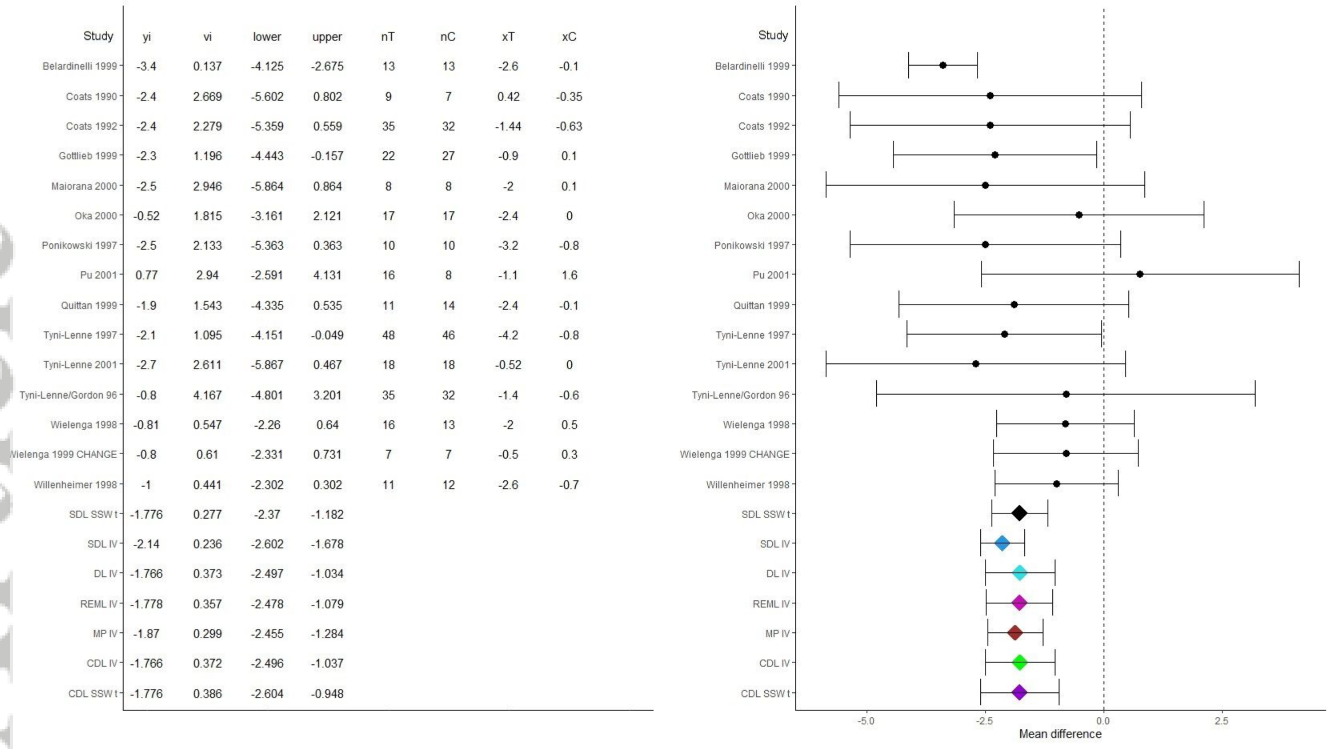
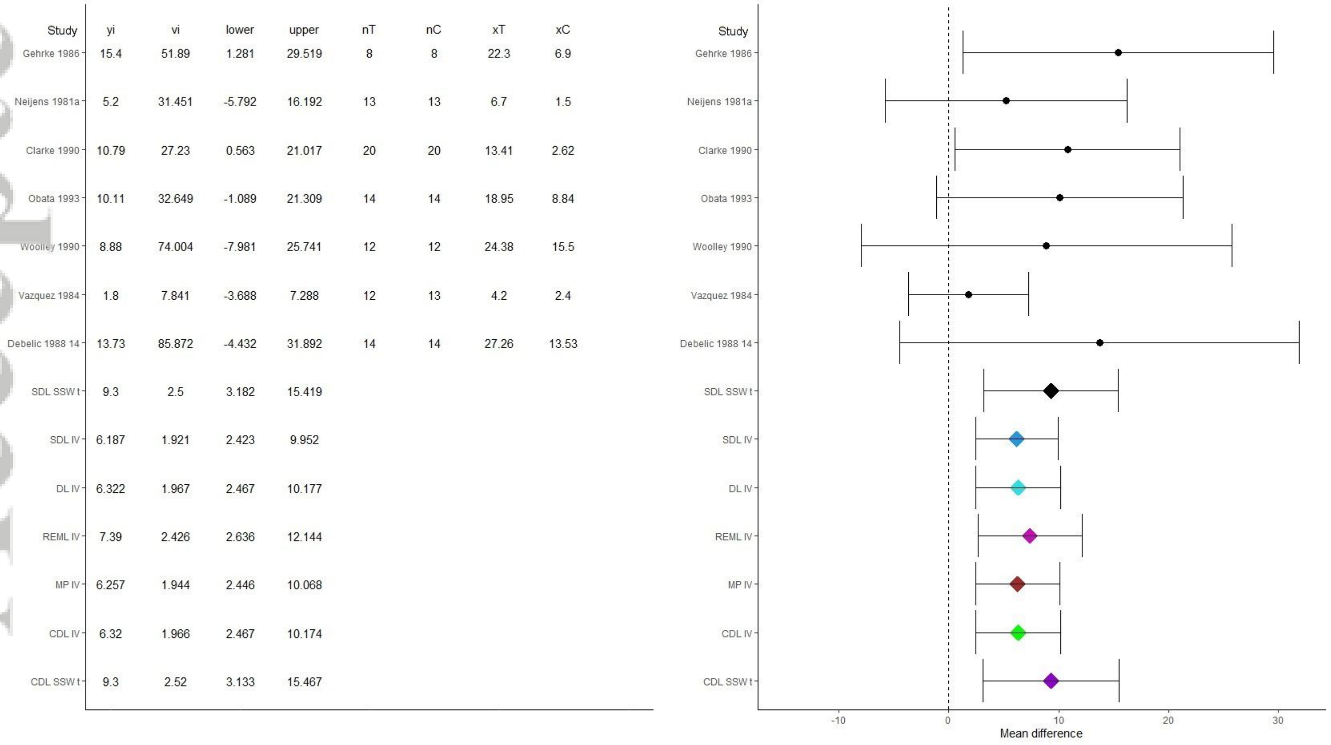


FIGURE 5 Bias in estimation of between-study variance τ^2 by five methods: SDL, DL, REML, MP, and CDL. First row: unequal sample sizes, $\bar{n} = 13$, $\sigma_C^2 = \sigma_T^2 = 1$, $f = .5$; second and subsequent rows: equal sample sizes, $\sigma_C^2 = 1$. Second row: $n = 20$, $\sigma_T^2 = 1$, $f = .5$; third row: $n = 20$, $\sigma_T^2 = 1$, $f = .75$; fourth row: $n = 40$, $\sigma_T^2 = 2$, $f = .75$. (The scale on the vertical axis varies among the rows.)



Study	yi	vi	lower	upper	nT	nC	xT	xC
Belardinelli 1999	-3.4	0.137	-4.125	-2.675	13	13	-2.6	-0.1
Coats 1990	-2.4	2.669	-5.602	0.802	9	7	0.42	-0.35
Coats 1992	-2.4	2.279	-5.359	0.559	35	32	-1.44	-0.63
Gottlieb 1999	-2.3	1.196	-4.443	-0.157	22	27	-0.9	0.1
Maiorana 2000	-2.5	2.946	-5.864	0.864	8	8	-2	0.1
Oka 2000	-0.52	1.815	-3.161	2.121	17	17	-2.4	0
Ponikowski 1997	-2.5	2.133	-5.363	0.363	10	10	-3.2	-0.8
Pu 2001	0.77	2.94	-2.591	4.131	16	8	-1.1	1.6
Quittan 1999	-1.9	1.543	-4.335	0.535	11	14	-2.4	-0.1
Tyni-Lenne 1997	-2.1	1.095	-4.151	-0.049	48	46	-4.2	-0.8
Tyni-Lenne 2001	-2.7	2.611	-5.867	0.467	18	18	-0.52	0
Tyni-Lenne/Gordon 96	-0.8	4.167	-4.801	3.201	35	32	-1.4	-0.6
Wielenga 1998	-0.81	0.547	-2.26	0.64	16	13	-2	0.5
Wielenga 1999 CHANGE	-0.8	0.61	-2.331	0.731	7	7	-0.5	0.3
Willenheimer 1998	-1	0.441	-2.302	0.302	11	12	-2.6	-0.7
SDL SSW t	-1.776	0.277	-2.37	-1.182				
SDL IV	-2.14	0.236	-2.602	-1.678				
DL IV	-1.766	0.373	-2.497	-1.034				
REML IV	-1.778	0.357	-2.478	-1.079				
MP IV	-1.87	0.299	-2.455	-1.284				
CDL IV	-1.766	0.372	-2.496	-1.037				
CDL SSW t	-1.776	0.386	-2.604	-0.948				

FIGURE 6 Data and forest plot for Rees (2004) meta-analysis on exercise-related changes in VO_2max



Study	yi	vi	lower	upper	nT	nC	xT	xC
Gehrke 1986	15.4	51.89	1.281	29.519	8	8	22.3	6.9
Neijens 1981a	5.2	31.451	-5.792	16.192	13	13	6.7	1.5
Clarke 1990	10.79	27.23	0.563	21.017	20	20	13.41	2.62
Obata 1993	10.11	32.649	-1.089	21.309	14	14	18.95	8.84
Woolley 1990	8.88	74.004	-7.981	25.741	12	12	24.38	15.5
Vazquez 1984	1.8	7.841	-3.688	7.288	12	13	4.2	2.4
Debelic 1988 14	13.73	85.872	-4.432	31.892	14	14	27.26	13.53
SDL SSW t	9.3	2.5	3.182	15.419				
SDL IV	6.187	1.921	2.423	9.952				
DL IV	6.322	1.967	2.467	10.177				
REML IV	7.39	2.426	2.636	12.144				
MP IV	6.257	1.944	2.446	10.068				
CDL IV	6.32	1.966	2.467	10.174				
CDL SSW t	9.3	2.52	3.133	15.467				

FIGURE 7 Data and forest plot for Spooner et al.(2003) meta-analysis on drugs for prevention of exercise-induced asthma attacks

Appendix

This appendix assembles the more-technical information related to evaluating and approximating the distribution of Q. Section A.1 discusses approaches, in the broader context of quadratic forms in normal random variables. Section A.2 explains the form of the two-moment and three-moment approximations. Then Section A.3 presents derivations for the variance and third moment of Q. The resulting expressions involve the first six unconditional moments of Θ_i. Section A.4 develops those moments for a general effect measure, and Section A.5 applies and simplifies them for the mean difference.

A.1 | APPROXIMATIONS TO THE DISTRIBUTION OF QUADRATIC FORMS IN NORMAL VARIABLES

The Q statistic, Equation (2), is a quadratic form in the random variables Θ_i. We can write $Q = \Theta^T A \Theta$ for a symmetric matrix A of rank $K-1$ with the elements $a_{ij} = q_i\delta_{ij} - q_i q_j$, $1 \leq i, j \leq K$, where δ_{ij} is the Kronecker delta. In this section we assume constant weights unless stated otherwise. Unconditionally, the Θ_i are centered at 0, but they are not, in general, normally distributed. However, for large sample sizes n_i, their distributions are approximately normal. Normality holds exactly for the mean difference (MD). In this case the exact distribution of the quadratic form is that of a weighted sum of central chi-square variables. but the cumulative distribution function of Q needs to be evaluated numerically. Therefore, we consider suitable approximations.

Quadratic forms in normal variables have an extensive literature. When the vector Θ has the multivariate normal distribution $N(\mu, \Sigma)$, the exact distribution of Q is $\sum_{r=1}^{m} \lambda_r \chi^2_{h_r}(\delta_r^2)$, where the λ_r are the eigenvalues of $A\Sigma$, the h_r are their multiplicities, and the δ_r^2 are the non-centrality parameters for the independent chi-square variables $\chi^2_{h_r}(\delta_r^2)$ with h_r degrees of freedom. (The δ_r are linear combinations of $\mu_1, \ldots, \mu_K$.)

Interest typically centers on the upper-tail probabilities $P(Q > x)$. Moment-based approximations match a particular distribution, often a gamma distribution or, equivalently, a scaled chi-square distribution, to several moments of Q. These methods include the well-known Welch-Satterthwaite approximation, which uses $c\chi^2_p$ and matches the first two moments[22,23]. Imhof[24] investigated an approximation to the distribution of a quadratic form in noncentral normal variables by matching a central chi-square distribution to three moments (including the skewness). The approximation has the form $Q \sim (\chi^2_{h'} - h')(2h')^{-\frac{1}{2}} \sqrt{\text{Var}(Q)} + E(Q)$. Pearson[25] first suggested this approach to approximate a noncentral chi-square distribution. Liu et al.[26] proposed a four-moment noncentral chi-square approximation. To approximate the probability that a standardized Q exceeds t^*, they use the probability that a standardized noncentral chi-square exceeds t^*, equating the skewness of the two distributions and matching the kurtosis as closely as possible.

Yuan and Bentler[27] studied, by simulation, the Type I errors of a Q test with the critical values based on the Welch-Satterthwaite approximation. They concluded that this approximation is satisfactory when the eigenvalues do not have too large a coefficient of variation, preferably less than 1. For larger CV, the Type I errors may be larger than nominal.

For the general case of a noncentral quadratic form, the distribution of Q can be approximated by the distribution of cU^r, where the distribution of U can depend on one or two parameters. The choice of c, r, and the parameters of U then permits matching the necessary moments. Solomon and Stephens[7] consider three moment-based approximations: a four-moment approximation by a Type III Pearson curve and two three-moment approximations, one with $U \sim N(\mu, \sigma^2)$ and the other with $U \sim \chi^2_p$. They recommend the latter as fitting better in the lower tail, partly because it necessarily starts at zero, whereas the other approximations do not. This approximation matches the constants c, r, and p to the first three moments of Q. For $c(\chi^2_p)^r$ the moments about 0 are $\mu'_k = c^k 2^{kr} \Gamma(kr + p/2)/\Gamma(p/2)$.

Other, more-complicated methods include relying on numerical inversion of the characteristic function[24]; this can be made very accurate, with bounds on accuracy. The algorithm of Sheil and O'Muircheartaigh[28], improved by Farebrother[5], represents the value of the c.d.f. for a noncentral quadratic form by an infinite sum of central chi-square probabilities. Kuonen[29] proposes a saddlepoint approximation, and Zghoul[30] and Ha and Provost[31] consider approximations by Hermite and Laguerre polynomials. The first two methods are nearly exact and perform better than Pearson's three-moment approximation by a central chi-square distribution or, in the noncentral case, the four-moment approximation by a Type III Pearson curve[24,6]. Bodenham and Adams[32] and Chen and Lumley[33] discuss the behavior of various approximations when K is large.

We are aware of only one paper[34] on the asymptotic ($K \to \infty$) distribution of quadratic forms in non-normal iid random variables with finite sixth moment. This distribution can be approximated by that of a second-order polynomial in normal variables.

In meta-analysis, approximations to the distribution of Q have usually been sought only for the Q_{IV} version with non-constant inverse-variance weights. Typically, the chi-square distribution with $K-1$ degrees of freedom is used indiscriminately as the null distribution of Q_{IV}. For MD, Kulinskaya et al.[35] introduced an improved two-moment approximation to this version of Q based on the Welch[36] test in the heteroscedastic ANOVA. The distribution of this Welch test for MD is approximated under the null by a rescaled F distribution, and under alternatives by a shifted chi-square distribution. Kulinskaya and coauthors also explored improved moment-based approximations for some other effect measures[37,38,39], using two-moment approximations with a scaled chi-square distribution to the null distribution of Q_{IV}. Biggerstaff and Jackson[8] used the Farebrother approximation to the distribution of a quadratic form in normal variables as the "exact" distribution of Q_{IV}. This is not correct when the weights are the reciprocals of estimated variances, but with constant weights it is correct for MD. When $\tau^2 = 0$, the Biggerstaff and Jackson approximation to the distribution of Q_{IV} is the χ^2_{K-1} distribution.

A.2 | TWO- AND THREE-MOMENT APPROXIMATIONS TO THE DISTRIBUTION OF Q

The two- and three-moment approximations to the distribution of Q use the distribution of a transformed chi-square random variable $c(\chi^2_p)^r$. The parameters c, r, and p are found by matching the first two or three moments.

The kth moment about zero for $c(\chi^2_p)^r$ is

$$\mu'_k = \frac{c^k 2^{kr}\Gamma(kr + p/2)}{\Gamma(p/2)}.$$

A.2.1 | Two-moment approximation

The two-moment approximation by[23] and[22] sets $r = 1$, so $Q \sim c(\chi^2_p)$. Matching the first moment μ'_1 to E(Q), we obtain

$$\frac{2c\Gamma(1 + p/2)}{\Gamma(p/2)} = E[Q].$$

Since $\Gamma(n+1) = n\Gamma(n)$, the above equation reduces to

$$cp = E[Q]. \tag{A.1}$$

For the second moment μ'_2,

$$\frac{4c^2\Gamma(2 + p/2)}{\Gamma(p/2)} = E[Q^2],$$

which reduces to

$$c^2 p(p+2) = E[Q^2]. \tag{A.2}$$

Solving for c in equation (A.1) and substituting the result into (A.2) yield

$$c = E[Q]/p, \quad p = 2\left[\frac{E[Q^2]}{E[Q]^2} - 1\right]^{-1}.$$

A.2.2 | Three-moment approximation

For the three-moment approximations we have $Q \sim c(\chi^2_p)^r$. Similar to the two-moment case, we set $k = 1, 2, 3$ to obtain the following system of equations

$$(\mu'_1): \quad \frac{2^r c\Gamma(r + p/2)}{\Gamma(p/2)} = E[Q];$$

$$(\mu'_2): \quad \frac{2^{2r} c^2\Gamma(2r + p/2)}{\Gamma(p/2)} = E[Q^2];$$

$$(\mu'_3): \quad \frac{2^{3r} c^3\Gamma(3r + p/2)}{\Gamma(p/2)} = E[Q^3].$$

Dividing μ'_2 by μ'_1, we obtain the following expression for c:

$$c = \frac{E[Q^2]\Gamma(r + p/2)}{2^r E[Q]\Gamma(2r + p/2)}.$$

To eliminate c, define $A = \mu_2'/(\mu_1')^2$ and $B = \mu_3'/(\mu_1')^3$. Then we have the following two nonlinear equations:

$$A = \frac{\Gamma(2r + p/2)\Gamma(p/2)}{\Gamma^2(r + p/2)}, \quad B = \frac{\Gamma(3r + p/2)\Gamma^2(p/2)}{\Gamma^3(r + p/2)}.$$

We solve this system for p and r by using the function *'multiroot'* in the R package *rootSolve*[40] with the starting values $r = 1$ and c and p from the two-moment approximation.

A.3 | VARIANCE AND THIRD MOMENT OF Q

For approximations based on the first two or three moments, we need the second and the third moments of Q under the REM introduced in Section 1.

We distinguish between the conditional distribution of Q (given the θ_i) and the unconditional distribution, and the respective moments of Θ_i. For instance, the conditional second moment of Θ_i is $M_{2i}^c = v_i^2$, and the unconditional second moment is $M_{2i} = \mathrm{E}(\Theta_i^2) = \mathrm{Var}(\hat{\theta}_i) = \mathrm{E}(v_i^2) + \tau^2$. Similarly, $M_{4i} = \mathrm{E}(\Theta_i^4)$ is the fourth (unconditional) central moment of $\hat{\theta}_i$. These two moments are required to calculate the variance of Q, given by

$$W^{-2}\mathrm{Var}(Q) = \sum_i q_i^2(1 - q_i)^2(M_{4i} - M_{2i}^2) + 2\sum_{i \neq j} q_i^2 q_j^2 M_{2i} M_{2j}. \tag{A.3}$$

Section A.3.1 gives the details. When the weights are not related to the effect, these expressions for the mean and variance of Q are the same as in Kulinskaya et al.[37].

For (known) inverse-variance weights $w_i = v_i^{-2}$, and assuming that each $\hat{\theta}_i$ is normally distributed and $\tau^2 = 0$, so that $M_{2i} = v_i^2$ and $M_{4i} = 3v_i^4$, the first moment of Q is $K - 1$, and the variance is $2(K - 1)$, as it should be for a chi-square distribution with $K - 1$ degrees of freedom.

In general, the unconditional moments M_{2i} and M_{4i} depend on the effect measure (through its second and fourth conditional moments) and on the REM that defines the unconditional moments. Section A.4 gives the details.

In the null distribution $\tau^2 = 0$, and the unconditional moments of Q coincide with its conditional moments.

The derivation for the unconditional third moment of Q

$$W^{-3}\mathrm{E}(Q^3) = \mathrm{E}\{[\sum q_i(1 - q_i)\Theta_i^2 - \sum\sum_{i \neq j} q_i q_j \Theta_i \Theta_j]^3\}$$

parallels that for the second moment, starting from Equation (2). Section A.3.2 gives the details of the derivation.

Importantly, $M_{3i} = \mathrm{E}(\Theta_i^3)$ and $M_{6i} = \mathrm{E}(\Theta_i^6)$, the third and the sixth unconditional central moments of $\hat{\theta}_i$, are required for this calculation, in addition to the second and the fourth central moments used in calculating the second moment of Q.

Unconditional central moments of $\hat{\theta}_i$ are linear combinations of expected values of conditional moments, their cross-products, and powers of τ^2. Section A.4 provides the requisite expressions for the first six unconditional central moments for a general effect measure. Calculations of unconditional moments are much simpler for the mean difference (MD), as we show in Section A.5.

A.3.1 | Calculation of the second moment of Q

The second moment of Q (times W^{-2}) is

$$\begin{aligned} W^{-2}\mathrm{E}(Q^2) &= \mathrm{E}\left[\sum q_i(1 - q_i)\Theta_i^2\right]^2 \\ &\quad -2\mathrm{E}\left(\left[\sum q_k(1 - q_k)\Theta_k^2\right]\left[\sum_{i \neq j} q_i q_j \Theta_i \Theta_j\right]\right) \\ &\quad +\mathrm{E}\left[\sum_{i \neq j} q_i q_j \Theta_i \Theta_j\right]^2 \\ &= A - 2B + C. \end{aligned}$$

The first term,

$$\begin{aligned} A &= \mathrm{E}\left(\sum_{i,j} q_i(1-q_i)q_j(1-q_j)\Theta_i^2\Theta_j^2\right) \\ &= \mathrm{E}\left(\sum_{i} q_i^2(1-q_i)^2\Theta_i^4\right) + \mathrm{E}\left(\sum_{i\neq j} q_i(1-q_i)q_j(1-q_j)\Theta_i^2\Theta_j^2\right) \\ &= \sum_i q_i^2(1-q_i)^2(M_{4i}-M_{2i}^2) + \left[\sum q_i(1-q_i)M_{2i}\right]^2, \end{aligned}$$

where $M_{2i} = \mathrm{E}(\Theta_i^2) = \mathrm{Var}(\Theta_i) = \mathrm{E}(v_i^2) + \tau^2$ is the variance, and $M_{4i} = \mathrm{E}(\Theta_i^4)$ is the fourth central moment of $\hat{\theta}_i$.

The second term, $B = 0$ because its terms $\mathrm{E}(\Theta_k^2\Theta_i\Theta_j)$, with $i \neq j$, always include a first-order moment of Θ_i for some i.

In the third term, $C = \mathrm{E}[\sum_{i\neq j}\sum_{k\neq l} q_iq_jq_kq_l\Theta_i\Theta_j\Theta_k\Theta_l]$, the only nonzero terms have $i = k$ and $j = l$ or $i = l$ and $j = k$, so $C = 2\sum_{i\neq j} q_i^2q_j^2M_{2i}M_{2j}$.

To obtain W^{-2} times the variance of Q, we subtract the square of its mean, given by Equation (3), which is exactly the second term of A:

$$W^{-2}\mathrm{Var}(Q) = \sum_i q_i^2(1-q_i)^2(M_{4i}-M_{2i}^2) + 2\sum_{i\neq j} q_i^2q_j^2M_{2i}M_{2j}.$$

A.3.2 | Calculation of the third moment of Q

For the derivation of the third moment of Q, we record selected steps. We have

$$\begin{aligned} W^{-3}\mathrm{E}(Q^3) &= \mathrm{E}\{[\sum q_i(1-q_i)\Theta_i^2 - \sum\sum_{i\neq j} q_iq_j\Theta_i\Theta_j]^3\} \\ &= \mathrm{E}\{[\sum q_i(1-q_i)\Theta_i^2]^3\} \\ &\quad -3\mathrm{E}\{[\sum q_i(1-q_i)\Theta_i^2]^2[\sum\sum_{i\neq j} q_iq_j\Theta_i\Theta_j]\} \\ &\quad +3\mathrm{E}\{[\sum q_i(1-q_i)\Theta_i^2][\sum\sum_{i\neq j} q_iq_j\Theta_i\Theta_j]^2\} \\ &\quad -\mathrm{E}\{[\sum\sum_{i\neq j} q_iq_j\Theta_i\Theta_j]^3\} \\ &= A - 3B + 3C - D \end{aligned}$$

The terms A, B, C, and D are obtained below.

$$\begin{aligned} A &= \mathrm{E}[\textstyle\sum_i\sum_j\sum_k q_i(1-q_i)q_j(1-q_j)q_k(1-q_k)\Theta_i^2\Theta_j^2\Theta_k^2] \\ &= \textstyle\sum q_i^3(1-q_i)^3M_{6i} \\ &\quad +3\textstyle\sum\sum_{i\neq j} q_i^2(1-q_i)^2q_j(1-q_j)M_{4i}M_{2j} \\ &\quad +\textstyle\sum\sum\sum_{i\neq j\neq k} q_i(1-q_i)q_j(1-q_j)q_k(1-q_k)M_{2i}M_{2j}M_{2k} \\ &= \textstyle\sum q_i^3(1-q_i)^3M_{6i} \\ &\quad +3\{[\textstyle\sum_i q_i^2(1-q_i)^2M_{4i}][\sum_j q_j(1-q_j)M_{2j}] - \sum q_i^3(1-q_i)^3M_{4i}M_{2i}\} \\ &\quad +\{[\textstyle\sum_i q_i(1-q_i)M_{2i}]^3 - 3[\sum_i q_i^2(1-q_i)^2M_{2i}^2][\sum_j q_j(1-q_j)M_{2j}] + 2\sum_i q_i^3(1-q_i)^3M_{2i}^3\} \\ &= \textstyle\sum q_i^3(1-q_i)^3[M_{6i} - 3M_{4i}M_{2i} + 2M_{2i}^3] \\ &\quad +3[\textstyle\sum_j q_j(1-q_j)M_{2j}][\sum_i q_i^2(1-q_i)^2(M_{4i} - M_{2i}^2] \\ &\quad +[\textstyle\sum_i q_i(1-q_i)M_{2i}]^3 \end{aligned}$$

$$
\begin{aligned}
B &= \mathrm{E}\{[\sum_i \sum_j q_i(1-q_i)q_j(1-q_j)\Theta_i^2\Theta_j^2][\sum\sum_{i\neq j} q_i q_j \Theta_i \Theta_j]\} \\
&= \mathrm{E}\{[\sum_i q_i^2(1-q_i)^2\Theta_i^4 + \sum\sum_{i\neq j} q_i(1-q_i)q_j(1-q_j)\Theta_i^2\Theta_j^2][\sum\sum_{i\neq j} q_i q_j \Theta_i \Theta_j]\} \\
&= \mathrm{E}\{2\sum\sum_{i\neq j} q_i^3(1-q_i)^2 q_j \Theta_i^5 \Theta_j + \sum\sum\sum_{i\neq j\neq k} q_i^2(1-q_i)^2 q_j q_k \Theta_i^4 \Theta_j \Theta_k \\
&+ \sum\sum_{i\neq j}\sum\sum_{l\neq k} q_i(1-q_i)q_j(1-q_j)q_k q_l \Theta_i^2\Theta_j^2\Theta_k\Theta_l\} \\
&= 2\sum\sum_{i\neq j} q_i^3(1-q_i)^2 q_j \mathrm{E}(\Theta_i^5)\mathrm{E}(\Theta_j) + \sum\sum\sum_{i\neq j\neq k} q_i^2(1-q_i)^2 q_j q_k \mathrm{E}(\Theta_i^4)\mathrm{E}(\Theta_j)\mathrm{E}(\Theta_k) \\
&+ \sum\sum_{i\neq j}\sum\sum_{l\neq k} q_i(1-q_i)q_j(1-q_j)q_k q_l \mathrm{E}(\Theta_i^2\Theta_j^2\Theta_k\Theta_l)
\end{aligned}
$$

The first two summations are zero because $\mathrm{E}(\Theta_j) = 0$. In the third summation, however, some terms have (for example) $i = k$ and $j = l$, yielding $\mathrm{E}(\Theta_i^3)\mathrm{E}(\Theta_j^3)$. It is straightforward, but somewhat tedious, to identify those terms, The result is

$$
B = 2\sum\sum_{i\neq j} q_i^2(1-q_i)q_j^2(1-q_j)M_{3i}M_{3j}
$$

$$
\begin{aligned}
C &= \mathrm{E}\{[\sum_i q_i(1-q_i)\Theta_i^2][\sum\sum_{k\neq j}\sum\sum_{m\neq l} q_j q_k q_l q_m \Theta_j\Theta_k\Theta_l\Theta_m]\} \\
&= \sum_i \sum\sum_{k\neq j}\sum\sum_{m\neq l} q_i(1-q_i)q_j q_k q_l q_m \mathrm{E}(\Theta_i^2\Theta_j\Theta_k\Theta_l\Theta_m)
\end{aligned}
$$

As in B, this summation contains some terms that do not vanish. Identifying those yields

$$
\begin{aligned}
C &= 4\sum\sum_{i\neq j} q_i^3(1-q_i)q_j^2 M_{4i}M_{2j} \\
&+2\sum\sum\sum_{i\neq j\neq k} q_i(1-q_i)q_j^2 q_k^2 M_{2i}M_{2j}M_{2k}
\end{aligned}
$$

$$
D = \mathrm{E}[\sum\sum_{i\neq j}\sum\sum_{k\neq l}\sum\sum_{m\neq n} q_i q_j q_k q_l q_m q_n \Theta_i\Theta_j\Theta_k\Theta_l\Theta_m\Theta_n]
$$

As above, removing the terms that vanish leaves

$$
\begin{aligned}
D &= 4\sum\sum_{i\neq j} q_i^3 q_j^3 M_{3i}M_{3j} \\
&+8\sum\sum\sum_{i\neq j\neq k} q_i^2 q_j^2 q_k^2 M_{2i}M_{2j}M_{2k}.
\end{aligned}
$$

Finally, assembling the four parts (with some simplification) yields

$$
\begin{aligned}
W^{-3}\mathrm{E}(Q^3) &= \textstyle\sum_i q_i^3(1-q_i)^3(M_{6i} - 3M_{4i}M_{2i} + 2M_{2i}^3) \\
&+3[\textstyle\sum_j q_j(1-q_j)M_{2j}][\sum_i q_i^2(1-q_i)^2(M_{4i} - M_{2i}^2)] \\
&+[\textstyle\sum_i q_i(1-q_i)M_{2i}]^3 \\
&-6\textstyle\sum\sum_{i\neq j} q_i^2(1-q_i)q_j^2(1-q_j)M_{3i}M_{3j} \\
&+12\textstyle\sum\sum_{i\neq j} q_i^3(1-q_i)q_j^2 M_{4i}M_{2j} \\
&+6\textstyle\sum\sum\sum_{i\neq j\neq k} q_i(1-q_i)q_j^2 q_k^2 M_{2i}M_{2j}M_{2k} \\
&-4\textstyle\sum\sum_{i\neq j} q_i^3 q_j^3 M_{3i}M_{3j} \\
&-8\textstyle\sum\sum\sum_{i\neq j\neq k} q_i^2 q_j^2 q_k^2 M_{2i}M_{2j}M_{2k}.
\end{aligned}
$$

A.4 | UNCONDITIONAL MOMENTS OF Θ

The unconditional moments of Θ_i for $\theta_i \sim N(\theta, \tau^2)$ are given by

$$M_{ri} = \mathrm{E}[(\hat{\theta}_i - \theta)^r] = \sum_{j=0}^{r} \binom{r}{j} \mathrm{E}[(\hat{\theta}_i - \theta_i)^j (\theta_i - \theta)^{r-j}] = \sum_{j=0}^{r} \binom{r}{j} \mathrm{E}[M^c_{ji}(\theta_i - \theta)^{r-j}], \qquad \text{(A.4)}$$

for conditional central moments $M^c_{ji} = \mathrm{E}[(\hat{\theta}_i - \theta_i)^j | \theta_i]$ with $M^c_{0i} = 1$ and $M^c_{2i} = v_i^2$. For unbiased estimators $\hat{\theta}_i$,
$M_{1i} = M^c_{1i} = 0$,
$M_{2i} = \mathrm{E}(v_i^2) + \tau^2$,
$M_{3i} = \mathrm{E}(M^c_{3i}) + 3\mathrm{E}(v_i^2(\theta_i - \theta))$,
$M_{4i} = \mathrm{E}(M^c_{4i}) + 4\mathrm{E}(M^c_{3i}(\theta_i - \theta)) + 6\mathrm{E}(v_i^2(\theta_i - \theta)^2) + 3\tau^4$,
$M_{5i} = \mathrm{E}(M^c_{5i}) + 5\mathrm{E}(M^c_{4i}(\theta_i - \theta)) + 10\mathrm{E}(M^c_{3i}(\theta_i - \theta)^2) + 10\mathrm{E}(v_i^2(\theta_i - \theta)^3)$,
$M_{6i} = \mathrm{E}(M^c_{6i}) + 6\mathrm{E}(M^c_{5i}(\theta_i - \theta)) + 15\mathrm{E}(M^c_{4i}(\theta_i - \theta)^2) + 20\mathrm{E}(M^c_{3i}(\theta_i - \theta)^3) + 15\mathrm{E}(v_i^2(\theta_i - \theta)^4) + 15\tau^6$.

A.5 | UNCONDITIONAL CENTRAL MOMENTS OF $\hat{\theta}$ FOR MEAN DIFFERENCE

Assume that each of K studies consists of two groups whose data are normally distributed with sample sizes n_{iC} and n_{iT} and means μ_{iC} and $\mu_{iT} = \mu_{iC} + \Delta_i$, and possibly different variances σ^2_{iC} and σ^2_{iT}. Then the mean difference Δ_i in Study i is estimated by

$$\hat{\Delta}_i = \bar{X}_{iT} - \bar{X}_{iC}, \qquad \text{(A.5)}$$

and its (conditional) variance $v_i^2 = \sigma^2_{iT}/n_{iT} + \sigma^2_{iC}/n_{iC}$. The conditional distribution of $\hat{\Delta}_i$ is $N(\Delta_i, v_i^2)$, so its odd central moments are zero, and its even moments are $M^c_{i,2r} = [(2r)!/(2^r r!)]v_i^{2r}$. As the conditional moments do not involve Δ_i, it is easy to write out the unconditional moments:
$M_{2i} = v_i^2 + \tau^2$,
$M_{4i} = 3v_i^4 + 6v_i^2\tau^2 + 3\tau^4$,
$M_{6i} = 15v_i^6 + 15 * 3v_i^4\tau^2 + 15v_i^2 * 3\tau^4 + 15\tau^6$. The first three moments of Q can be calculated by substituting these moments into Equations (3), (A.3), and the expression for the third moment.