the british psychological society
promoting excellence in psychology

British Journal of Mathematical and Statistical Psychology (2022)

published by John Wiley & Sons Ltd
on behalf of British Psychological Society

www.wileyonlinelibrary.com

On the Q statistic with constant weights for standardized mean difference

Ilyas Bakbergenuly[1], David C. Hoaglin[2] and Elena Kulinskaya[1]*

[1]School of Computing Sciences, University of East Anglia, Norwich, UK
[2]Department of Population and Quantitative Health Sciences, UMass Chan Medical School, Worcester, Massachusetts, USA

Cochran's Q statistic is routinely used for testing heterogeneity in meta-analysis. Its expected value is also used in several popular estimators of the between-study variance, τ^2. Those applications generally have not considered the implications of its use of estimated variances in the inverse-variance weights. Importantly, those weights make approximating the distribution of Q (more explicitly, Q_{IV}) rather complicated. As an alternative, we investigate a new Q statistic, Q_F, whose constant weights use only the studies' effective sample sizes. For the standardized mean difference as the measure of effect, we study, by simulation, approximations to distributions of Q_{IV} and Q_F, as the basis for tests of heterogeneity and for new point and interval estimators of τ^2. These include new DerSimonian–Kacker-type moment estimators based on the first moment of Q_F, and novel median-unbiased estimators. The results show that: an approximation based on an algorithm of Farebrother follows both the null and the alternative distributions of Q_F reasonably well, whereas the usual chi-squared approximation for the null distribution of Q_{IV} and the Biggerstaff–Jackson approximation to its alternative distribution are poor; in estimating τ^2, our moment estimator based on Q_F is almost unbiased, the Mandel – Paule estimator has some negative bias in some situations, and the DerSimonian–Laird and restricted maximum likelihood estimators have considerable negative bias; and all 95% interval estimators have coverage that is too high when $\tau^2 = 0$, but otherwise the Q-profile interval performs very well.

1. Introduction

When the individual studies assembled for a meta-analysis report means for their treatment and control arms, but those data are on different scales or come from different instruments, the customary measure of effect is the standardized mean difference (SMD). The SMD is considered to be the most appropriate effect-size index in psychological research (Sánchez-Meca & Marín-Martínez, 2010), and was also found to be more generalizable than the mean difference (Takeshima et al., 2014). In studying estimation of the overall effect in random-effects meta-analyses of SMD, we found that

Correspondence should be addressed to Elena Kulinskaya, School of Computing Sciences, University of East Anglia, Norwich Research Park, Norwich NR4 7TJ, UK (email: e.kulinskaya@uea.ac.uk).

DOI:10.1111/bmsp.12263

SSW, a weighted mean whose weights involve only the studies' arm-level sample sizes, performed well, avoiding shortcomings associated with estimators that use inverse-variance weights based on estimated variances (Bakbergenuly, Hoaglin, & Kulinskaya, 2020).

The present paper takes a natural further step by investigating a version of Cochran's Q statistic (Cochran, 1954) for assessment of heterogeneity that uses those constant weights. This work also draws on favourable results for Q with sample-size-based weights when the measure of effect is the mean difference (MD), which is less common but more tractable (Kulinskaya, Hoaglin, Bakbergenuly, & Newman, 2021). From this version of the Q statistic we also derive new point and interval estimators of the between-study variance, τ^2.

Simulation of the actual distribution of Q enables us to study the accuracy of approximations for the null distribution ($\tau^2 = 0$), the empirical level when $\tau^2 = 0$ and when $\tau^2 > 0$, the bias of point estimators of τ^2, and the coverage of confidence intervals for τ^2. For comparison we include the usual version of Q (based on inverse-variance weights) and familiar point and interval estimators of τ^2.

Section 2 briefly reviews study-level estimation of SMD. Section 3 reviews the random-effects model (REM) and describes the Q statistic. Section 4 introduces new point and interval estimators for τ^2. Section 5 discusses approximations to the distribution of Q. Section 6 describes the simulation design and summarizes the results. Section 7 provides an example of meta-analysis using SMD. Section 8 offers a summary and discussion. An Appendix gives the derivation of conditional and unconditional moments of Hedges's estimator of study-level SMD.

2. Study-level estimation of standardized mean difference

Consider a meta-analysis of K comparative studies, each consisting of two arms, treatment (T) and control (C), with sample sizes n_{iT} and n_{iC}. The total sample size in study i is $n_i = n_{iT} + n_{iC}$, and the ratio of the control sample size to the total is $f_i = n_{iC}/n_i$. The subject-level data in each arm are assumed to be normally distributed with means μ_{iT} and μ_{iC} and equal variances σ_i^2. The sample means are $\overline{x}_{ij}$, and the sample variances are s_{ij}^2, for $i = 1, \ \ldots, \ K$ and $j = C, \ T$.

The SMD effect measure is

$$\delta_i = \frac{\mu_{iT} - \mu_{iC}}{\sigma_i}.$$

The unbiased estimator of δ_i is Hedges's g, given by

$$g_i = J(m_i)\frac{\overline{x}_{iT} - \overline{x}_{iC}}{s_i}, \tag{1}$$

where the standard deviation, σ_i, is estimated by the square root of the pooled sample variance s_i^2, $m_i = n_{iT} + n_{iC} - 2$, and the factor $J(m) = \Gamma\left(\frac{m}{2}\right)/\sqrt{\frac{m}{2}}\Gamma\left(\frac{m-1}{2}\right)$ corrects for bias.

For the variance of g_i we use the unbiased estimator.

$$v_i^2 = \frac{n_{iT} + n_{iC}}{n_{iT} n_{iC}} + \left(1 - \frac{(m_i - 2)}{m_i J(m_i)^2}\right) g_i^2, \tag{2}$$

derived by Hedges (1983). The literature contains several other estimators of the variance of g_i and its biased counterpart, d_i. Lin and Aloe (2021) provide a comprehensive assessment.

Define $\tilde{n}_i = n_{iC} n_{iT} / n_i$, the effective sample size in study i. The sample SMD g_i has a scaled non-central t distribution with non-centrality parameter $\gamma_i = \tilde{n}_i^{1/2} \delta_i$ (Hedges & Olkin, 1985):

$$\sqrt{\tilde{n}_i} J(m_i)^{-1} g_i \sim t_{m_i}\left(\tilde{n}_i^{1/2} \delta_i\right). \tag{3}$$

3. Random-effects model and the Q statistic

We consider a generic REM. For study i ($i = 1, \ldots, K$) the estimate of the effect is $\hat{\theta}_i \sim G\left(\theta_i,\ v_i^2\right)$, 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 estimates of the true conditional effects θ_i, and the $v_i^2 = \mathrm{Var}(\hat{\theta}_i | \theta_i)$ are the true conditional variances.

Cochran's Q statistic is a weighted sum of the squared deviations of the estimated effects $\hat{\theta}_i$ from their weighted mean $\overline{\theta}_w = \sum w_i \hat{\theta}_i / \sum w_i$:

$$Q = \sum w_i (\hat{\theta}_i - \overline{\theta}_w)^2. \tag{4}$$

In Cochran (1954), w_i is the reciprocal of the *estimated* variance of $\hat{\theta}_i$. We denote this traditional version of Q with inverse-variance weights by Q_{IV}. In meta-analysis those w_i come from the fixed-effect model. In what follows, we examine the version of Q, discussed by DerSimonian and Kacker (2007) and further studied by Kulinskaya et al. (2021), in which the w_i are arbitrary positive constants. We denote this Q statistic with fixed weights by Q_{F}.

Define $W = \sum w_i$, $q_i = w_i / W$ and $\Theta_i = \hat{\theta}_i - \theta$. In this notation, and expanding $\overline{\theta}_w$, equation (4) can be written as

$$Q = W\left[\sum q_i (1 - q_i) \Theta_i^2 - \sum_{i \neq j} q_i q_j \Theta_i \Theta_j\right]. \tag{5}$$

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 $M_{2i} = E\left(\Theta_i^2\right) = \mathrm{Var}(\hat{\theta}_i) = E\left(v_i^2\right) + \tau^2$.

Under the above REM, it is straightforward to obtain the first moment of Q_{F} as.

$$E(Q_F) = W[\sum q_i (1 - q_i) \mathrm{Var}(\Theta_i)] = W\left[\sum q_i (1 - q_i)\left(E\left(v_i^2\right) + \tau^2\right)\right]. \tag{6}$$

This expression is similar to equation (4) in DerSimonian and Kacker (2007); they use the conditional variance v_i^2 instead of its unconditional mean $E(v_i^2)$.

Kulinskaya et al. (2021) also provide expressions for the second and third moments of Q_F, but these moments require higher moments of Θ, up to the sixth moment. For Hedges's g the expressions for these higher central moments are rather complicated; we provide them in the Appendix.

4. Point and interval estimators of τ^2

4.1. Point estimators

Rearranging the terms in equation (5) gives the moment-based estimator of τ^2:

$$\hat{\tau}^2_M = \frac{Q/W - \sum q_i(1-q_i)\hat{E}\left(v_i^2\right)}{\sum q_i(1-q_i)}. \tag{7}$$

DerSimonian and Kacker (2007) obtain a similar result; they use the conditional estimate, $\hat{v}_i^2$, instead of the unconditional estimate, $\hat{E}\left(v_i^2\right)$. For MD the two estimators are the same, because then $E\left(v_i^2\right) = v_i^2$. For SMD we study both estimators with effective-sample-size weights w. With the conditional estimated variances in equation (7), we denote the estimator by SSC; with the unconditional estimated variances, it is SSU.

The estimator $\hat{\tau}^2_M$ arose from setting the observed value of Q equal to its expected value and solving for τ^2. Instead of the expected value, one could use the median of the distribution of Q given τ^2. If the true (or approximate) cumulative distribution function is $F(\cdot|\tau^2)$, a point estimator of τ^2 can be found as

$$\hat{\tau}^2_{\text{med}} = \max\left(0,\ \{\tau^2 : F(Q|\tau^2) = 0.5\}\right).$$

In the Farebrother approximation to the distribution of Q (Section 5), one can use either the conditional estimated variances or the unconditional estimated variances. We denote the resulting estimators by SMC and SMU, respectively.

For comparison our simulations (Section 6) include four estimators that use inverse-variance weights: DerSimonian and Laird (1986) (DL), REML, Mandel and Paule (1970) (MP), and an estimator (KDB) based on the work of Kulinskaya, Dollinger, and Bjørkestøl (2011a) and discussed by Bakbergenuly et al. (2020). KDB uses an improved non-null first moment of Q and has better performance than most other estimators. In their review of methods for estimating the between-study variance, Veroniki et al. (2016) explain that DL is (by default) the most widely used, and they conclude that both REML and MP are better alternatives.

4.2. Interval estimators

Straightforward use of $F(\cdot|\tau^2)$ also yields a $100(1-\alpha)\%$ confidence interval for τ^2:

$$\{\tau^2 \geq 0 : F(Q|\tau^2) \in [\alpha/2,\ 1-\alpha/2]\}.$$

We use both the conditional estimated variances and the unconditional estimated variances in the Farebrother approximation to F; we refer to the resulting profile estimators as FPC and FPU. Jackson (2013) introduced a similar approach using conditional variances.

Our simulations (Section 6) also include the *Q*-profile (QP) interval (Viechtbauer, 2007b), the profile-likelihood (PL) interval (Hardy & Thompson, 1996), and the KDB interval, which is based on the chi-squared distribution with the corrected first moment developed by Kulinskaya, Dollinger, and Bjørkestøl (2011b).

5. Approximations to the distribution of Q

For meta-analysis of MD, Kulinskaya et al. (2021) considered the distribution of Q_{F}, a quadratic form in normal variables, which has the form $Q = \Theta^T A \Theta$ for a symmetric matrix A of rank $K - 1$. Because the vector Θ has a multivariate normal distribution, $N(\mu,\ \Sigma)$, the distribution of Q can be obtained by the algorithm of Farebrother (1984) (after determining the eigenvalues of $A\Sigma$ and some other inputs). If the variances in Σ are the true variances, Farebrother's algorithm evaluates the exact distribution of Q. In practice (as in our simulations), it is necessary to plug in estimated variances. Encouragingly, the resulting approximation is quite accurate for MD. Kulinskaya et al. (2021) also considered a two-moment approximation and a three-moment approximation. The three-moment approximation regularly encountered numerical problems, so we do not include it here.

For SMD, Q_{F} is a quadratic form in t variates. The Farebrother algorithm may provide a satisfactory approximation, especially for larger sample sizes. To apply it, we again plug in estimated variances. We investigate the quality of that approximation, which we denote by F SW, and the two-moment approximation (M2 SW), which is based on the gamma distribution.

The null distribution of Q_{IV} is usually approximated by the chi-squared distribution with $K - 1$ degrees of freedom. For MD and SMD, however, this approximation is not accurate for small sample sizes (Viechtbauer, 2007a). For SMD, Kulinskaya, Dollinger, and Bjørkestøl (2011a) provided an improved approximation to the null distribution of Q_{IV} based on a chi-squared distribution with degrees of freedom equal to the estimate of the corrected first moment; we denote this approximation by KDB. Biggerstaff and Jackson (2008) used the Farebrother approximation to the distribution of a quadratic form in normal variables as the 'exact' distribution of Q_{IV}. We denote this approximation by BJ. Jackson, Turner, Rhodes, and Viechtbauer (2014) extended this approach to a Q with arbitrary weights in a meta-regression setting. When $\tau^2 = 0$, the BJ approximation to the distribution of Q_{IV} is the χ^2_{K-1} distribution. For comparison, our simulations include these three approximations.

6. Simulation design and results

6.1. Simulation design

Our simulation design follows that described in Bakbergenuly et al. (2020). Briefly, we varied five parameters: the overall true SMD (δ), the between-studies variance (τ^2), the number of studies (K), the studies' total sample size (n and $\overline{n}$) and the proportion of observations in the control arm (f). Table 1 lists the values of each parameter.

The values of δ (0, 0.2, 0.5, 1, 2) aim to represent the range containing most values encountered in practice. Our choices align well with the results of Rubio-Aparicio et al. (2018), who reanalysed the data of 41 meta-analyses on the effectiveness of clinical psychology treatments. Their pooled estimates of δ ranged from 0.068 to 1.075, with a median at 0.409; many of those meta-analyses included study-level estimates of δ that were somewhat larger, occasionally exceeding 4. An illustrative meta-analysis of the efficacy of

Table 1. Values of parameters in the simulations for Q with constant weights and SMD as the measure of effect

Parameter	Equal study sizes	Unequal study sizes
K (number of studies)	5, 10, 30	5, 10, 30
n or $\overline{n}$ (average size of individual study) – total of the two arms	20, 40, 100, 250	30 (12,16,18,20,84)
	30, 50, 60, 70	60 (24,32,36,40,168)
For $K = 10$ and $K = 30$, the same set of unequal study sizes is used twice or six times, respectively		100 (64,72,76,80,208)
		160 (124,132,136,140,268)
f (proportion of each study in the control arm)	1/2, 3/4	1/2, 3/4
δ(true value of the SMD)	0, 0.2, 0.5, 1, 2	0, 0.2, 0.5, 1, 2
τ^2(variance of random effects)	0, 0.5, 1, 1.5, 2, 2.5	0, 0.5, 1, 1.5, 2, 2.5

treatments for obsessive-compulsive disorder (Sánchez-Meca & Marín-Martínez, 2010) involved 24 studies; half of the study-level SMDs were less than 1 in magnitude. For the social sciences more broadly, Ferguson (2009) proposed 0.41, 1.15 and 2.70 as benchmarks for small, medium and large effects.

The values of τ^2 (0, 0.5, 1, 1.5, 2, 2.5) systematically cover a reasonable range. We know little about actual values of τ^2 for SMD, primarily because accurate estimation of variances requires quite large samples. Rubio-Aparicio et al. reported estimates of τ^2 from 0 to 0.789.

The numbers of studies ($K = 5$, 10, 30) reflect the sizes of many meta-analyses and have yielded valuable insights in previous work. Rubio-Aparicio et al. (2018) reported numbers of studies ranging from 7 (their minimum for inclusion) to 70, with 28 of the 41 between 10 and 24.

In practice, many studies' total sample sizes fall in the ranges covered by our choices ($n = 20,\ 40,\ 100,\ 250$ when all studies have the same n, supplemented by 30, 50, 60, 70; and $\overline{n} = 30,\ 60,\ 100,\ 160$ when sample sizes vary among studies). The choices of $\overline{n}$ follow a suggestion of Sánchez-Meca and Marín-Martínez (2000), who constructed the studies' sample sizes to have skewness 1.464, which they regarded as typical in behavioural and health sciences. The meta-analyses studied by Rubio-Aparicio et al. had median study-level sample sizes from 16 to 87.5; within those meta-analyses, the sample sizes varied substantially.

Many studies allocate subjects equally to the two groups ($f = 1/2$), and rough equality holds more widely (as in the studies analysed by Rubio-Aparicio et al.). Unequal allocations, either planned or observed, are not uncommon. To investigate potential impacts of such situations, we also used $f = 3/4$, a substantial departure from equality.

We generated the true effect sizes δ_i from a normal distribution: $\delta_i \sim N(\delta,\ \tau^2)$. We generated the values of Hedges's estimator g_i directly from the appropriately scaled non-central t distribution, given by equation (3). We used a total of 10,000 repetitions for each combination of parameters.

R statistical software (R Core Team, 2016) was used for simulations. The user-friendly R programs implementing our methods and analysing the example in Section 7 are available at https://osf.io/3gytv.

6.2. *Simulation results*

In tests based on either version of Q, heterogeneity corresponds to large values of Q. Thus, we focused on the upper tail of the distribution. For each configuration of parameters and for each generated value of Q, we used each approximation to calculate 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 $\hat{p} = \#(\tilde{p} < p)/10,\ 000$ at p =.001,.0025,.005,.01,.025,.05,.1,.25,.5 and, for completeness, the complementary values.75,. . .,.999. Thus, $\hat{p}$ estimates the upper-tail probability $P(\hat{F}(Q) > 1 - p)$.

The values of τ^2 included both null ($\tau^2 = 0$) and non-null ($\tau^2 > 0$) values (Table 1). 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 probability–probability (P-P) plots (vs. the true null distribution) for two approximations to the distribution of Q with effective-sample-size weights (F SW and M2 SW) and two approximations to the distribution of Q with IV weights (chi-squared/BJ and KDB (for $\tau^2 = 0$ only)) and for estimating their null levels and their non-null empirical tail areas. We also estimate the bias of eight point

estimators of τ^2 (DL, REML, MP, KDB, SSC, SSU, SMC and SMU) and the coverage of five interval estimators of τ^2 (QP, PL, KDB, FPC and FPU).

Our full simulation results are available as an arXiv e-print (Bakbergenuly, Hoaglin, & Kulinskaya, 2021).

6.3. P-P plots for $\tau^2 = 0$

To compare an approximation for a distribution function of Q against the theoretical distribution function, we use P-P plots (Wilk & Gnanadesikan, 1968). Evaluating two distribution functions, F_1 and F_2, to obtain upper-tail probabilities at x yields $p_1 = 1 - F_1(x)$ and $p_2 = 1 - F_2(x)$. In the usual plot of p_2 versus p_1, equality of the two distributions corresponds to the line $p_2 = p_1$. To make departures from that reference pattern more visible, we flatten the plot by subtracting the line; that is, we plot $p_2 - p_1$ versus p_1.

When $\delta = 0$, P-P plots show that the KDB and F SW approximations perform reasonably well for small sample sizes, but the χ^2_{K-1} and M2 SW approximations do not (Figure 1). This difference is especially pronounced for larger K values. It appears that KDB performs better overall for small K, and F SW for large K. All four approximations perform reasonably well for $n \geq 100$. When δ increases, the performance of KDB deteriorates somewhat in the upper half, though it may be somewhat better than F SW in the lower half. Both the χ^2_{K-1} and M2 SW approximations deteriorate further. Results are similar for equal and unequal sample sizes.

6.4. Empirical levels when $\tau^2 = 0$

To better visualize the quality of the approximations as the basis for a test for heterogeneity at the 0.05 level, we plot their empirical levels under the null $\tau^2 = 0$ against sample size. Figure 2 presents typical results at the 0.05 level. Figure 3 depicts the quality of the approximations at the 0.95 level.

For small sample sizes, the error rate of the test based on F SW somewhat exceeds the nominal 5% (up to 6%), and the error rate of the test based on KDB is somewhat low (between 4% and 3%). Both the χ^2_{K-1} and M2 SW approximations result in tests with error rates that are noticeably too low (in that order). For all approximations, departures from the nominal level increase for larger K and larger δ, especially when sample sizes are unequal. The χ^2_{K-1} approximation has empirical levels of about 2.5% (vs. the nominal 5%) when $K = 30$. For unequal sample sizes or unbalanced samples, the results are similar. The chi-squared approximation provides reasonable results by $n = 100$.

The picture is similar in the lower tail. All approximations except M2 SW, which produces extremely high empirical levels when n is small, work well for $K = 5$. However, increasing values of K and, to a lesser degree, of δ, result in decreasing empirical levels for χ^2_{K-1} (down to 92%) and hence larger error rates, and, to a lesser degree, for F SW (to 93.5%) when $K = 30$. KDB exhibits the best performance in the lower tail.

6.5. Empirical levels when $\tau^2 > 0$

To understand how the approximations behave as τ^2 increases, we plot the empirical p-values ($\hat{p}$) versus τ^2 for the F SW, M2 SW and BJ approximations for the nominal level 0.05 (Figure 4). F SW provides robust though somewhat high (for larger K) levels at all values

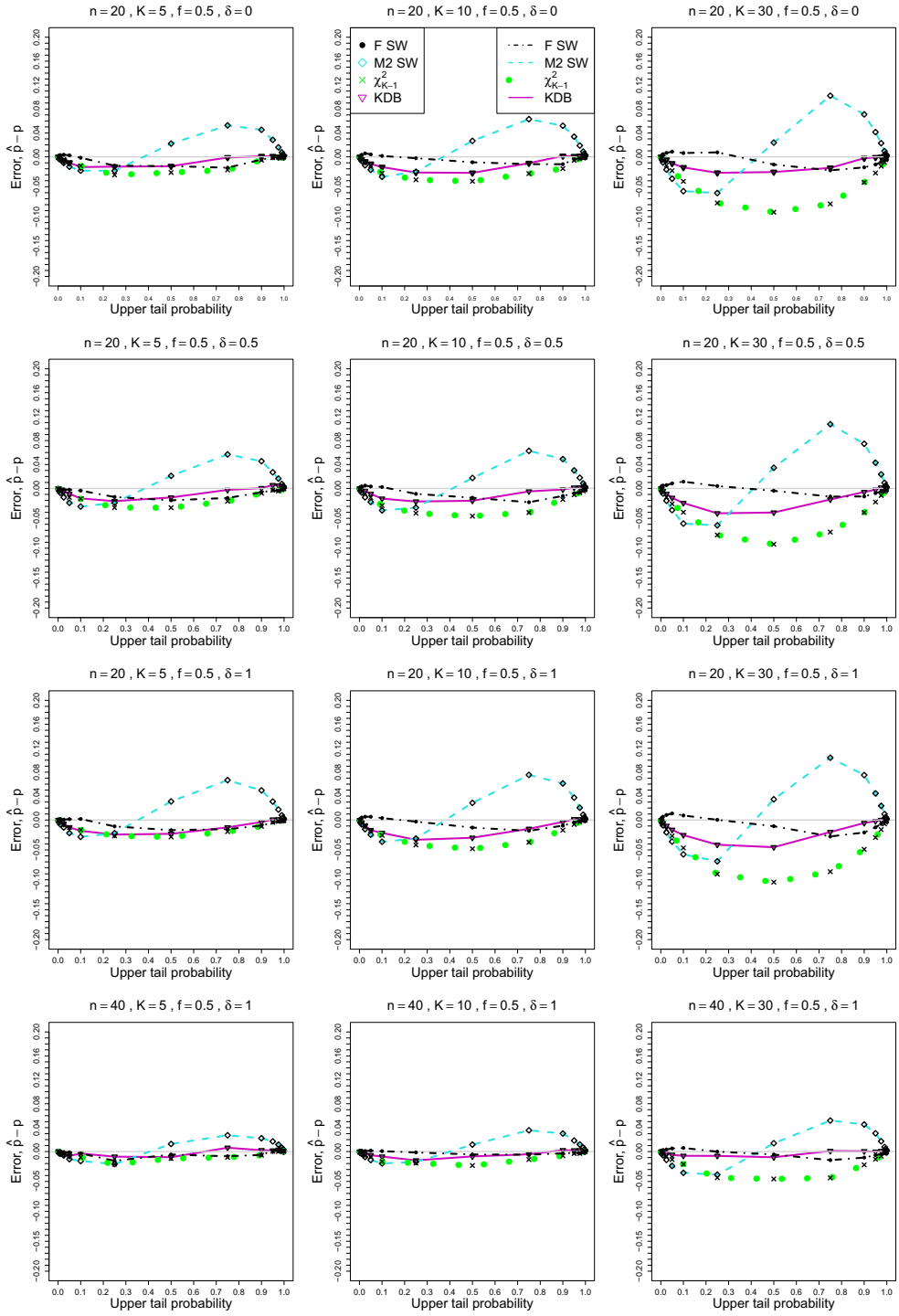

Figure 1. Flattened P-P plots of upper-tail probabilities for the Farebrother and M2 approximations to the null distribution of Q with sample-size-based weights, and for the chi-squared and KDB approximations to the null distribution of Q with IV-based weights. First three rows: equal sample sizes, $n = 20$, $f = 0.5$, $\delta = 0,\ 0.5,\ 1$. Fourth row: $n = 40$, $\delta = 1$, $f = 0.5$

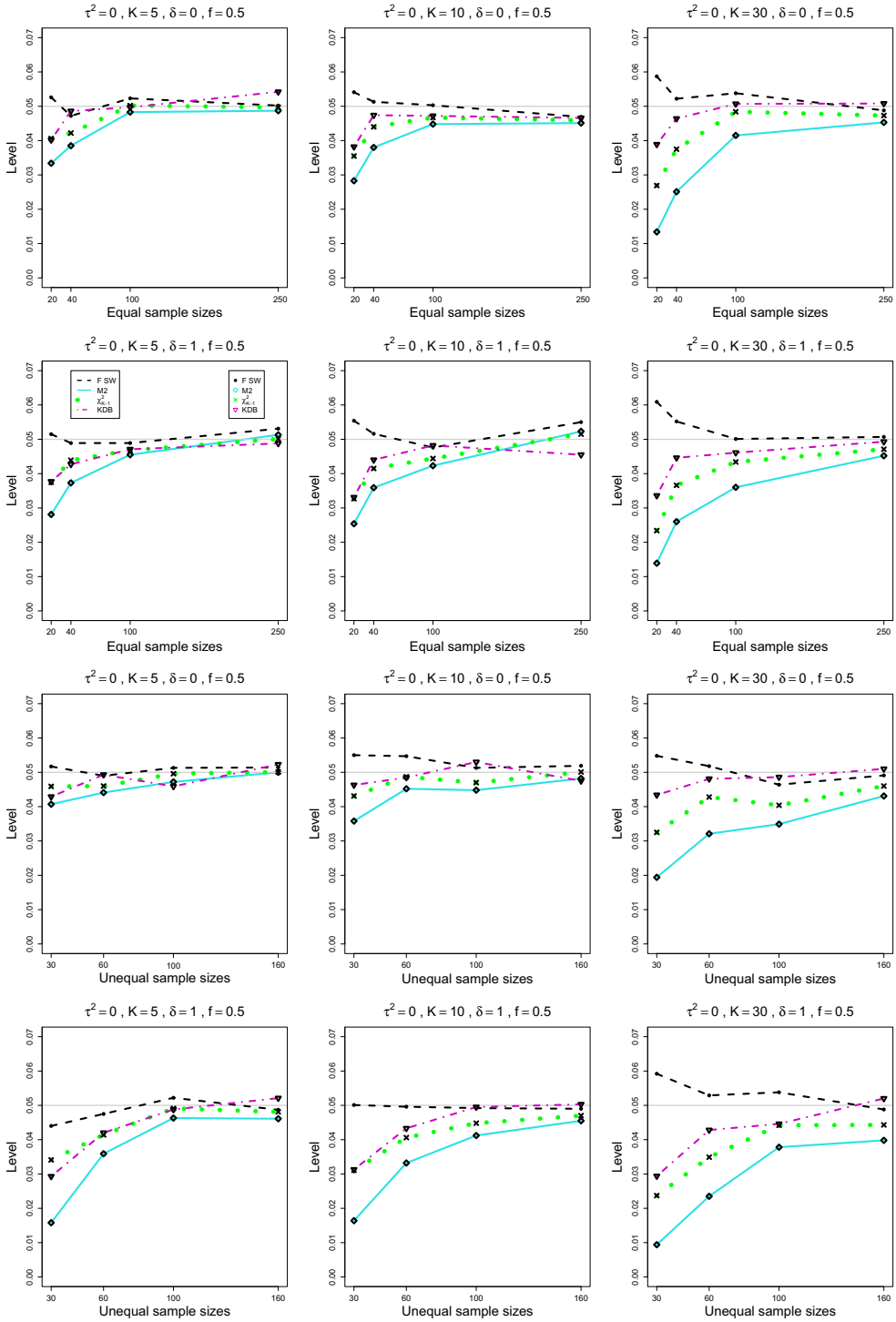

Figure 2. Empirical levels of approximations to the null distribution of Q with sample-size-based or IV weights at nominal 0.05 level against sample size n. In all plots, $\tau^2 = 0$ and $f = 0.5$. Top two rows: equal sample sizes, $\delta = 0$ and $\delta = 1$. Bottom two rows: unequal sample sizes, $\delta = 0$ and $\delta = 1$

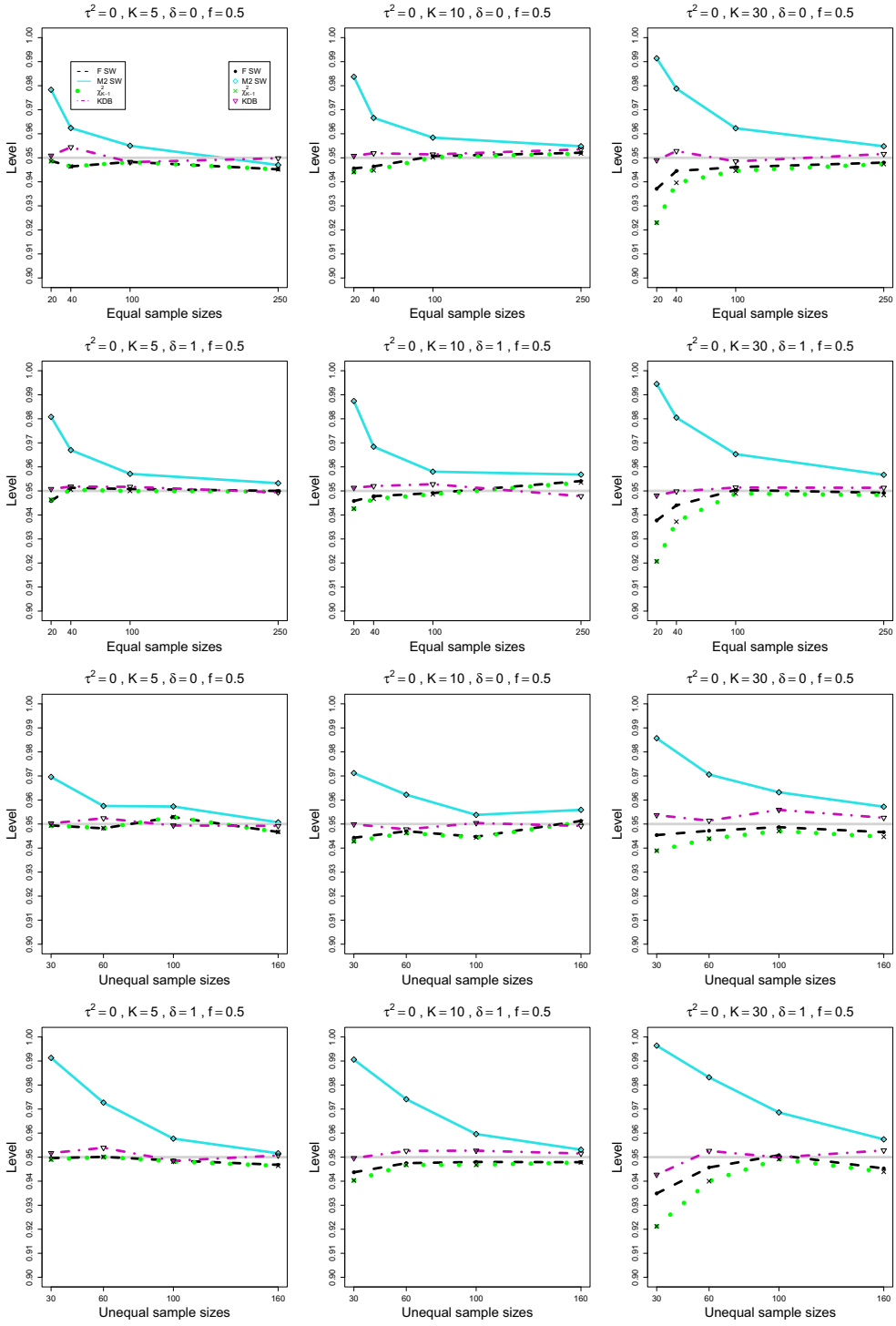

Figure 3. Empirical levels of approximations to the null distribution of Q with sample-size-based or IV weights at nominal 0.95 level against sample size n. In all plots, $\tau^2 = 0$ and $f = 0.5$. Top two rows: equal sample sizes, $\delta = 0$ and $\delta = 1$. Bottom two rows: unequal sample sizes, $\delta = 0$ and $\delta = 1$

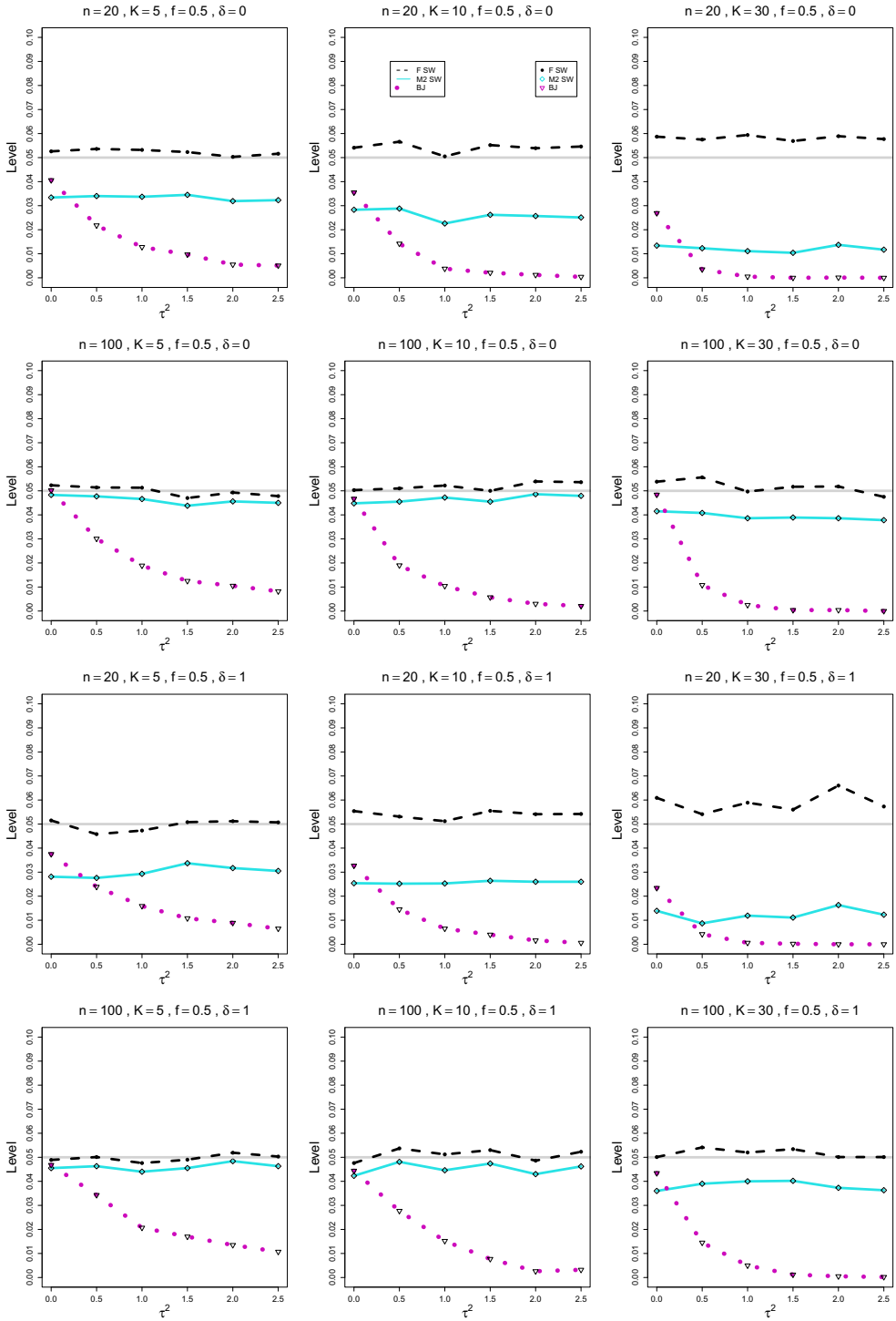

Figure 4. Empirical p-values of approximations to the distribution of Q with sample-size-based or IV weights at the nominal 0.05 level against between-study variance τ^2. All plots have equal sample sizes and $f = 0.5$. Top two rows: $\delta = 0$, $n = 20$ and $n = 100$. Bottom two rows: $\delta = 1$, $n = 20$ and $n = 100$

of τ^2 and δ. This is also true for unequal sample sizes and unbalanced studies. M2 SW results in lower error rates; its level decreases further for larger δ but does not depend on τ^2. The BJ approximation has even lower error rates, and it deteriorates further as τ^2 increases.

6.6. Bias in estimation of τ^2

Here we compare eight point estimators of τ^2: three well-known estimators (DL, REML, MP), the less-well-known KDB, and four new estimators (SSC, SSU, SMC, SMU). Figure 5 depicts the biases of the eight estimators for small sample sizes.

SSC is the best estimator overall; it is almost unbiased under all studied conditions even for very small and unbalanced sample sizes (Figure 5). DL is clearly the worst; it has considerable negative bias, which increases in τ^2. REML is the second worst; its bias is similar to DL but less pronounced. MP is the best of the established estimators; this agrees with the recommendation of Veroniki et al. (2016). It is also negatively biased for larger K and τ^2 values, but not by much. The bias of SSU is similar to that of MP for $K = 5$, and it is smaller than that of MP for larger values of K and δ, though it is larger than that of SSC. This makes sense, as the unconditional variance is calculated from averages over K values, so it is estimated more precisely for larger K. Estimators SMC and SMU are positively biased; their bias increases in τ^2 but decreases in K. SMU is less biased than SMC. By design, these two estimators are expected to be median-unbiased; we discuss them further in Section 8. Finally, KDB is somewhat positively biased, and its bias increases in τ^2. We recommended MP and KDB in our previous work (Bakbergenuly et al., 2020), and our current results agree with the previous ones.

To summarize, SSC provides very precise estimation of τ^2 and should be exclusively used in practice.

6.7. Coverage in interval estimation of τ^2

Here we compare the coverage of five interval estimators of τ^2 (QP, PL, KDB, FPC, FPU) at the 95% nominal level of confidence. Figure 6 depicts the coverage of the five estimators for small sample sizes.

All interval estimators have coverage that is too high for $\tau^2 < 0.5$. For larger values of τ^2, QP and KDB typically have somewhat excessive coverage (KDB more so than QP), and PL, FPC and FPU typically have somewhat deficient coverage. Coverage of FPC is very slightly above that of FPU. The coverage of these two new estimators is close to nominal when $K = 5$ and $\tau^2 \geq 0.5$, but it decreases to about 94% for $K = 30$. Coverage of PL may be erratic, especially for small K, even for large sample sizes, and we do not recommend its use.

Overall, QP performs quite impressively, and we recommend its use in practice. This finding is counter-intuitive, as we saw previously that the χ^2_{K-1} approximation does not hold levels well. However, it works for confidence levels. The explanation is that the confidence intervals provided by QP are not symmetrical for small n, especially for large values of K and δ.

To explain this point, consider a QP confidence interval at the 90% level. In the discussion of empirical levels in Section 6.4 and Figures 2 and 3 for levels 0.05 and 0.95, we noted that when $K = 30$, the χ^2_{K-1} approximation has empirical level about 2.5% at the nominal 5% level and empirical level about 92% at the nominal 95% level. The QP 90%

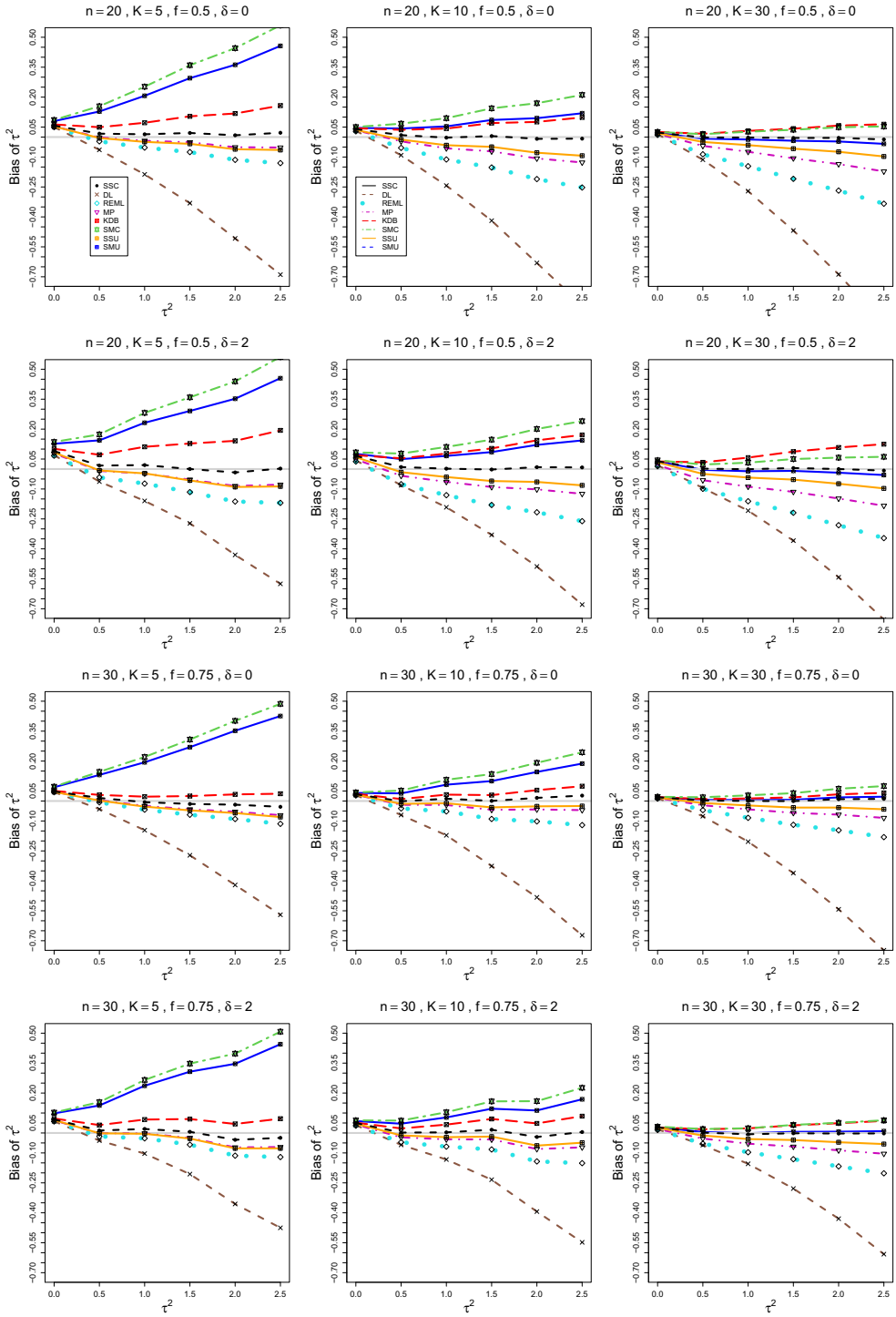

Figure 5. Bias in estimation of between-study variance τ^2 by eight methods: DL, REML, MP, KDB, SSC, SSU, SMC and SMU. First two rows: equal sample sizes, $n = 20$, $f = 0.5$, $\delta = 0$ and $\delta = 2$. Second two rows: unequal sample sizes, $\overline{n} = 30$, $f = 0.75$, $\delta = 0$ and $\delta = 2$

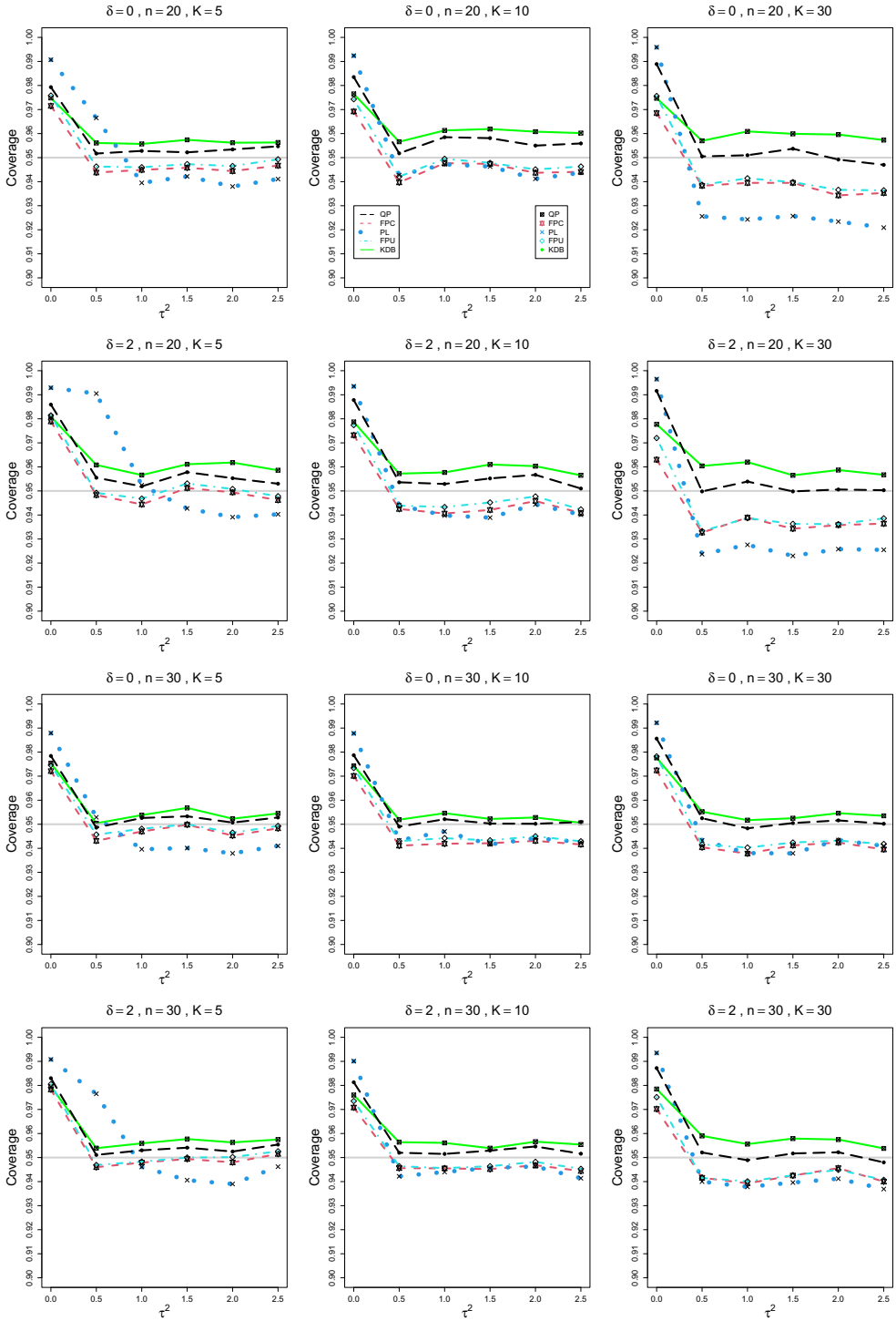

Figure 6. Coverage at 95% nominal level of confidence of five interval estimators of between-study variance τ^2: QP, PL, KDB, FPC and FPU. First two rows: equal sample sizes, $n = 20$, $f = 0.5$, $\delta = 0$ and $\delta = 2$. Second two rows: unequal sample sizes, $\overline{n} = 30$, $f = 0.75$, $\delta = 0$ and $\delta = 2$

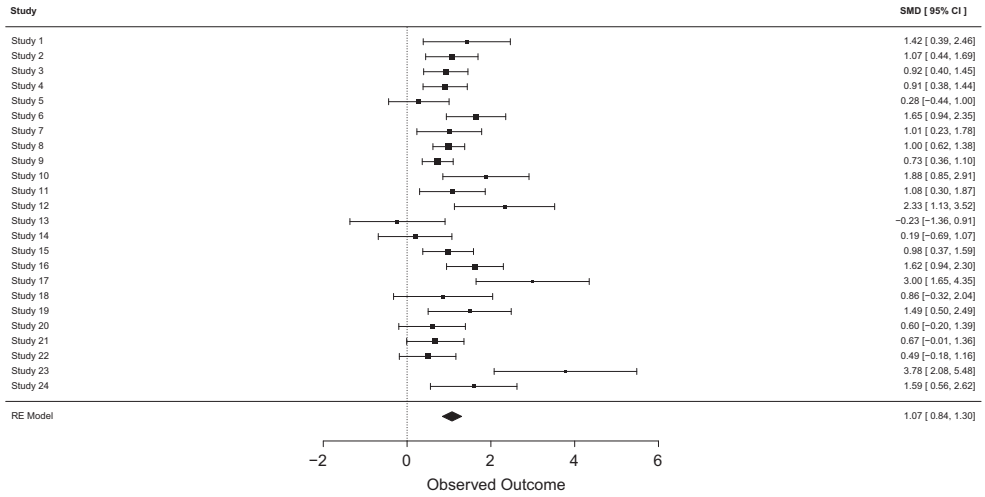

Figure 7. Forest plot for meta-analysis of the data from Sánchez-Meca and Marín-Martínez (2010) on the efficacy of psychological treatments for obsessive-compulsive disorder. The estimate of the overall effect obtained by using REML is included for illustration

confidence interval consists of $\{\tau^2 : q_{0.95} \geq Q_{w(\hat{\tau}^2)} \geq q_{0.05}\}$, where the q_α are critical values of χ^2_{K-1} and $w(\hat{\tau}^2)$ are the IV weights $w(\hat{\tau}^2) = (v_i^2 + \hat{\tau}^2)^{-1}$. This interval would have about 8% probability below the lower limit and about 2.5% above the upper limit. The results for levels .025 and .975 (not shown) also show non-symmetric patterns similar to those in Figures 2 and 3. QP intervals at the 95% level can have about 4% in the left tail, and 1% in the right tail.

7. Example

We use data, previously considered by (Sánchez-Meca & Marín-Martínez, 2010) and subsequently by Bakbergenuly et al. (2020), on the efficacy of psychological treatments for obsessive-compulsive disorder. These data consist of 24 trials with mostly small sample sizes, ranging from 12 to 121 patients. The effect measure is SMD, and positive values correspond to lower levels of obsessions and compulsions in the treatment group. The data appear in table 4 of Bakbergenuly et al. (2020), and our Figure 7 shows a forest plot.

Heterogeneity in these data is rather high. The value of Q_{IV} is 53.45 resulting in a p-value of .00032 for the χ^2_{K-1} approximation and a p-value of .00010 for the KDB approximation. The value 83.44 for Q_{F} results in a p-value of .00003 for F SW and in a much higher p-value, .0139 for the two-moment approximation, M2 SW. Table 4 of Bakbergenuly et al. (2020) includes the year of the study (one in 1980 and the rest from 1993 to 2006) and whether the study design was experimental or quasi-experimental (four studies). The values of g_i, however, are not systematically related to either of these variables. Further examination of potential sources of heterogeneity would consider details of the studies' designs and variation among the means of the control arms.

Table 5 of Bakbergenuly et al. (2020) shows results for several point and interval estimators of τ^2 that we also consider here, and for the corresponding estimators of δ, with $\hat{\delta}$ values from 1.07 to 1.12. Our Table 2 includes all estimators of τ^2 in our simulation

Table 2. Point and confidence-interval estimates for the heterogeneity parameter τ^2 in the example of efficacy of psychological treatments for obsessive-compulsive disorder

Method	$\widehat{\tau}^2$	L	U	Length of CI
DL&QP	0.1697	0.0992	1.1002	1.0010
REML&PL	0.1622	0	0.6029	0.6029
MP&QP	0.3722	0.0992	1.1002	1.0010
KDB&KDB	0.4539	0.2167	0.9033	0.6866
SMC&FPC	0.2879	0.0997	0.7415	0.6418
SMU&FPU	0.2671	0.0844	0.7056	0.6212
SSC	0.2698			
SSU	0.2504			

Note L and U denote the lower and upper limits of the 95% confidence interval.

study. For comparison, our simulations include data patterns with $K \geq 20$, $\delta = 1$ and $\tau^2 \leq 0.5$.

Point estimates of τ^2 are lowest for REML and DL, in agreement with our simulations (Figure 5). All our new estimators give rather similar values. SSC, at 0.2698, is somewhat higher than SSU, at 0.2504; and SMC, at 0.2879, is somewhat higher than SMU. KDB is highest, also in agreement with our simulations. MP is unexpectedly high, but this may be due to the comparatively low value of τ^2, as all estimators have positive bias at $\tau^2 = 0$.

The lower limits of the QP, FPC and FPU confidence intervals are rather similar, whereas PL is lowest, at 0, and KDB is highest, at 0.2167. The lengths of the confidence intervals are also rather similar, except for QP, which is widest at 1.001. This may be due to the shift of the upper limit of the QP interval, further into the right tail, as discussed in Section 6.7.

8. Discussion

The Q statistic serves as the basis for two main steps in random-effects meta-analysis: testing for the presence of heterogeneity and estimating the between-study variance. In its customary form, with inverse-variance weights based on estimated fixed-effect variances, Q_{IV} contributes to a variety of shortcomings. Encouraged by the favourable performance, with the mean difference as the measure of effect, of a version, Q_{F}, whose weights are based only on the studies' arm-based sample sizes, we studied key features of its performance when the measure of effect is the SMD. Aspects of performance included accuracy of approximations for the distribution of Q_{F} (or Q_{IV}), empirical levels when $\tau^2 = 0$ and when $\tau^2 > 0$, bias of point estimators of τ^2, and coverage of interval estimators for τ^2. On most of these aspects, Q_{F} and related estimators performed better than Q_{IV} and their other counterparts.

The P-P plots show that the Farebrother approximation (F SW) usually comes close to the actual null distribution of Q_{F}, much closer than the two-moment approximation. This result should not be surprising, because F SW makes more detailed use of the study-level variances than M2. It is encouraging, because F SW assumes that the variables in the quadratic form have normal distributions, instead of the actual t distributions. For Q_{IV}, the KDB approximation is consistently better than χ^2_{K-1}, especially as K increases when n is small. Having the correct first moment of the null distribution can make a big

difference. It may be possible to improve both approximations, using data from simulations on the empirical distributions of Q_F and Q_{IV}. We aim to pursue that topic in further work.

The new SSC estimator of τ^2 provides very precise estimation of τ^2, and we recommend its exclusive use in practice. We also introduced two new estimators of τ^2, SMC and SMU, by solving for the value that equates the observed value of Q with the median of F SW. Both have positive bias. When $K = 5$, that bias grows rapidly as τ^2 increases. When $K = 10$, however, the growth is much less rapid; and when $K = 30$, SMU has little bias. The bias of these estimators is due, in large part, to the difference between the mean and the median of the (skewed) distribution of Q_F. We infer that the skewness decreases for larger K, similarly to the skewness of the χ^2_{K-1} distribution, which approximates the distribution of Q for large n. We expect both estimators to be almost median-unbiased, and we intend to check this in future work. If median unbiasedness holds, SMC and SMU would be the first estimators of τ^2 with this property. Whether mean unbiasedness or median unbiasedness is more important is a moot point.

So far, we have demonstrated that the Q statistic with constant weights is useful in testing for and estimating heterogeneity when the effect measure is the mean difference or standardized mean difference. Its usefulness in meta-analysis of binary outcomes such as the log-odds ratio or risk difference is less clear. In these cases the Farebrother approximation seems less likely to provide a good fit to the distribution of Q_F. We intend to investigate binary outcomes further in future research.

Acknowledgements

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

Conflicts of interest

All authors declare no conflict of interest.

Author contributions

Ilyas Bakbergenuly: Methodology (equal); Software (equal); Visualization (equal); Writing – review & editing (equal). **David C Hoaglin:** Conceptualization (equal); Methodology (equal); Writing – original draft (equal); Writing – review & editing (equal). **Elena Kulinskaya:** Conceptualization (equal); Funding acquisition (equal); Methodology (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

Data availability statement

Our full simulation results are available as an arXiv e-print (arXiv:2103.03272v1). The user-friendly R program implementing the QF test for heterogeneity in meta-analysis of SMD and the SSC, SSU, SMC and SMU estimators of $\tau 2$ with related confidence intervals is available at https://osf.io/3gytv. The file also includes the example discussed in Section 7.

References

Bakbergenuly, I., Hoaglin, D. C., & Kulinskaya, E. (2020). Estimation in meta-analyses of mean difference and standardized mean difference. *Statistics in Medicine*, *39*, 171–191. https://doi.org/10.1002/sim.8422

Bakbergenuly, I., Hoaglin, D. C. & Kulinskaya, E. (2021). Simulation study of Q statistic with constant weights for testing and estimation of heterogeneity of standardized mean differences in meta-analysis. eprint arXiv:2103.03272v1 [stat.ME]. https://arxiv.org/abs/2103.03272

Biggerstaff, B. J., & Jackson, D. (2008). The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. *Statistics in Medicine*, *27*, 6093–6110. https://doi.org/10.1002/sim.3428

Cochran, W. G. (1954). The combination of estimates from different experiments. *Biometrics*, *10*, 101–129. https://doi.org/10.2307/3001666

DerSimonian, R., & Kacker, R. (2007). Random-effects model for meta-analysis of clinical trials: An update. *Contemporary Clinical Trials*, *28*, 105–114.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Cinical Trials*, 7, 177–188. https://doi.org/10.1016/0197-2456(86)90046-2

Farebrother, R. W. (1984). Algorithm AS 204: The distribution of a positive linear combination of χ^2 random variables. *Journal of the Royal Statistical Society, Series C*, *33*, 332–339. https://doi.org/10.2307/2347721

Ferguson, C. J. (2009). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research & Practice*, *40*, 532–538. https://doi.org/10.1037/a0015808

Hardy, R. J., & Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in Medicine*, *15*, 619–629. https://doi.org/10.1002/(SICI)1097-0258(19960330)15:6<619:AID-SIM188>3.0.CO;2-A

Hedges, L. V. (1983). A random effects model for effect sizes. *Psychological Bulletin*, *93*, 388–395. https://doi.org/10.1037/0033-2909.93.2.388

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.

Jackson, D. (2013). Confidence intervals for the between-study variance in random effects meta-analysis using generalised Cochran heterogeneity statistics. *Research Synthesis Methods*, *4*, 220–229. https://doi.org/10.1002/jrsm.1081

Jackson, D., Turner, R., Rhodes, K., & Viechtbauer, W. (2014). Methods for calculating confidence and credible intervals for the residual between-study variance in random effects meta-regression models. *BMC Medical Research Methodology*, *14*, 103. https://doi.org/10.1186/1471-2288-14-103

Johnson, N. L., Kotz, S., & Balakrishnan, N. (1995). *Continuous univariate distributions* (Vol. *2*., 2nd ed.). Hoboken, NJ: Wiley.

Kulinskaya, E., Dollinger, M. B., & Bjørkestøl, K. (2011a). On the moments of Cochran's Q statistic under the null hypothesis, with application to the meta-analysis of risk difference. *Research Synthesis Methods*, *2*, 254–270. https://doi.org/10.1002/jrsm.54

Kulinskaya, E., Dollinger, M. B., & Bjørkestøl, K. (2011b) Testing for homogeneity in meta-analysis I. The one-parameter case: Standardized mean difference. *Biometrics*, *67*, 203–212. https://doi.org/10.1111/j.1541-0420.2010.01442.x

Kulinskaya, E., Hoaglin, D. C., Bakbergenuly, I., & Newman, J. (2021). A *Q* statistic with constant weights for assessing heterogeneity in meta-analysis. *Research Synthesis Methods*, *12*, 711–730. https://doi.org/10.1002/jrsm.1491

Lin, L., & Aloe, A. M. (2021). Evaluation of various estimators for standardized mean difference in meta-analysis. *Statistics in Medicine*, *40*, 403–426. https://doi.org/10.1002/sim.8781

Mandel, J., & Paule, R. C. (1970). Interlaboratory evaluation of a material with unequal numbers of replicates. *Analytical Chemistry*, *42*, 1194–1197. https://doi.org/10.1021/ac60293a019

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Rubio-Aparicio, M., López-López, J. A., Sánchez-Meca, J., Marín-Martínez, F., Viechtbauer, W., & Van den Noortgate, W. (2018). Estimation of an overall standardized mean difference in random-effects meta-analysis if the distribution of random effects departs from normal. *Research Synthesis Methods*, *9*, 489–503 (2018). https://doi.org/10.1002/jrsm.1312

Sánchez-Meca, J., & Marin-Martnez, F. (2000). Testing the significance of a common risk difference in meta-analysis. *Computational Statistics & Data Analysis*, *33*, 299–313. https://doi.org/10.1016/S0167-9473(99)00055-9

Sánchez-Meca, J., & Marin-Martnez, F. (2010). Meta-analysis in psychological research. *International Journal of Psychological Research*, *3*, 150–162. https://doi.org/10.21500/20112084.860

Takeshima, N., Sozu, T., Tajika, A., Ogawa, Y., Hayasaka, Y., & Furukawa, T. A. (2014). Which is more generalizable, powerful and interpretable in meta-analyses, mean difference or standardized mean difference? *BMC Medical Research Methodology*, *14*, 30. https://doi.org/10.1186/1471-2288-14-30

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., … Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods*, *7*, 55–79. https://doi.org/10.1002/jrsm.1164

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

Viechtbauer, W. (2007b). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, *26*, 37–52. https://doi.org/10.1002/sim.2514

Wilk, M. B., & Gnanadesikan, R. (1968). Probability plotting methods for the analysis of data. *Biometrika*, *55*, 1–17. https://doi.org/10.2307/2334448

Received 8 March 2021

Appendix :

Derivation of the moments of Hedges's g

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

$$M_{ri} = E[(\hat{\theta}_i - \theta)^r] = \sum_{j=0}^{r} \binom{r}{j} E[(\hat{\theta}_i - \theta_i)^j (\theta_i - \theta)^{r-j}] = \sum_{j=0}^{r} \binom{r}{j} E[M_{ji}^c (\theta_i - \theta)^{r-j}], \qquad (A.1)$$

for conditional central moments $M_{ji}^c = E[(\hat{\theta}_i - \theta_i)^j | \theta_i]$ with $M_{2i}^c = v_i^2$.

For unbiased estimators $\hat{\theta}_i$, $M_{i1} = M_{i1}^c = 0$ and $M_{2i} = E(v_i^2) + \tau^2$.

The scaled sample SMD, $\sqrt{\tilde{n}} J^{-1} \hat{g}$, has the non-central t distribution $t_{n-2}(\gamma)$ with non-centrality parameter $\gamma = \sqrt{\tilde{n}}\delta$ (Hedges & Olkin, 1985, p. 79).

In what follows, we suppress the subscript i on all variables pertaining to study.i We require the central moments $E[(\hat{g} - \delta)^r]$ for $r = 1,\ \ldots,\ 6$. To ensure that these moments exist, we assume that.$n > 8$ (The usual chi-squared approximation requires $n > 4$ for the variance of g to exist.)

From Johnson, Kotz, and Balakrishnan (1995, p. 512), the moments of $t_{n-2}(\gamma)$ about zero are given by

$$\mu_r = E\left[t^r_{n-2}(\gamma)\right] = \left(\frac{n-2}{2}\right)^{r/2} \frac{\left[\frac{n-2-r}{2}\right]}{\left[\frac{n-2}{2}\right]} \sum_{j=0}^{r/2} \binom{r}{2j} \frac{(2j)!}{2^j j!} \mu_{r-2j}. \tag{A.2}$$

The first moment of $t_{n-2}(\gamma)$, denoted by μ_1, is

$$\mu_1 = \left(\frac{n-2}{2}\right)^{1/2} \frac{\Gamma\left[\frac{n-3}{2}\right]}{\Gamma\left[\frac{n-2}{2}\right]} \gamma \tag{A.3}$$

The second moment is

$$\mu_2 = \frac{n-2}{n-4}(1+\gamma^2).$$

The (conditional) central moments of Hedges's $\widehat{g}$ are

$$E[(\hat{g}-\delta)^r)] = \left(\frac{J}{\sqrt{\tilde{n}}}\right)^r E[(t_{n-2}(\gamma)-\mu_1)^r] = \left(\frac{J}{\sqrt{\tilde{n}}}\right)^r \sum_{s=0}^{r} (-1)^{r-s} \binom{r}{s} \mu_s \mu_1^{r-s}, \tag{A.4}$$

where μ_s is the sth moment $E\left[t^s_{n-2}(\gamma)\right]$ given by equation (A.2). Substituting the result from equation (A.2) and expressions for γ and μ_1, the conditional central moments of Hedges's g are$\widehat{g}$ are

$$\begin{aligned}
M^c_r &= \left(\tfrac{J}{\sqrt{\tilde{n}}}\right)^r \sum_{s=0}^{r} (-1)^{r-s} \binom{r}{s} s_1^{r-s} \\
&= \left(\tfrac{J}{\sqrt{\tilde{n}}}\right)^r \sum_{s=0}^{r} (-1)^{r-s} \binom{r}{s} \left(\tfrac{n-2}{2}\right)^{s/2} \frac{\left[\frac{n-2-s}{2}\right]}{\left[\frac{n-2}{2}\right]} \sum_{j=0}^{s/2} \binom{s}{2j} \frac{(2j)!}{2^j j!} s_1^{s-2j} r_1^{r-s} \\
&= \left(\tfrac{J}{\sqrt{\tilde{n}}}\right)^r \sum_{s=0}^{r} (-1)^{r-s} \binom{r}{s} \left(\tfrac{n-2}{2}\right)^{s/2} \frac{\left[\frac{n-2-s}{2}\right]}{\left[\frac{n-2}{2}\right]} \sum_{j=0}^{s/2} \binom{s}{2j} \frac{(2j)!}{2^j j!} \left(\frac{n-2}{2}\right)^{\frac{r-s}{2}} \left(\frac{\left[\frac{n-3}{2}\right]}{\left[\frac{n-2}{2}\right]}\right)^{r-s} \sim n^{\frac{r-2j}{2}} r_1^{r-2j} \\
&= \sum_{s=0}^{r} (-1)^{r-s} C_{rs} \sum_{j=0}^{s/2} B_{sj} r_1^{r-2j},
\end{aligned}$$

where $C_{rs} = J^r \left(\frac{n-2}{2}\right)^{\frac{r}{2}} \binom{r}{s} \left(\frac{\left[\frac{n-3}{2}\right]}{\left[\frac{n-2}{2}\right]}\right)^{r-s} \frac{\left[\frac{n-2-s}{2}\right]}{\left[\frac{n-2}{2}\right]}$ and $B_{sj} = \binom{s}{2j} \frac{(2j)!}{2^j j!} \sim n^{-j}$.

Now we apply these results to study i by restoring the subscript i on variables pertaining to study i and substituting the conditional moments into equation (A.1):

$$
\begin{aligned}
M_{ti} &= \sum_{r=0}^{t} \binom{t}{r} E(M_{ri}^{c}(\delta_i - \delta)^{t-r}) = \sum_{r=0}^{t} \binom{t}{r} \sum_{k=0}^{t-r} (-1)^k \binom{t-r}{k} E\left(M_{ri}^{c}\delta_i^{t-r-k}\right)\delta^k \\
&= \sum_{r=0}^{t} \binom{t}{r} \sum_{k=0}^{t-r} \binom{t-r}{k} \sum_{s=0}^{r} (-1)^{k+r-s} C_{rs} \sum_{j=0}^{\lfloor s/2 \rfloor} B_{sj} E\left(\delta_i^{t-2j-k}\right)\delta^k \\
&= \sum_{r=0}^{t} \binom{t}{r} \sum_{k=0}^{t-r} \binom{t-r}{k} \sum_{s=0}^{r} (-1)^{k+r-s} C_{rs} \sum_{j=0}^{\lfloor s/2 \rfloor} B_{sj} \sum_{m=0}^{t-2j-k} \binom{t-2j-k}{m} E_m \delta^{t-2j-m},
\end{aligned}
\tag{A.5}
$$

where E_m is the mth central moment of the $N(0,\ \tau^2)$ distribution. Define $E_0 = 1$. All odd central moments are zero, and the even moments are $E_m = \tau^m(m-1)!!$.