

Testing Categorical Moderators in Mixed-Effects Meta-analysis in Presence of Heteroscedasticity

Journal:	<i>The Journal of Experimental Education</i>
Manuscript ID	01-18-071.R2
Manuscript Type:	Measurement, Statistics, and Research Design
Keywords:	Effect Size , Meta-Analysis , Simulation Studies, residual between-studies variance, mixed-effects model, subgroup analyses
Abstract:	Mixed-effects models can be used to examine the association between a categorical moderator and the magnitude of the effect size. Two approaches are available to estimate the residual between-studies variance, τ_{res}^2 , namely separate estimation within each category of the moderator versus pooled estimation across all categories. We examine, by means of a Monte Carlo simulation study, both approaches for τ_{res}^2 estimation in combination with two methods to test the statistical significance of the moderator, namely the Wald-type χ^2 and F tests. Results suggest that the F test using a pooled estimate of τ_{res}^2 across categories is the best option in most conditions, although the F test using separate estimates of τ_{res}^2 is preferable if the residual heterogeneity variances are heteroscedastic.
Note: The following files were submitted by the author for peer review, but cannot be converted to PDF. You must view these files (e.g. movies) online.	
Supplementary file 1.R	

Testing Categorical Moderators in Mixed-Effects Meta-analysis in Presence of Heteroscedasticity

Abstract

Mixed-effects models can be used to examine the association between a categorical moderator and the magnitude of the effect size. Two approaches are available to estimate the residual between-studies variance, τ^2_{res} , namely separate estimation within each category of the moderator versus pooled estimation across all categories. We examine, by means of a Monte Carlo simulation study, both approaches for τ^2_{res} estimation in combination with two methods to test the statistical significance of the moderator, namely the Wald-type χ^2 and F tests. Results suggest that the F test using a pooled estimate of τ^2_{res} across categories is the best option in most conditions, although the F test using separate estimates of τ^2_{res} is preferable if the residual heterogeneity variances are heteroscedastic.

Keywords: meta-analysis; mixed-effects model; subgroup analyses; residual between-studies variance.

Introduction

Meta-analysis has emerged as the standard methodology for quantitatively integrating the results of a set of primary studies examining a common research question (Borenstein, Hedges, Higgins, & Rothstein, 2009; Cooper, Hedges, & Valentine, 2009; Schmidt & Hunter, 2015). Two of the main purposes of a meta-analysis are to calculate an overall effect estimate across studies and to assess the amount of variability among the individual effect sizes. If the amount of variability is larger than expected based on sampling error alone, then this is typically taken to indicate that the underlying true effects are heterogeneous. The amount of between-studies variance in the true effects can then be estimated using a random-effects model (DerSimonian & Laird, 1986). A further goal then consists of searching for study-level characteristics (often called ‘moderator variables’) that may be able to explain at least part of that variability (Lau, Ioannidis, & Schmid, 1998; Thompson, 1994).

In the present paper, we are particularly interested in the use of subgroup analyses, which are commonly used to examine the association between categorical moderator variables and the magnitude of the effect size. Based on a subgroup analysis, we can estimate the (average) effect size for each level of the moderator and test for between-group differences. Such analyses may provide valuable insights regarding the influence of qualitative moderators, as well as under which conditions an educational intervention is more effective.

A general recommendation when conducting such moderator analyses is to adopt a mixed-effects model which explicitly models potential ‘residual heterogeneity’ in the effects, that is, heterogeneity in the true effects not accounted for by the moderator variable(s) included in the model (Thompson & Higgins, 2002). For models

with a categorical moderator, residual heterogeneity simply denotes heterogeneity in the true effects within the various levels of the moderator.

Two approaches can be used to estimate the amount of residual heterogeneity in the context of such models. One is to allow for and estimate a different between-studies variance component (denoted by τ_{res}^2) within each level of the moderator, while the other consists of assuming a common amount of residual heterogeneity across categories and to calculate a pooled estimate thereof (Borenstein et al., 2009).

Rubio-Aparicio, Sánchez-Meca, López-López, Marín-Martínez, and Botella (2017) recently carried out a simulation study to compare the statistical performance of the omnibus Wald-type χ^2 test for between-group differences in the (average) effect sizes (here denoted as the Q_B test) in terms of its Type I error and statistical power rates when the two alternative procedures for estimating τ_{res}^2 (i.e., separate vs. pooled estimation) are used. The results indicated that pooled estimation is preferable for most scenarios, unless τ_{res}^2 is different across categories and the number of studies in each category is large enough to obtain precise separate estimates. However, the Type I error rate of the Q_B test was not nominal for many of the conditions examined, regardless of the approach used in the estimation of τ_{res}^2 . A potential explanation is that the test does not take into account the uncertainty derived from the estimation process of τ_{res}^2 , which typically results in inflated rejection rates under the null hypothesis.

Hartung, Makambi, and Argaç (2001), and Hartung, Argaç, and Makambi (2002) proposed an alternative method that accounts for the imprecision in the estimated amount of residual heterogeneity in subgroup analyses.

Knapp and Hartung (2003) proposed an improved method for meta-regression based on the same rationale that underlies the Hartung and colleagues' method (2001, 2002). In meta-regression, this method has repeatedly been found to provide adequate

control of the Type I error rate in several simulation studies (Huizenga, Visser, & Dolan, 2011; Knapp & Hartung, 2003; Sidik & Jonkman, 2005; Viechtbauer, López-López, Sánchez-Meca, & Marín-Martínez, 2015) and is routinely recommended nowadays (Gonzalez-Mulé & Aguinis, 2017). Nonetheless, the implementation of the alternative method is still relatively uncommon when testing for categorical moderators in contrast with growing popularity of the improved method for continuous moderators. It is important to note that the issue of estimating τ_{res}^2 separately for each category of the moderator or by means of a pooled estimate is specific to qualitative moderators, as continuous moderators are typically analyzed assuming a common τ_{res}^2 . Therefore, the performance of the improved method proposed by Hartung and colleagues (2001, 2002) when using pooled or separate estimates of τ_{res}^2 , and the conditions under which one approach should be recommended over the other have not yet been studied.

The purpose of the present study was to examine the Type I error and statistical power rates of the improved method proposed by Hartung and colleagues (Hartung et al., 2001, 2002; Knapp, & Hartung, 2003) to test the statistical significance of a qualitative moderator under a mixed-effects model when using pooled versus separate estimates of the residual heterogeneity variance. In addition, we compared Hartung and colleagues' method (2001, 2002) to the standard Q_B test. In sum, we compared the performance of four statistical tests: Hartung and colleagues (2001, 2002) versus standard Q_B tests in combination with pooled versus separate estimates of τ_{res}^2 . The results of this simulation study can shed light on where pooled or separate estimates of τ_{res}^2 should be preferred given the characteristics of the meta-analytic database.

In the next section, the mixed-effects model is outlined, followed by a description of the two hypothesis tests for categorical moderators and the different estimators of τ_{res}^2 either using pooled or separate estimates across categories. Then, the

methods and results from a Monte Carlo simulation study comparing the performance of the different procedures are detailed. Last, a discussion of the main results and implications arising from them is provided.

Mixed-effects model

In a meta-analysis with k studies grouped into m mutually exclusive categories of the moderator variable, let k_j denote the number of effect sizes of category j ($j = 1, \dots, m$; with $k_j > 1$ for all j), so that $k = \sum_j k_j$. The mixed-effects model assumes a random-effects model for the study-specific true effects within each category of the moderator variable and hence the statistical model is given by

$$T_{ij} = \mu_{\theta_j} + \epsilon_{ij} + e_{ij}, \quad (1)$$

where T_{ij} denotes the i th effect size estimate within the j th category, μ_{θ_j} represents the mean true effect size of the j th category, and ϵ_{ij} and e_{ij} represent the within-study and between-studies errors, respectively. It is common to assume that these two errors are normally distributed and independent of each other, and therefore, the estimated effect sizes are normally distributed as $T_{ij} \sim N(\mu_{\theta_j}, \sigma_{ij}^2 + \tau_{res(j)}^2)$, with σ_{ij}^2 being the within-study variance for the i th study in the j th category of the moderator and $\tau_{res(j)}^2$ denoting the residual between-studies variance in the j th category. The model also implies that the true effects in the j th category, θ_{ij} , follow a normal distribution with mean μ_{θ_j} and between-studies variance $\tau_{res(j)}^2$, that is, $\theta_{ij} \sim N(\mu_{\theta_j}, \tau_{res(j)}^2)$. Therefore, in a mixed-effects model a random sampling process underlies the standard random-effects model in each category of the moderator.

One of the main objectives in a subgroup analysis is to test the statistical association of the moderator with the effect sizes, which is accomplished by comparing

the mean effect sizes from each category of the moderator. For that aim, we first estimate the mean effect size of the j th category of the moderator, μ_{θ_j} , with

$$\bar{T}_j = \frac{\sum_i \hat{w}_{ij} T_{ij}}{\sum_i \hat{w}_{ij}}, \quad (2)$$

where the weights $\hat{w}_{ij} = 1/(\hat{\sigma}_{ij}^2 + \hat{\tau}_{res(j)}^2)$ are computed with $\hat{\sigma}_{ij}^2$ denoting the estimated within-study variance of the i th effect size of the j th category and $\hat{\tau}_{res(j)}^2$ an estimate of the residual between-studies variance of the j th category. Two strategies can be applied to estimate $\tau_{res(j)}^2$: (a) by pooling the estimated residual between-studies variances of the categories ($\hat{\tau}_{res(+)}^2$) or (b) by means of separate estimates of the residual between-studies variance (e.g., $\hat{\tau}_{res(1)}^2$ and $\hat{\tau}_{res(2)}^2$ for a dichotomous moderator, or $\hat{\tau}_{res(1)}^2$, $\hat{\tau}_{res(2)}^2$, and $\hat{\tau}_{res(3)}^2$ for a moderator with three categories). Note that one of main purposes of our investigation was to examine the extent to which pooled or separate estimates of the residual between-studies variance can affect the performance of statistical tests in a subgroup analysis.

An estimate of the variance of \bar{T}_j can be obtained with

$$Var[\bar{T}_j] = \frac{1}{\sum_i \hat{w}_{ij}}. \quad (3)$$

Tests of between-groups differences

The statistical association of a categorical moderator with the effect sizes can be tested by means of a standard Wald-type χ^2 test (Borenstein et al., 2009)

$$Q_B = \sum_{j=1}^m \hat{w}_{+j} (\bar{T}_j - \bar{T})^2, \quad (4)$$

where $\hat{w}_{+j} = 1/Var[\bar{T}_j]$, with $Var[\bar{T}_j]$ defined in Equation 3, and \bar{T} represents the weighted average of all effect sizes and is computed with

$$\bar{T} = \frac{\sum_j \sum_i \hat{w}_{ij} T_{ij}}{\sum_j \sum_i \hat{w}_{ij}}, \quad (5)$$

where $\hat{w}_{ij} = 1/(\hat{\sigma}_{ij}^2 + \hat{\tau}_{res(j)}^2)$ and T_{ij} denotes the i th effect size estimate of the j th category of the moderator. Note that, as mentioned above, $\hat{\tau}_{res(j)}^2$ can be calculated in two ways: as a pooled ($\hat{\tau}_{res(+)}^2$) or as a separate ($\hat{\tau}_{res(j)}^2$) estimate.

Under the null hypothesis that the m categories share the same true mean effect size ($H_0: \mu_{\theta_1} = \mu_{\theta_2} = \dots = \mu_{\theta_m}$), the Q_B statistic follows asymptotically a χ^2 distribution with $m - 1$ degrees of freedom (requiring both large within-study sample sizes and large k_j for $j = 1, \dots, m$).

An alternative method to test the statistical significance of a categorical moderator is computed with (Hartung et al., 2001, 2002)

$$F = \frac{\frac{Q_B}{m-1}}{\frac{Q_W}{k-m}}, \quad (6)$$

where $Q_W = \sum_j Q_{w_j}$ and

$$Q_{w_j} = \sum_{i=1}^{k_j} \hat{w}_{ij} (T_{ij} - \bar{T}_j)^2, \quad (7)$$

with $\hat{w}_{ij} = 1/(\hat{\sigma}_{ij}^2 + \hat{\tau}_{res(j)}^2)$.

Under the null hypothesis of no difference between the mean effect sizes across categories ($H_0: \mu_{\theta_1} = \mu_{\theta_2} = \mu_{\theta_3} = \dots = \mu_{\theta_m}$), the F statistic follows asymptotically an F distribution with $(m - 1)$ and $(k - m)$ degrees of freedom. The F statistic proposed by Hartung and colleagues takes into account the uncertainty due to the estimation of the residual between-studies variance and, as a consequence, it is expected to outperform the standard Q_B statistic.

The F test for subgroup analyses can be considered to be a special case of the improved method for meta-regression. In the meta-regression context, Knapp and Hartung (2003) proposed a multiplicative adjustment factor for the estimated variances of the model coefficients, and suggested to truncate this factor to one if a smaller value was obtained, in order to minimize false positive findings. Several pieces of meta-analytic software currently incorporate such truncation, including Comprehensive Meta-Analysis 3.3 (Borenstein, Hedges, Higgins, & Rothstein, 2014) and the *metareg* macro for Stata (Harbord & Higgins, 2008), whereas other alternatives like the *metafor* package for R (Viechtbauer, 2010) use the untruncated factor by default. This adjustment factor is equal to the denominator of the F formula (see eq. 6), hence implementing the truncation in the context of a subgroup analysis would be straightforward. However, Viechtbauer et al. (2015) showed that the improved method for meta-regression provides an adequate adjustment of the nominal significance level without truncating, whereas overly conservative results may be obtained if the truncation is applied. Consequently, in the present study, we allowed the denominator of the F test to be smaller than one, and we generally would recommend this version of the test.

Estimating the residual between-studies variance

Several methods have been proposed to estimate τ^2 in the context of the random-effects model (Sánchez-Meca & Marín-Martínez, 2008; Viechtbauer, 2005). Most of these estimators have also been extended to the mixed-effects model, and we selected three methods that are commonly implemented and have been found to perform adequately in previous simulation studies (López-López, Marín-Martínez, Sánchez-Meca, Van den Noortgate, & Viechtbauer, 2014; Veroniki et al., 2016). In this section, we describe the

three estimators used in the present study and their computation using both separate estimates and pooled estimate of $\tau_{res(j)}^2$.

DerSimonian and Laird (DL) estimator

The estimator proposed by DerSimonian and Laird (1986), probably the most commonly used in meta-analysis, is derived from the method of moments. Applying this estimator, the residual between-studies variance for the j th category of the moderator, $\hat{\tau}_{res(j)}^2$, can be computed with the expression

$$\hat{\tau}_{res(j)DL}^2 = \frac{Q_{wj^*} - (k_j - 1)}{c_j}, \quad (8)$$

where Q_{wj^*} is computed with Equation 7, but using $\hat{w}_{ij}^* = 1/\hat{\sigma}_{ij}^2$ as the weights, and c_j is given by

$$c_j = \sum_i \hat{w}_{ij}^* - \frac{\sum_i (\hat{w}_{ij}^*)^2}{\sum_i \hat{w}_{ij}^*}. \quad (9)$$

Note that the Q_{wj} statistic defined in Equation 7 is not the same as the standard Q_{wj^*} statistic proposed by Hedges and Olkin (1985) to test the model misspecification in a mixed-effects model. Unlike Q_{wj} , the weights used to calculate the Q_{wj^*} statistics are a function of the within-study variance only. Should the DL estimate turn out to be negative, it is truncated to zero.

The pooled estimate of the residual between-studies variance applying DerSimonian and Laird is given by (Borenstein et al., 2009)

$$\hat{\tau}_{res(+)DL}^2 = \frac{\sum_j Q_{wj^*} - \sum_j (k_j - 1)}{\sum_j c_j}. \quad (10)$$

Restricted Maximum Likelihood (REML) estimator

An alternative for estimating $\tau_{res(j)}^2$ is based on restricted maximum likelihood estimation. The REML estimator for the j th category of the moderator can be obtained iteratively from

$$\hat{\tau}_{res(j)REML}^2 = \frac{\sum_i \hat{w}_{ij}^2 [(T_{ij} - \bar{T}_j)^2 - \hat{\sigma}_{ij}^2]}{\sum_i \hat{w}_{ij}^2} + \frac{1}{\sum_i \hat{w}_{ij}} \quad (11)$$

by first computing the right-hand side using initial values for the weights (e.g., by setting $\hat{\tau}_{res(j)}^2$ in $\hat{w}_{ij} = 1/(\hat{\sigma}_{ij}^2 + \hat{\tau}_{res(j)}^2)$ equal to the estimate obtained using the non-iterative DL estimator), then updating the weights (and hence also \bar{T}_j) using the estimate of $\hat{\tau}_{res(j)}^2$ obtained, and then iterating this process until convergence. Should $\hat{\tau}_{res(j)}^2$ ever become negative during this process, the estimate is truncated to zero.

The pooled REML estimate of the residual variance is again computed iteratively, but now using

$$\hat{\tau}_{res(+)REML}^2 = \frac{\sum_j \sum_i \hat{w}_{ij}^2 [(T_{ij} - \bar{T}_j)^2 - \hat{\sigma}_{ij}^2]}{\sum_j \sum_i \hat{w}_{ij}^2} + \frac{m}{\sum_j \sum_i \hat{w}_{ij}}, \quad (12)$$

with weights $\hat{w}_{ij} = 1/(\hat{\sigma}_{ij}^2 + \hat{\tau}_{res(+)}^2)$.

Paule and Mandel (PM) estimator

The third estimator that we included in our simulation study was proposed by Paule and Mandel (1982), and it is sometimes labelled as empirical Bayes estimator (Morris, 1983). A recent review of simulation studies concluded recommending use of the PM estimator in meta-analysis (Langan, Higgins, & Simmonds, 2017). Furthermore, another simulation study comparing seven methods in the context of meta-regression found that the PM, DL and REML estimators yielded the best results across conditions (López-López et al., 2014).

The PM estimate for the j th category is given by the solution to

$$\sum_i \hat{w}_{ij} (T_{ij} - \bar{T}_j)^2 - (k_j - 1) = 0. \quad (13)$$

The left-hand side of Equation (13) is a monotonically decreasing function of $\hat{\tau}_{res(j)}^2$ and can be easily solved for 0 using any standard root finding algorithm. We denote the resulting estimate with $\hat{\tau}_{res(j)PM}^2$. Should Equation (13) be negative for $\hat{\tau}_{res(j)}^2 = 0$, then the estimate is truncated to zero.

To obtain the pooled estimate for the PM estimator, $\hat{\tau}_{res(+)PM}^2$, we must solve

$$\sum_j \sum_i \hat{w}_{ij} (T_{ij} - \bar{T}_j)^2 - \sum_j (k_j - 1) = 0, \quad (14)$$

with weights $\hat{w}_{ij} = 1/(\hat{\sigma}_{ij}^2 + \hat{\tau}_{res(+)}^2)$.

Method

In the previous section, we presented two methods for testing the statistical significance of a categorical moderator (i.e., the Q_B and F tests) and three methods (i.e., the DL, REML, and PM estimators) which can be used to obtain either a pooled estimate or separate estimates for τ_{res}^2 . This yields 12 different ways of testing the statistical significance of a categorical moderator in a mixed-effects model subgroup analysis, namely the $Q_{B(S)}$ test using separate estimates of the heterogeneity variance combined with either the DL, REML, or PM estimator ($Q_{B(S)DL}$, $Q_{B(S)REML}$, and $Q_{B(S)PM}$, respectively), the $Q_{B(P)}$ test when using a pooled estimate using either the DL, REML, or PM estimator ($Q_{B(P)DL}$, $Q_{B(P)REML}$, and $Q_{B(P)PM}$, respectively), the $F_{(S)}$ test using separate estimates ($F_{(S)DL}$, $F_{(S)REML}$, and $F_{(S)PM}$, respectively), and the $F_{(P)}$ test when using a pooled estimate ($F_{(P)DL}$, $F_{(P)REML}$, and $F_{(P)PM}$, respectively). To compare the performance of these methods, we conducted a Monte Carlo simulation study

programmed in R using the *metafor* package (Viechtbauer, 2010). Supplementary file 1 contains the full R code of our simulation study.

Meta-analyses of k studies were simulated with the standardized mean difference as the effect size index. Each individual study included in a meta-analysis compared two groups (experimental and control) with respect to some continuous outcome. For a given study, values of the outcome were sampled from normal distributions with equal variances (i.e., $N(\mu_E, \sigma^2)$ and $N(\mu_C, \sigma^2)$). For each study, the population standardized mean difference, θ , was defined as (Hedges & Olkin, 1985)

$$\theta = \frac{\mu_E - \mu_C}{\sigma}. \quad (15)$$

Without loss of generality, the normal distributions of the experimental and control populations were defined as $N(\theta, 1)$ and $N(0, 1)$, respectively.

The effect size was estimated by means of the nearly unbiased estimator proposed by Hedges and Olkin (1985, p. 81)

$$T = c(m) \frac{\bar{y}_E - \bar{y}_C}{s}, \quad (16)$$

where \bar{y}_E and \bar{y}_C are the sample means of the experimental and control groups, s is the pooled standard deviation computed with

$$s = \sqrt{\frac{(n_E - 1)s_E^2 + (n_C - 1)s_C^2}{n_E + n_C - 2}}, \quad (17)$$

n_E and n_C being the experimental and control group sample sizes, respectively, s_E^2 and s_C^2 the variances of the two groups, and $c(m)$ is a correction factor for small sample sizes given by

$$c(m) = 1 - \frac{3}{4N - 9}, \quad (18)$$

where $N = n_E + n_C$. The estimated within-study variance of T , assuming equal variances and normality within each study, is given by

$$\hat{\sigma}^2 = \frac{n_E + n_C}{n_E n_C} + \frac{T^2}{2(n_E + n_C)}. \quad (19)$$

The k studies were assumed to fall into two or three categories (with k_1 and k_2 studies in each group for a dichotomous moderator, and k_1 , k_2 , and k_3 for a moderator with three categories). The true standardized mean differences within each subgroup were simulated from $N(\mu_{\theta_j}, \tau_{res(j)}^2)$ according to a mixed-effects model.

For a dichotomous moderator, we set the number of studies, k , to values of 12, 20, 40, and 60. For a moderator with three categories, we set k to values of 12, 24, 48, and 60. Choice of values for k was based on a review of meta-analyses undertaken by Anh, Ames, and Myers (2012) in the educational context where the first quartile, median, and third quartile of the empirical distribution of the number of studies were found to be 22, 38 and 67, respectively. Moreover, we manipulated how k was distributed within each category of the moderator, so that in some conditions there was a balanced distribution (i.e., $k_1 = k_2$, or $k_1 = k_2 = k_3$), while in the remaining conditions there was an unbalanced distribution (i.e., $k_1 \neq k_2$, or $k_1 \neq k_2 \neq k_3$) between the two or three categories. For a dichotomous moderator, an unbalanced distribution implied that the second category contained three times as many studies as the first category. For instance, when $k = 12$ we set $k_1 = k_2 = 6$ in the balanced conditions, and $k_1 = 3$ and $k_2 = 9$ in the unequal conditions. For a moderator with three categories, an unbalanced distribution implied that the second category contained twice as many studies as the first category, and the third category was three times as many studies as the first one. For instance, when $k = 12$ we set $k_1 = k_2 = k_3 = 4$ in balanced conditions, and $k_1 = 2$, $k_2 = 4$, and $k_3 = 6$ in the unequal conditions.

Furthermore, $\tau_{res(j)}^2$ was manipulated in two different ways. First, we considered three values for this parameter, 0.08, 0.16, and 0.32. Second, we simulated a set of scenarios with homoscedastic variances across categories ($\tau_{res(1)}^2 = \tau_{res(2)}^2$, or $\tau_{res(1)}^2 = \tau_{res(2)}^2 = \tau_{res(3)}^2$), as well as another set of heteroscedastic conditions ($\tau_{res(1)}^2 \neq \tau_{res(2)}^2$, or $\tau_{res(1)}^2 \neq \tau_{res(2)}^2 \neq \tau_{res(3)}^2$). In particular, under homogeneous conditions $\tau_{res(1)}^2 = \tau_{res(2)}^2 = 0.08$, $\tau_{res(1)}^2 = \tau_{res(2)}^2 = 0.16$, and $\tau_{res(1)}^2 = \tau_{res(2)}^2 = 0.32$ for a dichotomous moderator, and $\tau_{res(1)}^2 = \tau_{res(2)}^2 = \tau_{res(3)}^2 = 0.16$ for a moderator with three categories. Heteroscedastic variances were manipulated for a dichotomous moderator with pairs of values $\tau_{res(1)}^2 = 0.08$ and $\tau_{res(2)}^2 = 0.16$, $\tau_{res(1)}^2 = 0.16$ and $\tau_{res(2)}^2 = 0.08$, $\tau_{res(1)}^2 = 0.08$ and $\tau_{res(2)}^2 = 0.32$, $\tau_{res(1)}^2 = 0.32$ and $\tau_{res(2)}^2 = 0.08$, $\tau_{res(1)}^2 = 0.16$ and $\tau_{res(2)}^2 = 0.32$, and $\tau_{res(1)}^2 = 0.32$ and $\tau_{res(2)}^2 = 0.16$. For a moderator with three categories, the variance of the second category was always fixed at 0.16 ($\tau_{res(2)}^2 = 0.16$), and the variances of the first and the third categories were varied ($\tau_{res(1)}^2 = 0.08$ and $\tau_{res(3)}^2 = 0.16$, $\tau_{res(1)}^2 = 0.16$ and $\tau_{res(3)}^2 = 0.08$, $\tau_{res(1)}^2 = 0.08$ and $\tau_{res(3)}^2 = 0.08$, $\tau_{res(1)}^2 = 0.08$ and $\tau_{res(3)}^2 = 0.32$, $\tau_{res(1)}^2 = 0.32$ and $\tau_{res(3)}^2 = 0.08$, $\tau_{res(1)}^2 = 0.32$ and $\tau_{res(3)}^2 = 0.32$, $\tau_{res(1)}^2 = 0.16$ and $\tau_{res(3)}^2 = 0.32$, and $\tau_{res(1)}^2 = 0.32$ and $\tau_{res(3)}^2 = 0.16$).

The average total sample size of the individual studies \bar{N} was set to 20, 40, 60, 80, 200 and 2,000. These values were chosen following the revision of meta-analyses in education carried out by Ahn et al. (2012). In this review, the first quartile, median, and third quartile of the average total sample size distribution were 90, 185 and 1,900, respectively. The data in the primary studies were simulated assuming $n_E = n_C$. A χ^2 distribution with 4 degrees of freedom was used, so that the skewness of the distribution was +1.414. In addition, values equal to 16, 36, 56, 76, 196, or 1,996 were added to get the desired average value.

The mean effect size of each category of the moderator was also manipulated. Regarding a dichotomous moderator, in some conditions the two parametric mean effects were both equal to 0.5 ($\mu_{\theta_1} = \mu_{\theta_2} = 0.5$), whereas for other conditions they were set to different values: $\mu_{\theta_1} = 0.5$ and $\mu_{\theta_2} = 0.3$, $\mu_{\theta_1} = 0.5$ and $\mu_{\theta_2} = 0.1$, and $\mu_{\theta_1} = 0.7$ and $\mu_{\theta_2} = 0.1$. With respect to the moderator with three categories, in the equal conditions the three parametric mean effects were set to 0.3 ($\mu_{\theta_1} = \mu_{\theta_2} = \mu_{\theta_3} = 0.3$), while in the unequal conditions another set of values were manipulated: $\mu_{\theta_1} = 0.2$, $\mu_{\theta_2} = 0.3$ and $\mu_{\theta_3} = 0.4$, $\mu_{\theta_1} = 0.1$, $\mu_{\theta_2} = 0.3$ and $\mu_{\theta_3} = 0.5$, and $\mu_{\theta_1} = 0$, $\mu_{\theta_2} = 0.3$, and $\mu_{\theta_3} = 0.6$. Note that for both types of moderators (two and three categories) the difference between the largest mean effect size and the smallest one was fixed to 0.2, 0.4, and 0.6 across the unequal conditions. The manipulated conditions for the mean effect sizes covered a wide range of values around what can be considered as effect sizes of medium magnitude, following the benchmark of 0.5 proposed by Cohen (1988) in the behavioral sciences and the empirical value of 0.3 found by Lipsey and Wilson (1993) in the educational sciences. The conditions with equal mean effect sizes across categories allowed us to study the Type I error rate, whereas the conditions with different mean effect sizes enabled us to assess the statistical power.

To assess the Type I error rate, the total number of conditions was: 4 (number of studies) \times 2 (balanced-unbalanced number of studies in the two categories) \times 6 (average total sample size) \times 9 (residual between-studies variance) = 432. With respect to the statistical power, $432 \times 3 = 1,296$ conditions examined. Overall, the total number of conditions was therefore $1,728 \times 2$ (moderator with two and three categories) = 3,456 and for each condition we generated 10,000 replications. Thus, 34,560,000 meta-analyses were simulated. The 12 methods ($Q_{B(S)DL}$, $Q_{B(S)REML}$, $Q_{B(S)PM}$, $Q_{B(P)DL}$, $Q_{B(P)REML}$, $Q_{B(P)PM}$, $F_{(S)DL}$, $F_{(S)REML}$, $F_{(S)PM}$, $F_{(P)DL}$, $F_{(P)REML}$, and $F_{(P)PM}$) were applied to each one of

these replications. In each of the 3,456 conditions of our simulation study, the proportion of rejections of the null hypothesis of equality of the mean effect sizes across categories of the moderator was examined.

Results

In this section, we describe and compare the performance of the methods under the simulated conditions. For brevity, we only present the results for the PM estimator since the pattern of results was very similar for the remaining estimators. Nevertheless, Supplementary file 2 presents figures using the DL and REML estimators, and the full set of results can be obtained from the corresponding author upon request. This section is divided into two parts, corresponding to the Type I error and statistical power rates, respectively.

Type I Error

Setting $\mu_{\theta_1} = \mu_{\theta_2} = 0.5$ and $\mu_{\theta_1} = \mu_{\theta_2} = \mu_{\theta_3} = 0.3$ allowed comparing the methods in terms of their Type I error rates for a moderator with two and three categories, respectively. Figures in this section include dashed horizontal lines delimiting the range of values that can be considered as equivalent to the nominal significance level of 5%, after accounting for Monte Carlo error [.0543; .0457]. Therefore, empirical rejection rates within this interval indicated adequate control of the Type I error rate.

Figure 1 shows the average Type I error rates as a function of the number of studies, balanced and unbalanced distribution of number of studies within each category of the moderator, average sample size per study, and the amount of residual heterogeneity, in scenarios with homoscedastic residual between-studies variances across the two categories of the moderator. As k increased (Figure 1A), the proportion

of rejections of the null hypothesis of equality for $Q_{B(S)}$, $Q_{B(P)}$, and $F_{(S)}$, converged to the nominal significance level, whereas $F_{(P)}$ showed nominal levels regardless of the number of studies. Focusing on the balanced versus unbalanced distribution of the number of studies across categories (Figure 1B), $Q_{B(P)}$ and $F_{(P)}$ were not influenced by this factor, whereas $Q_{B(S)}$ and $F_{(S)}$ showed higher empirical rejection rates (above .05) when the number of studies was unbalanced across categories. Last, sample size and the amount of residual heterogeneity did not seem to have a strong influence on the rejection rates (Figures 1C and 1D), with $F_{(P)}$ consistently yielding the best control of the Type I error rate.

INSERT FIGURE 1

Figure 2 presents the average Type I error rates in conditions where the residual between-studies variances were heteroscedastic across the two categories of the moderator, and the category with less studies had the smaller variance. The influence of the number of studies (Figure 2A) was more pronounced for the Q_B test, with lower Type I error rates as k increased, and $Q_{B(S)}$ showing inflated rates with less than 40 studies. The F test was less affected, with $F_{(S)}$ showing an adequate control and $F_{(P)}$ yielding overly conservative results, regardless of the number of studies. Regarding the distribution of the number of studies (Figure 2B), $Q_{B(S)}$ and $F_{(S)}$ were not influenced by this factor, whereas $Q_{B(P)}$ and $F_{(P)}$ showed error rates below .05 under unbalanced distribution of the number of studies. Furthermore, results did not show important variations as a function of the average sample size and the amount of residual heterogeneity (Figures 2C and 2D), with $F_{(S)}$ and $Q_{B(P)}$ leading to a good adjustment to

the nominal level on average, $F_{(P)}$ yielding overconservative results, and $Q_{B(S)}$ showing inflated Type I error rates.

INSERT FIGURE 2

Figure 3 shows the average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the two categories of the moderator and larger variance for the category with less studies. When looking at the results as a function of the number of studies (Figure 3A), the rejection rates generally fell above the nominal significance level, with accurate rates provided only by $Q_{B(S)}$ and $F_{(S)}$ with at least 60 and 40 studies, respectively. Regarding the distribution of the number of studies in each category of the moderator, only $F_{(P)}$ and $F_{(S)}$ achieved good adjustment when the number of studies was balanced across categories, with inflated Type I error rates for all methods in the unbalanced scenarios (Figure 3B). The influence of the average sample size and the amount of residual heterogeneity were relatively minor (Figures 3C and 3D), and all methods yielded rejection rates that were too liberal. The $F_{(S)}$ test consistently provided the closest performance to the nominal significance level.

INSERT FIGURE 3

Figure 4 presents the average Type I error rates in scenarios with homoscedastic residual between-studies variances across the three categories of the moderator ($\tau_{res(1)}^2 = \tau_{res(2)}^2 = \tau_{res(3)}^2 = 0.16$). $F_{(P)}$ consistently yielded the best control of the Type I error rate in all situations for number of studies, balanced and unbalanced distribution of

number of studies across the three categories, and sample size (Figures 4A, 4B, and 4C, respectively). $Q_{B(S)}$ and $F_{(S)}$ yielded inflated rates above .15 under all scenarios.

INSERT FIGURE 4

Figure 5 shows the average Type I error rates in conditions where the residual between-studies variances were heteroscedastic across the three categories of the moderator, with smaller variance for the category with less studies. $Q_{B(P)}$ provided accurate rates as the number of studies increased, whereas $F_{(P)}$ yielded rates slightly under .05 regardless of the number of studies (Figure 5A). When looking at the results as a function of the distribution of the number of studies (Figure 5B), $F_{(P)}$ yielded a good adjustment to the nominal level under balanced distribution of the number of studies and $Q_{B(P)}$ did under unbalanced distribution. In addition, results did not show substantial variations as a function of the sample size and the amount of residual heterogeneity (Figures 5C and 5D), with $Q_{B(P)}$ yielding inflated error rates, and $F_{(P)}$ showing overconservative results. Last, $Q_{B(S)}$ and $F_{(S)}$ presented inflated rates above .15 across all the conditions (Figures 5A, 5B, 5C, and 5D).

INSERT FIGURE 5

Figure 6 presents the average Type I error rates in scenarios where the residual between-studies variances were heteroscedastic across the three categories of the moderator, with larger variance for the category with less studies. The influence of the conditions manipulated for the number of studies, balanced and unbalanced distribution of the number of studies, sample size and amount of residual heterogeneity was similar,

for all the methods, to the pattern found in Figure 5. In general, the adjustment to the Type I error rate of the $Q_{B(P)}$ and $F_{(P)}$ was deteriorated across all conditions (Figures 6A, 6B, 6C, and 6D), with $F_{(P)}$ performing closest to the nominal significance level. Once again, $Q_{B(S)}$ and $F_{(S)}$ presented the poorest adjustment under all conditions (Figures 6A, 6B, 6C, and 6D).

INSERT FIGURE 6

Finally, it is worth noting that, in general, the methods yielded a poorer adjustment to the error rate under scenarios with a moderator with three categories (see Figures 4-6) than under the dichotomous scenarios (see Figures 1-3).

Statistical Power

Statistical power reflects the probability of a method rejecting the null hypothesis that is in fact false. In general, power rates equal to or greater than 0.8 are often considered as acceptable in psychology and education (Cohen, 1988).

Figure 7 presents the average power rates in scenarios with homoscedastic residual between-studies variances across the two categories of the moderator. First, the influence of the different conditions manipulated was equivalent for $Q_{B(S)}$, $Q_{B(P)}$, $F_{(S)}$, and $F_{(P)}$ and, in most conditions, yielding statistical power below 0.8. As expected, for all methods, power increased as the number of studies (Figure 7A) and the magnitude of the difference between the mean effect sizes of the two categories (Figure 7E) increased, with at least 60 studies and a difference between the mean effect sizes equal to 0.6 ($\mu_{\theta_1} = 0.7$ and $\mu_{\theta_2} = 0.1$) being needed for the methods to provide power rates close to 0.8. Furthermore, larger residual heterogeneity resulted in lower power rates

(Figure 7D), whereas the distribution of the number of studies across categories (Figure 7B) and the average sample size per study (Figure 7C) did not show a substantial impact on the power rates of the methods under assessment. The Q_B test yielded slightly higher power rates than the F test across all manipulated conditions.

INSERT FIGURE 7

Figures 8 and 9 present the average power rates in scenarios where the residual between-studies variances were heteroscedastic across the two categories of the moderator, with the largest variance either falling in the category with more (Figure 8) or with less studies (Figure 9). The influence of the different conditions manipulated on the power rates of $Q_{B(S)}$, $Q_{B(P)}$, $F_{(S)}$, and $F_{(P)}$ was very similar to those under homoscedastic residual between-studies variances (see Figure 7), with larger k and larger differences among the mean effects leading to higher power rates. It is worth noting the effect of the residual between-studies variance on the power rates. On the one hand, when the category with less studies had less heterogeneous effect sizes (Figure 8D), $Q_{B(S)}$, $Q_{B(P)}$, $F_{(S)}$, and $F_{(P)}$ yielded power rates relatively higher under the condition of $\tau^2_{res(1)} = 0.08$ and $\tau^2_{res(2)} = 0.32$. On the other hand, when the category with less studies was more heterogeneous (Figure 9D), power rates for all of methods were slightly higher under the condition of $\tau^2_{res(1)} = 0.16$ and $\tau^2_{res(2)} = 0.08$.

INSERT FIGURES 8 AND 9

Figure 10 shows the average power rates in scenarios with homoscedastic residual between-studies variances across the three categories of the moderator ($\tau^2_{res(1)} = \tau^2_{res(2)} = \tau^2_{res(3)} = 0.16$). As in the dichotomous situation, the influence of the

different manipulated conditions was equivalent for all methods, with statistical power rates below 0.8 in most situations (Figures 10A, 10B, 10C, and 10D). As expected, power increased as the number of studies (Figure 10A) and the magnitude of the difference between the mean effect sizes of the first and third category of the moderator (Figure 10D) increased. The distribution of the number of studies (Figure 10B) and the sample size (Figure 10C) did not substantially affect the results. The $Q_{B(S)}$ and $F_{(S)}$ were the tests with the highest power rates across all manipulated conditions.

INSERT FIGURE 10

Figures 11 and 12 present the average power rates under the two heterogeneous situations of the residual between-studies variances across the three categories of the moderator, with the smallest variance either for the category with less (Figure 11) or with more studies (Figure 12). The impact of the conditions manipulated was very similar to the pattern observed under homoscedastic variances (see Figure 10), with higher power rates for the $Q_{B(S)}$ and $F_{(S)}$ tests. Regarding the effect of the residual between-studies variances on the results, power rates for all methods were slightly higher under the conditions of $\tau_{res(1)}^2 = .08$, $\tau_{res(2)}^2 = .16$, $\tau_{res(3)}^2 = .16$ (Figure 11D), and $\tau_{res(1)}^2 = .16$, $\tau_{res(2)}^2 = .16$, $\tau_{res(3)}^2 = .08$ (Figure 12D).

INSERT FIGURES 11 AND 12

Discussion

This study compared a variety of methods in the context of subgroup analyses using mixed-effects models. Specifically, two methods for testing the statistical significance

of the categorical moderator (i.e., the Q_B and F tests), two procedures for estimating the residual between-studies variance (pooled or separate estimates), and three residual heterogeneity variance estimators (DL, REML, and PM) were combined to provide twelve analysis approaches that were examined in a Monte Carlo simulation study, with standardized mean differences as the effect size measure. Two comparative criteria, empirical Type I error and statistical power rates, were considered for assessing the adequacy of each method across a wide variety of realistic scenarios in education.

Results were not found to be affected by the residual between-studies variance estimator used. However, some notable differences were observed depending on the method employed for testing the statistical association of a categorical moderator and on the approach implemented to estimate the amount of residual heterogeneity in each category (pooled versus separate estimates).

Some authors have criticized that the standard random-effects method does not take into account the uncertainty derived from the variance estimation process, which can lead to wrong statistical conclusions (e.g., Thompson & Higgins, 2002). This led to the development of improved hypothesis tests by Hartung and colleagues in the context of random-effects meta-analysis (Hartung, 1999) and mixed-effects meta-regression (Knapp & Hartung, 2003). These tests are known to outperform the standard methods in terms of their control of the Type I error rate (Huizenga et al., 2011; Sánchez-Meca & Marín-Martínez, 2008; Sidik & Jonkman, 2005; Viechtbauer et al., 2015) and are recommended for routine use nowadays. Hartung and colleagues (2001) also proposed an improved method for subgroup analyses using mixed-effects models using an F test, and we examined its performance compared to the typically implemented Q_B test, and using pooled or separate estimates of the residual heterogeneity variance. The empirical Type I error rates obtained by both methods suggest that, in general, the improved F test

has clear advantages over the standard Q_B test for moderators with two and three categories. As expected, this finding coincides with that obtained in previous studies for continuous moderators (Huizenga et al., 2011; Knapp, & Hartung, 2003; Viechtbauer et al., 2015). Therefore, this leads us to encourage meta-analysts who carry out subgroup analyses to apply the F test instead of the standard Q_B test in most situations.

When comparing the performance of the $F_{(P)}$ and $F_{(S)}$ tests, under homoscedastic variances across the two or three categories, $F_{(P)}$ yielded the best control of the Type I error rates, regardless of how the number of studies was distributed across the categories of the moderator.

Under heteroscedastic variances across categories, the performance of the $F_{(P)}$ and $F_{(S)}$ was different depending on whether the moderator was dichotomous or with three categories. Both $F_{(P)}$ and $F_{(S)}$ achieved adequate performance as long as the number of studies was distributed equally across the two categories of the moderator, whereas for a moderator with three categories, only $F_{(P)}$ showed a good performance. However, under an unbalanced distribution of the number of studies, the practical consequences of allowing for heteroscedastic residual between-studies variances were more evident. On the one hand, when the value of the smallest residual between-studies variance was associated with the category with the smallest number of studies the $F_{(S)}$ showed good adjustment for the dichotomous situation (see Figure 2), whereas for a moderator with three categories $Q_{B(P)}$ showed the best adjustment (see Figure 5). On the other hand, when the value of the largest residual between-studies variance was associated with the category with the smallest number of studies, all tests showed a poor adjustment to the nominal level for both moderators with two and three categories (see Figures 3 and 6, respectively).

These results allow us to recommend the use of the $F_{(p)}$ test in most conditions, except when the meta-analyst suspects that the true value of τ^2_{res} may vary across categories and the number of studies across categories is unbalanced. In that case, the $F_{(s)}$ and $Q_{B(p)}$ tests showed the best performance for moderators with two and three categories, respectively. Note that using a pooled estimate would be expected to provide more accurate results for most scenarios, as the estimate is then based on a larger number of studies. This can be particularly important if the total number of studies is small (e.g., $k < 20$), which has been found to be the case for most Cochrane Reviews (Davey, Turner, Clarke, & Higgins, 2011).

The statistical power of all methods was lower than .80 in most conditions, unless the magnitude of the difference between the mean effects across the two or three categories was equal to 0.6. As expected, statistical power rates increased with a larger number of studies, yielding rates close to .80 with at least 60 studies (see Figures 7-12). Note that the differences in the statistical power rates for the methods may also be caused by either inflated or overly conservative Type I error rates.

In summary, results of our simulation study suggest that out of the different alternatives considered in the present study, the improved F test computed using a pooled estimate is the most suitable option to test the statistical association between a categorical moderator and the effect sizes in most conditions. Nevertheless, if the meta-analyst suspects that the residual between-studies variances are heteroscedastic across categories of the moderator and the number of studies is unbalanced across categories, then the F test using separate estimates of the residual between-studies variance for a dichotomous moderator and the Q test using a pooled estimation of the residual between-studies variance for a moderator with three categories may be preferable.

These conclusions provide valuable information for applied researchers carrying out subgroup meta-analyses in the educational arena. Our empirical results enabled us to make several recommendations about which method (Q_B or F test) in combination with which procedure for estimating the residual between-studies variance (pooled or separate estimation) is the most suitable option depending on the characteristics of the meta-analytic database. Educational researchers should be aware of the practical consequences that the choice of one method or another could have for their meta-analytic results. For this reason, it is highly recommended to use software (e.g., the *metafor* package in R) that allows to choose between both tests and both estimation procedures when conducting a subgroup meta-analysis.

Limitations

Our study has several limitations. First, the present simulation study was conducted with standardized mean differences, but its results may be generalized to other effect size measures with (asymptotically) normal sampling distributions (e.g., mean difference, log odds ratio, log risk ratio, Fisher's Z-transformed correlation coefficients). Second, our results are limited to the manipulated conditions. Nevertheless, the values for the parameters were chosen to represent real meta-analyses in education. Lastly, an important limitation in this field is that the meta-analyst cannot determine whether the residual between-studies variances are homoscedastic or heteroscedastic across categories, as the parameters are unknown. In the absence of a formal statistic to test the homoscedasticity of the residual between-studies variances across categories, it is possible to compare the model fit using separate or pooled estimates.

Future Research Directions

Additional simulation studies are needed to assess the performance of the methods under more adverse conditions, such as a non-normal distribution for the true effects within each category of the moderator. Furthermore, in the future the development of a statistical test to assess the homoscedasticity of the residual between-studies variances across categories of a moderator could be an important contribution in the context of subgroup analyses in meta-analysis.

References

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A review of meta-analyses in education: Methodological strengths and weaknesses. *Review of Educational Research*, 82, 436-476. doi: 10.3102/0034654312458162

Borenstein, M., Hedges, L.V., Higgins, J. P. T., & Rothstein, H. R. (2017). *Comprehensive Meta-Analysis* (Vers. 3.3). Englewood, NJ: Biostat Inc.

Borenstein, M., Hedges, L.V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: Wiley. doi:10.1002/ 9780470743386.

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

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of research synthesis and meta-analysis* (2nd ed.). New York, NY: Russell Sage Foundation.

Davey, J., Turner, R. M., Clarke, M. J., & Higgins, J. P. T. (2011). Characteristics of meta-analyses and their component studies in the Cochrane Database of Systematic Reviews: A cross-sectional, descriptive analysis. *BMC Medical Research Methodology*, 11, 160. doi: 10.1186/1471-2288-11-160

- DerSimonian, R., & Laird, N. (1986). Meta-analysis of clinical trials. *Controlled Clinical Trials*, 7, 177–188. doi:10.1016/0197-2456(86)90046-2.
- Gonzalez-Mulé, E., & Aguinis, H. (2017). Advancing theory by assessing boundary conditions with metaregression: A critical review and best-practice recommendations. *Journal of Management*. Advance online publication. doi:10.1177/0149206317710723
- Harbord, R. M., & Higgins, J. P. T. (2008). Meta-regression in Stata. *The Stata Journal*, 8, 493-519.
- Hartung, J. (1999). An alternative method for meta-analysis. *Biometrical Journal*, 41, 901-916. doi:10.1002/(SICI)1521-4036(199912)41:8<901::AID-BIMJ901>3.0.CO;2-W.
- Hartung, J., Argac, D., & Makambi, K. H. (2002). Small sample properties of tests on homogeneity in one-way Anova and meta-analysis. *Statistical Papers*, 43, 197-235.
- Hartung, J., Makambi, K. H., & Argac, D. (2001). An extended ANOVA *F*-test with application to the heterogeneity problem in meta-analysis. *Biometrical Journal*, 43, 135-146. doi: 10.1002/1521-4036(200105)43:2<135::AID-BIMJ135>3.0.CO;2-H
- Hedges, L.V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. San Diego, CA: Academic Press.
- Huizenga, H. M., Visser, I., & Dolan, C. V. (2011). Testing overall and moderator effects in random effects meta-regression. *British Journal of Mathematical and Statistical Psychology*, 64, 1–19. doi:10.1348/000711010X52268

- Knapp, G., & Hartung J. (2003). Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, 22, 2693–2710.
doi:10.1002/sim.1482
- Langan, D., Higgins, J., & Simmonds, M. (2017). Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. *Research Synthesis Methods*, 8, 181-198.
- Lau, J., Ioannidis, J. P. A., & Schmid, C. H. (1998). Summing up evidence: One answer is not always enough. *Lancet*, 351, 123-127. doi: 10.1016/S0140-6736(97)08468-7
- Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American psychologist*, 48, 1181. doi: 10.1037/0003-066X.48.12.1181
- López-López, J. A., Marín-Martínez, F., Sánchez-Meca, J., Van den Noortgate, W., & Viechtbauer, W. (2014). Estimation of the predictive power of the model in mixed-effects meta-regression models: A simulation study. *British Journal of Mathematical and Statistical Psychology*, 67, 30–48. doi:10.1111/bmsp.12002
- Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78, 47-55.
- Paule, R. C., & Mandel, J. (1982). Consensus values and weighting factors. *Journal of Research of the National Bureau of Standards*, 87, 377–385.
- Rubio-Aparicio, M., Sánchez-Meca, J., López-López, J. A, Marín-Martínez, F., & Botella, J. (2017). Analysis of categorical moderators in mixed-effects meta-analysis: Consequences of using pooled vs. separate estimates of the residual between-studies variances. *British Journal of Mathematical and Statistical Psychology*, 70, 439-456. doi: 10.1111/bmsp.12092

- Sánchez-Meca, J., & Marín-Martínez, F. (2008). Confidence intervals for the overall effect size in random-effects meta-analysis. *Psychological Methods, 13*, 31–48. doi:10.1037/1082-989X.13.1.31
- Schmidt, F. L., & Hunter, J. E. (2015). *Methods of meta-analysis: Correcting error and bias in research findings* (3rd ed.). Los Angeles, CA: Sage.
- Sidik, K., & Jonkman, J. N. (2005). A note on variance estimation in random effects meta-regression. *Journal of Biopharmaceutical Statistics, 15*, 823-838. doi:10.1081/BIP-200067915
- Thompson, S. G. (1994). Why sources of heterogeneity in meta-analysis should be investigated. *British Medical Journal, 309*, 1351-1355.
- Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine, 21*, 1559–1573. doi:10.1002/sim.1187
- Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., Kuss, O., Higgins, J. P. T., Langan, D., & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research Synthesis Methods, 7*, 55–79. doi: 10.1002/jrsm.1164
- Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics, 30*, 261–293. doi:10.3102/10769986030003261
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*, 1–48.
- Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

regression models. *Psychological Methods*, 20, 360-374. doi:
10.1037/met0000023

For Peer Review

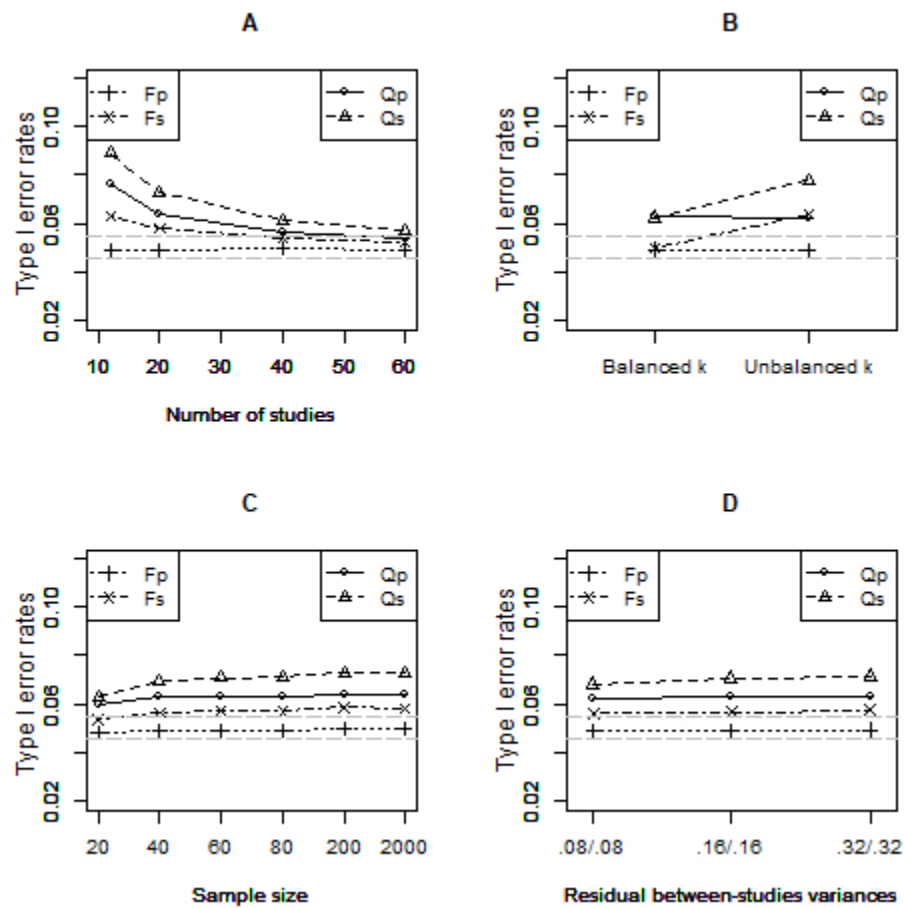


Figure 1. Average Type I error rates in scenarios with homoscedastic residual between-studies variances across the two categories of the moderator.

169x169mm (72 x 72 DPI)

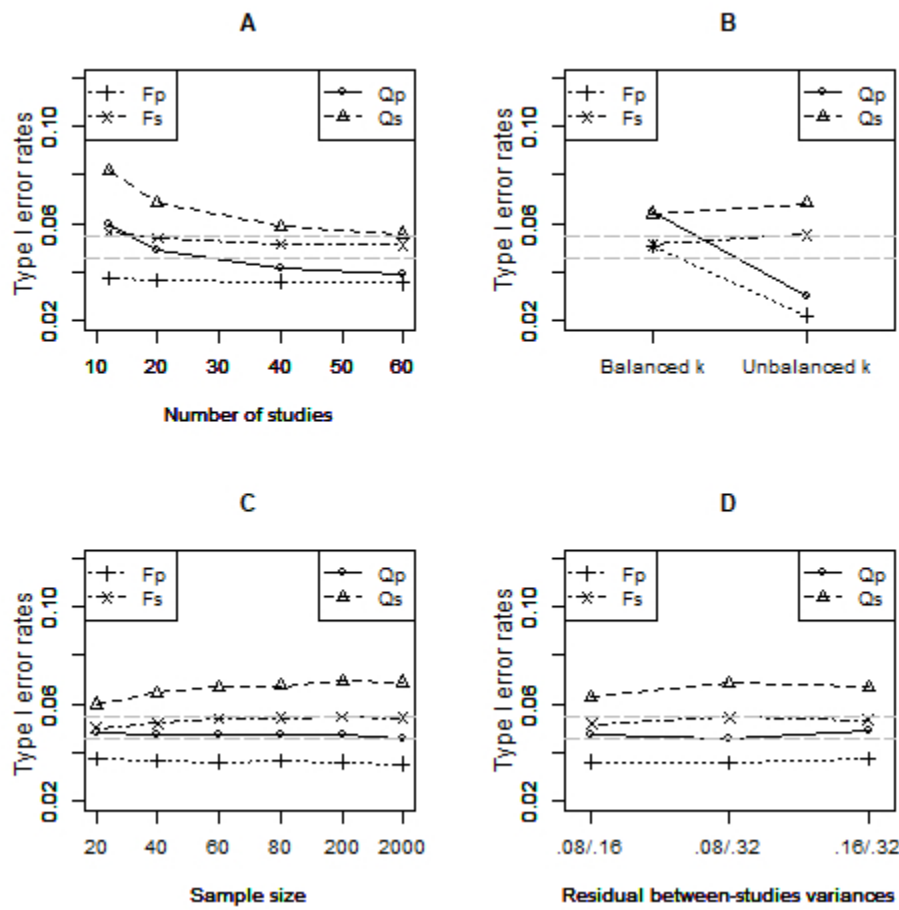


Figure 2. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the two categories of the moderator and smaller variance in the smaller category.

169x169mm (72 x 72 DPI)

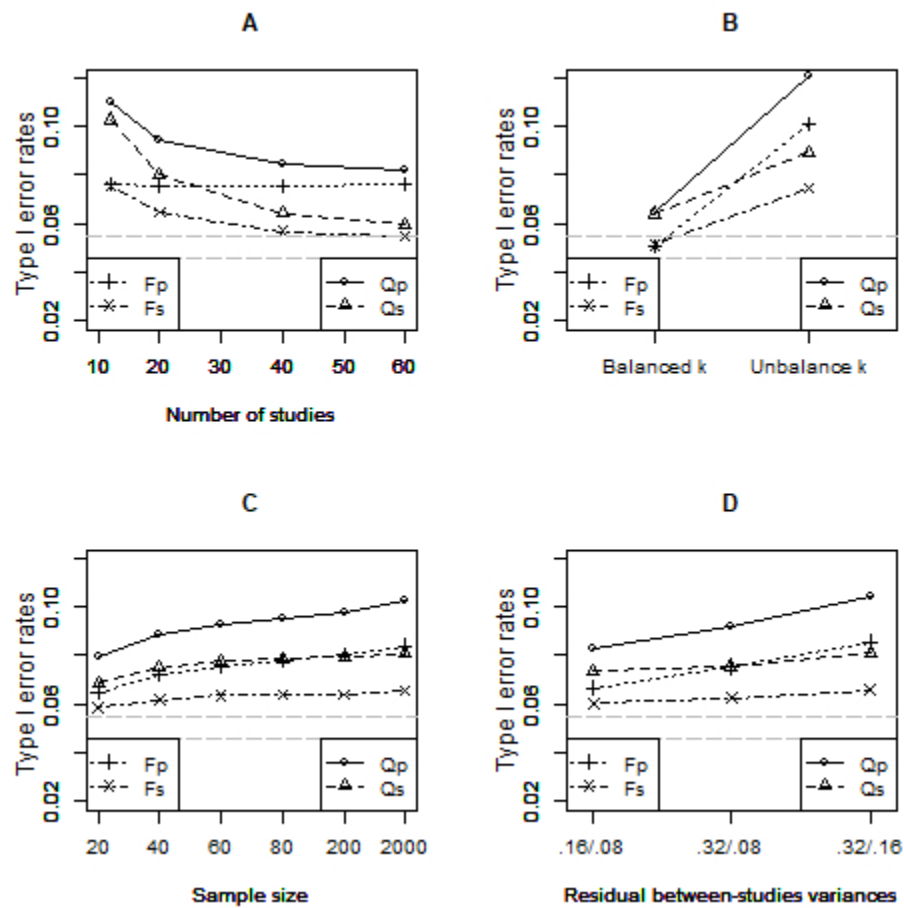


Figure 3. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the two categories of the moderator and larger variance in the smaller category.

169x169mm (72 x 72 DPI)

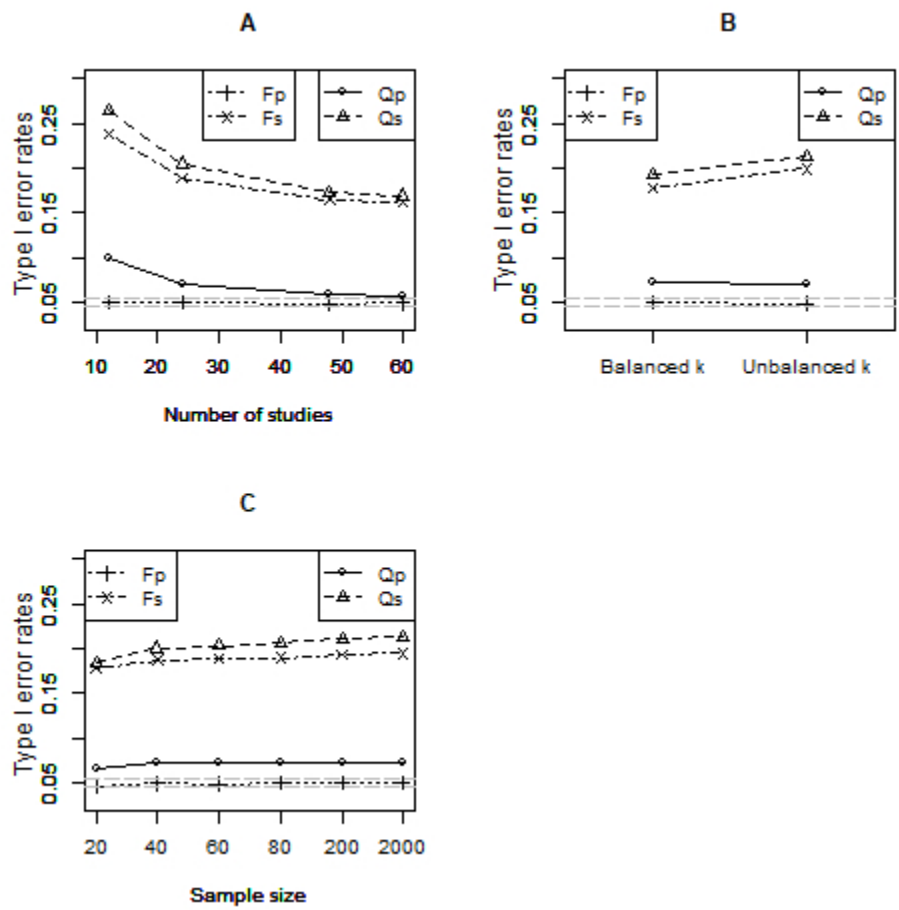


Figure 4. Average Type I error rates in scenarios with homoscedastic residual between-studies variances across the three categories of the moderator.

169x169mm (72 x 72 DPI)

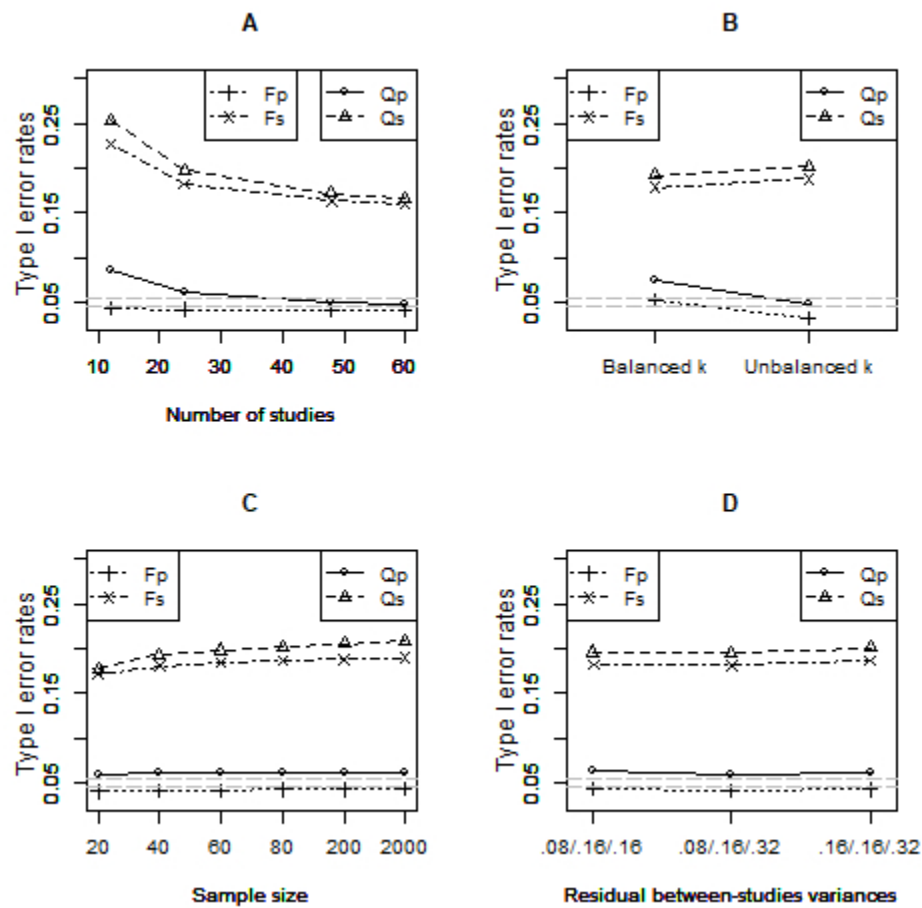


Figure 5. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the three categories of the moderator and smaller variance in the smaller category.

169x169mm (72 x 72 DPI)

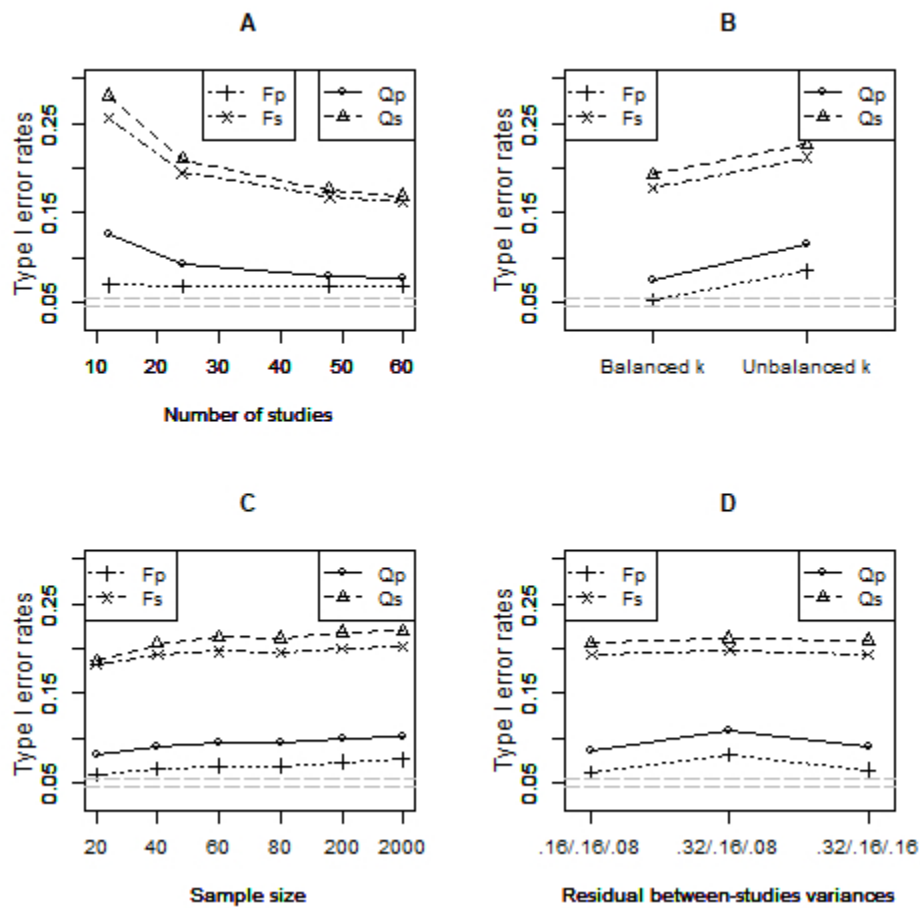


Figure 6. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the three categories of the moderator and larger variance in the smaller category.

169x169mm (72 x 72 DPI)

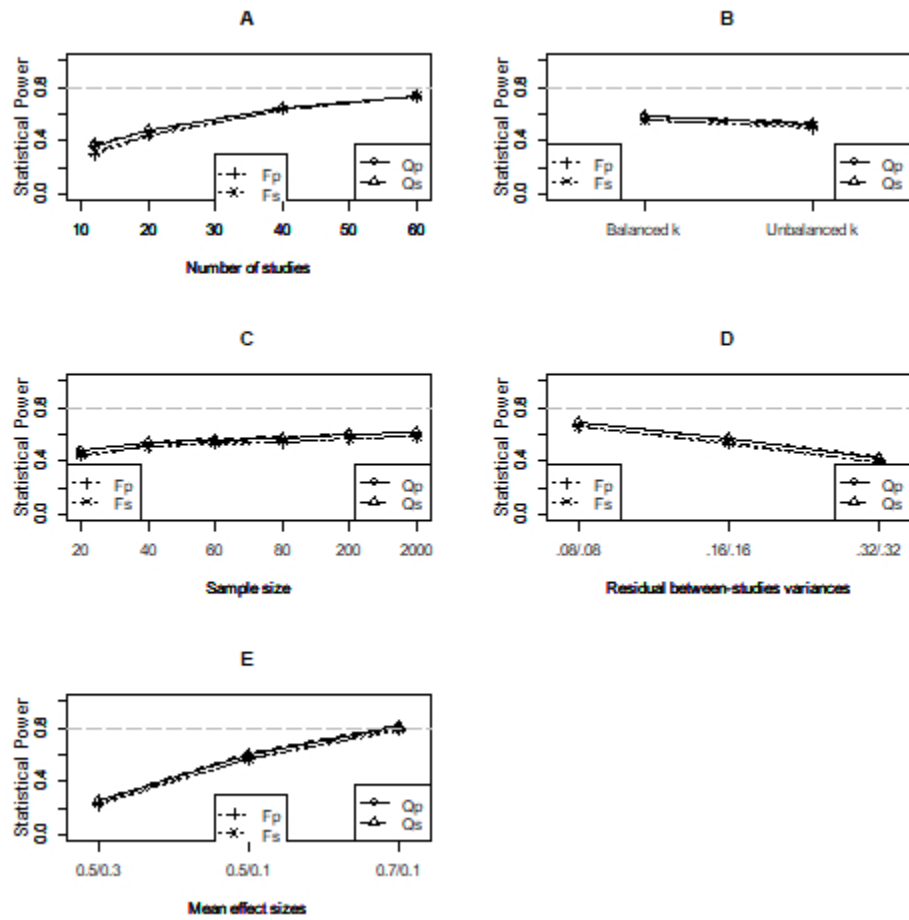


Figure 7. Average power rates in scenarios with homoscedastic residual between-studies variances across the two categories of the moderator.

169x169mm (72 x 72 DPI)

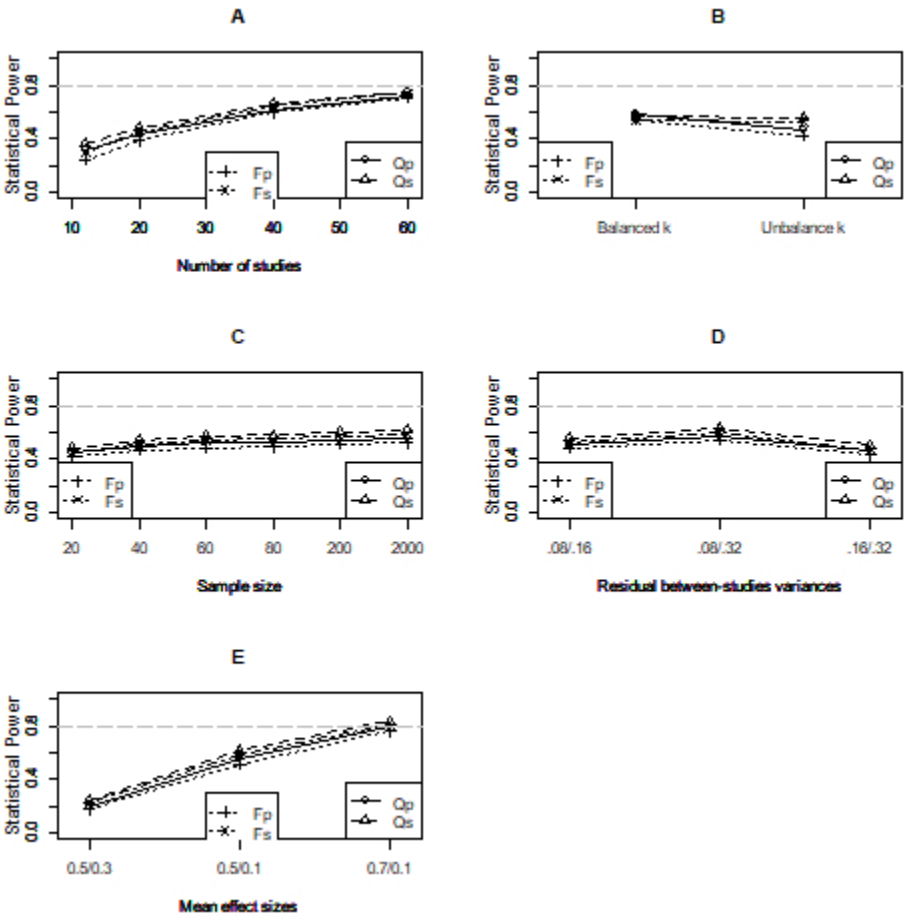


Figure 8. Average power rates in scenarios with heteroscedastic residual between-studies variances across the two categories of the moderator and smaller variance in the smaller category.

169x169mm (72 x 72 DPI)

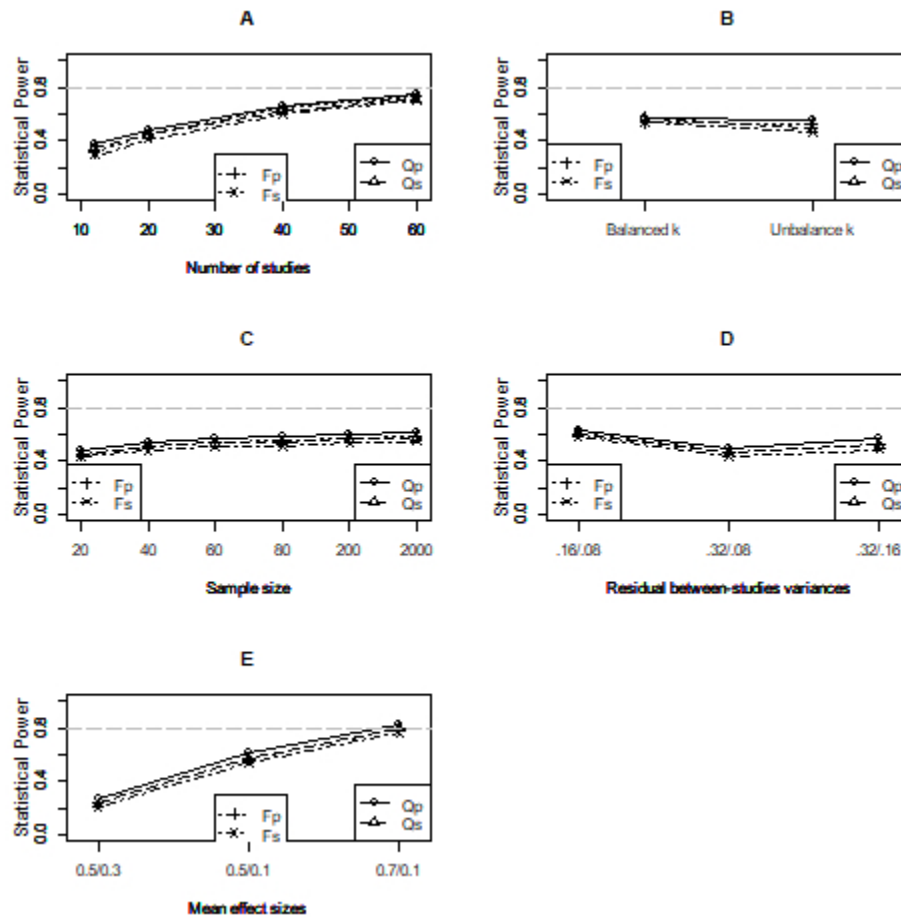


Figure 9. Average power rates in scenarios with heteroscedastic residual between-studies variances across the two categories of the moderator and larger variance in the smaller category.

169x169mm (72 x 72 DPI)

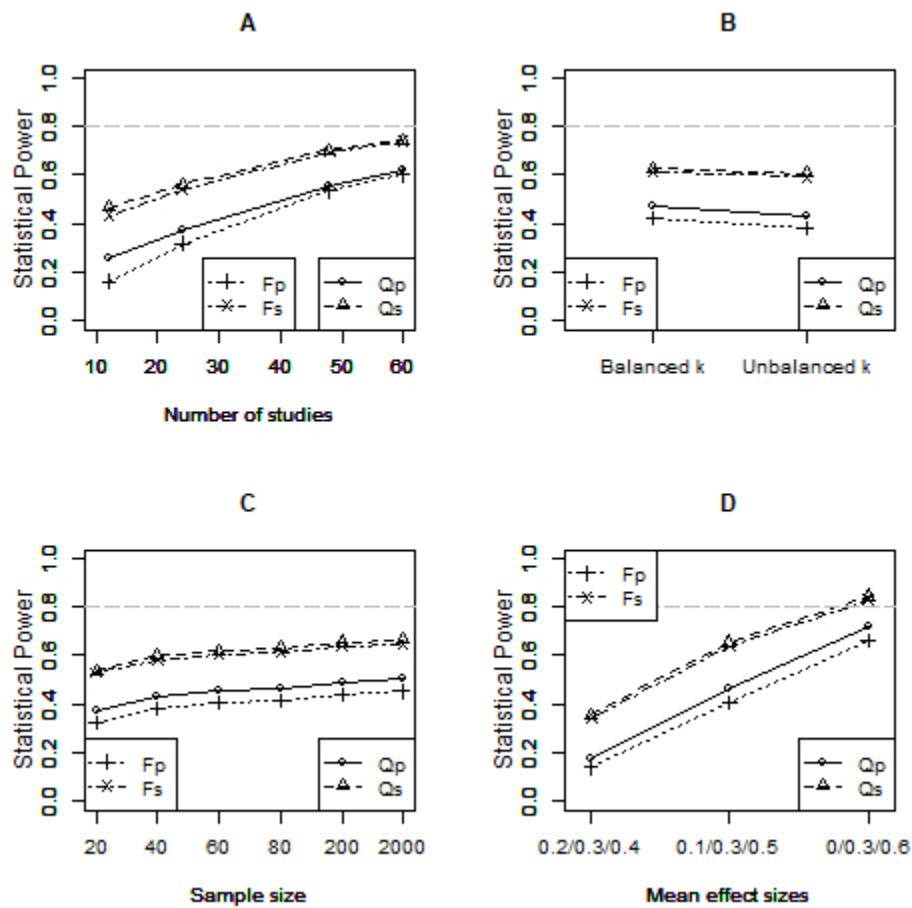


Figure 10. Average power rates in scenarios with homoscedastic residual between-studies variances across the three categories of the moderator.

169x169mm (72 x 72 DPI)

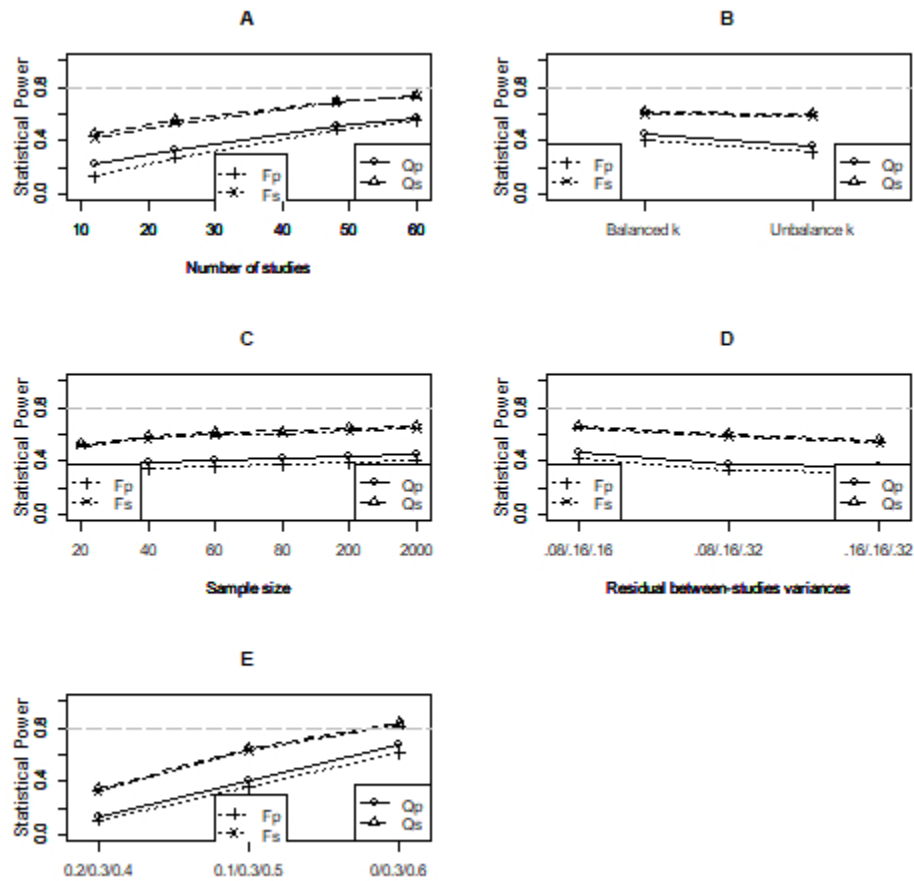


Figure 11. Average power rates in scenarios with heteroscedastic residual between-studies variances across the three categories of the moderator and smaller variance in the smaller category.

169x169mm (72 x 72 DPI)

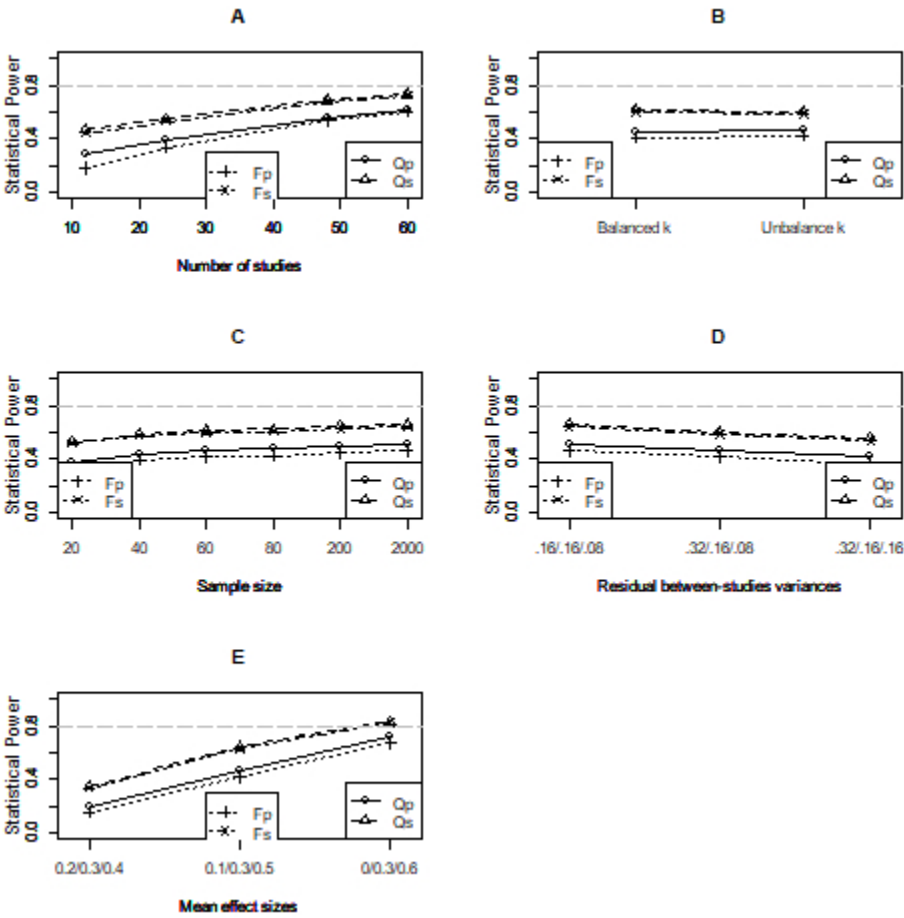
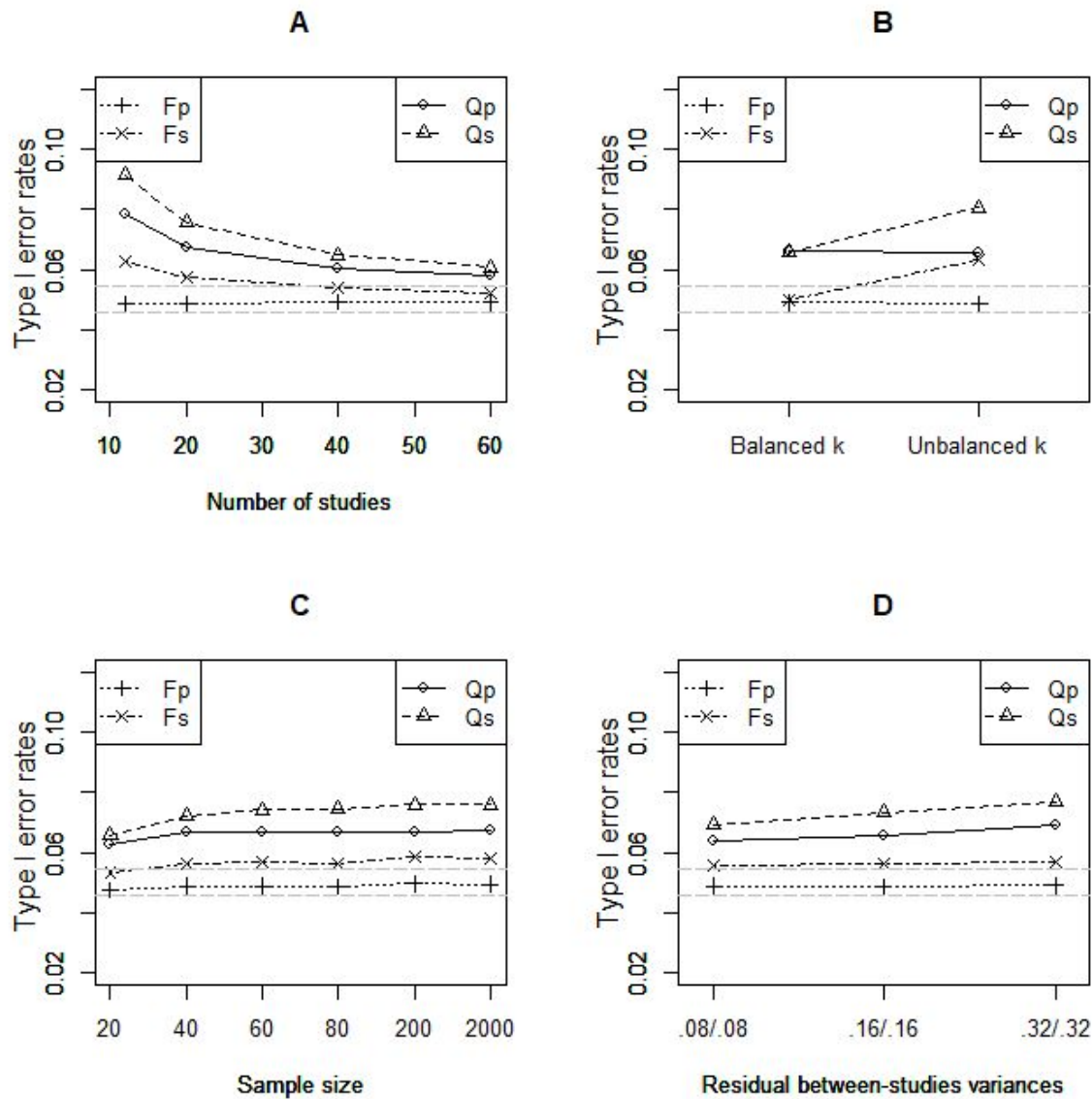


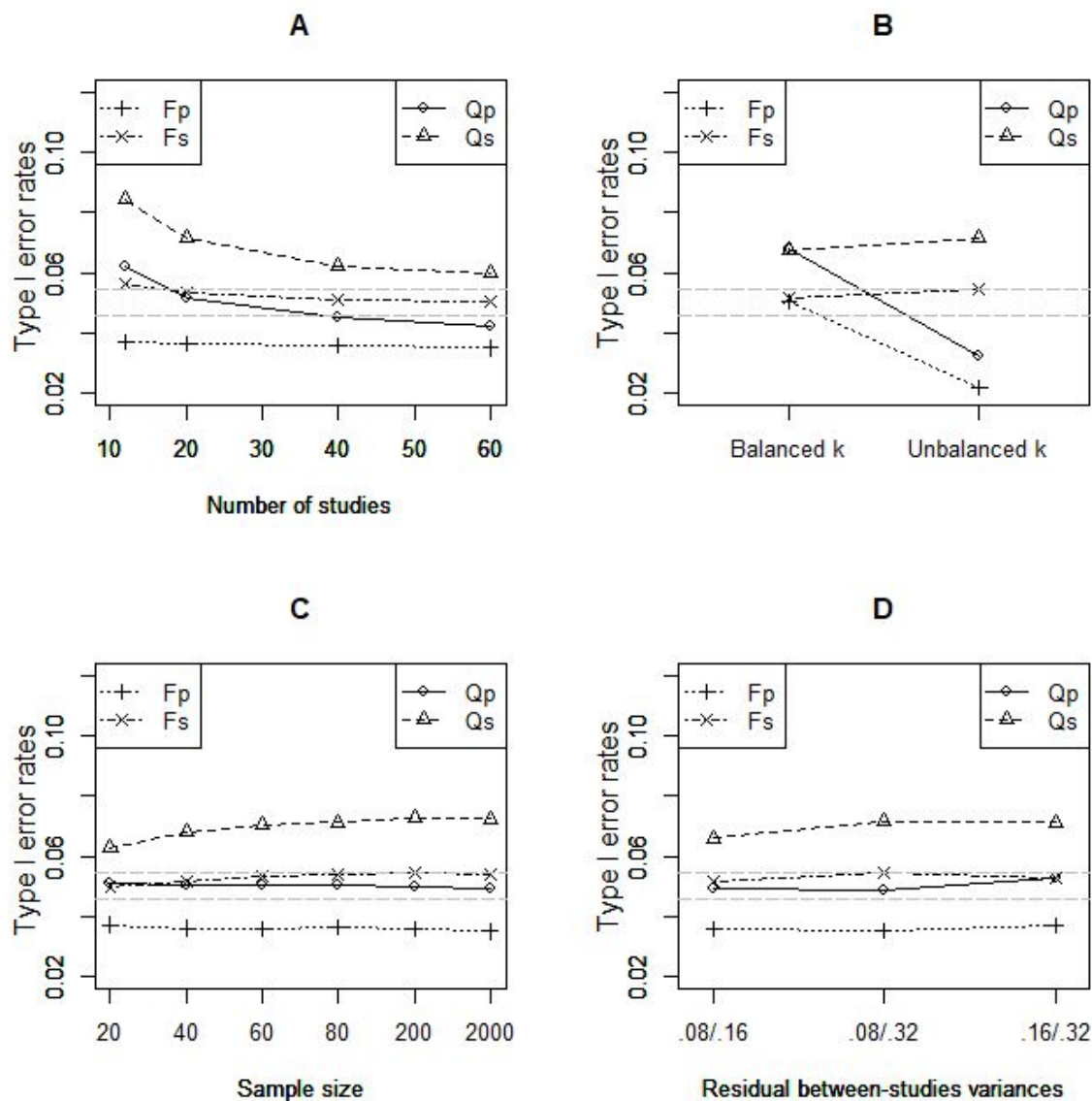
Figure 12. Average power rates in scenarios with heteroscedastic residual between-studies variances across the three categories of the moderator and larger variance in the smaller category.

169x169mm (72 x 72 DPI)

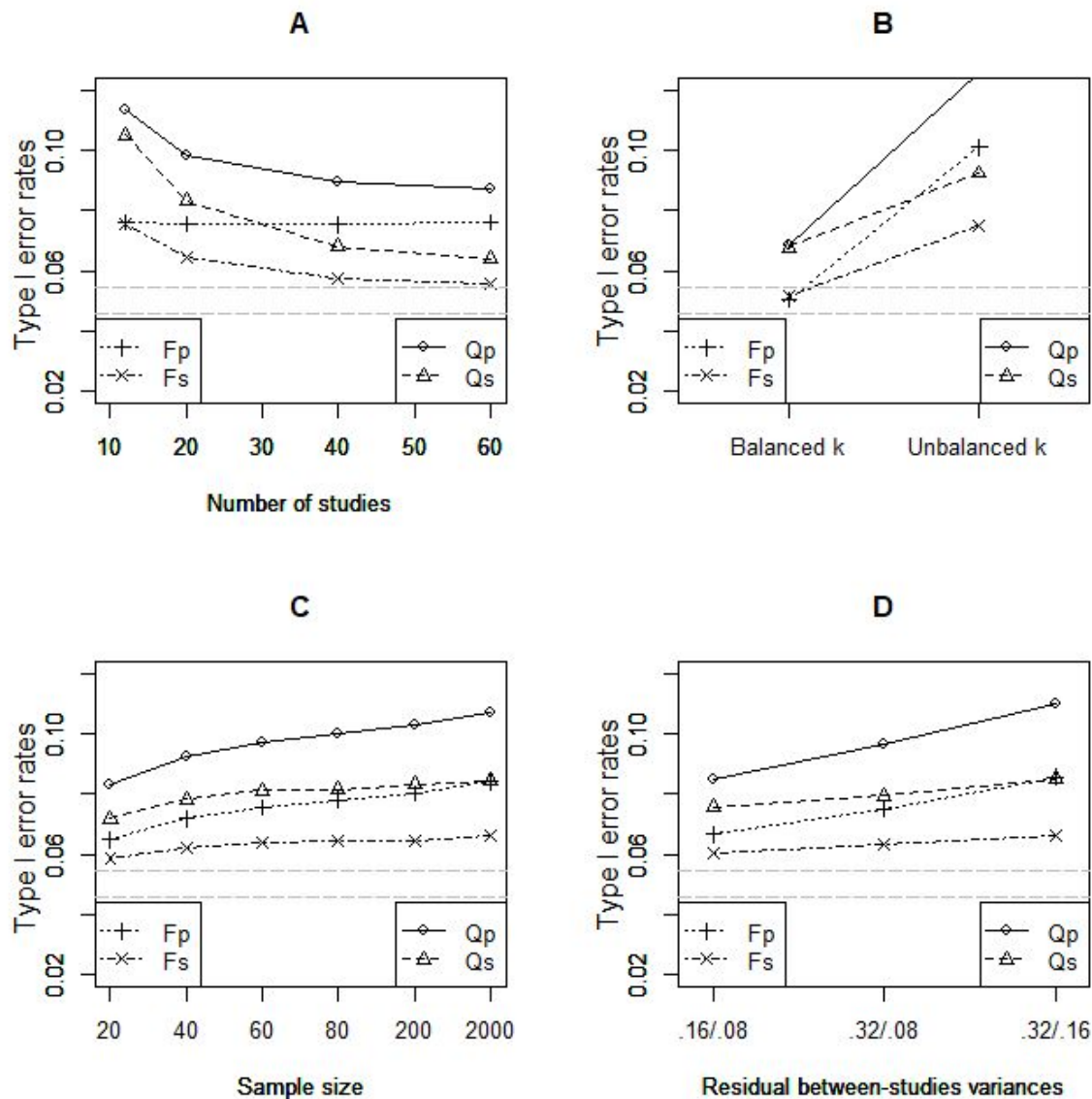
Supplementary file 2

Supplementary figures

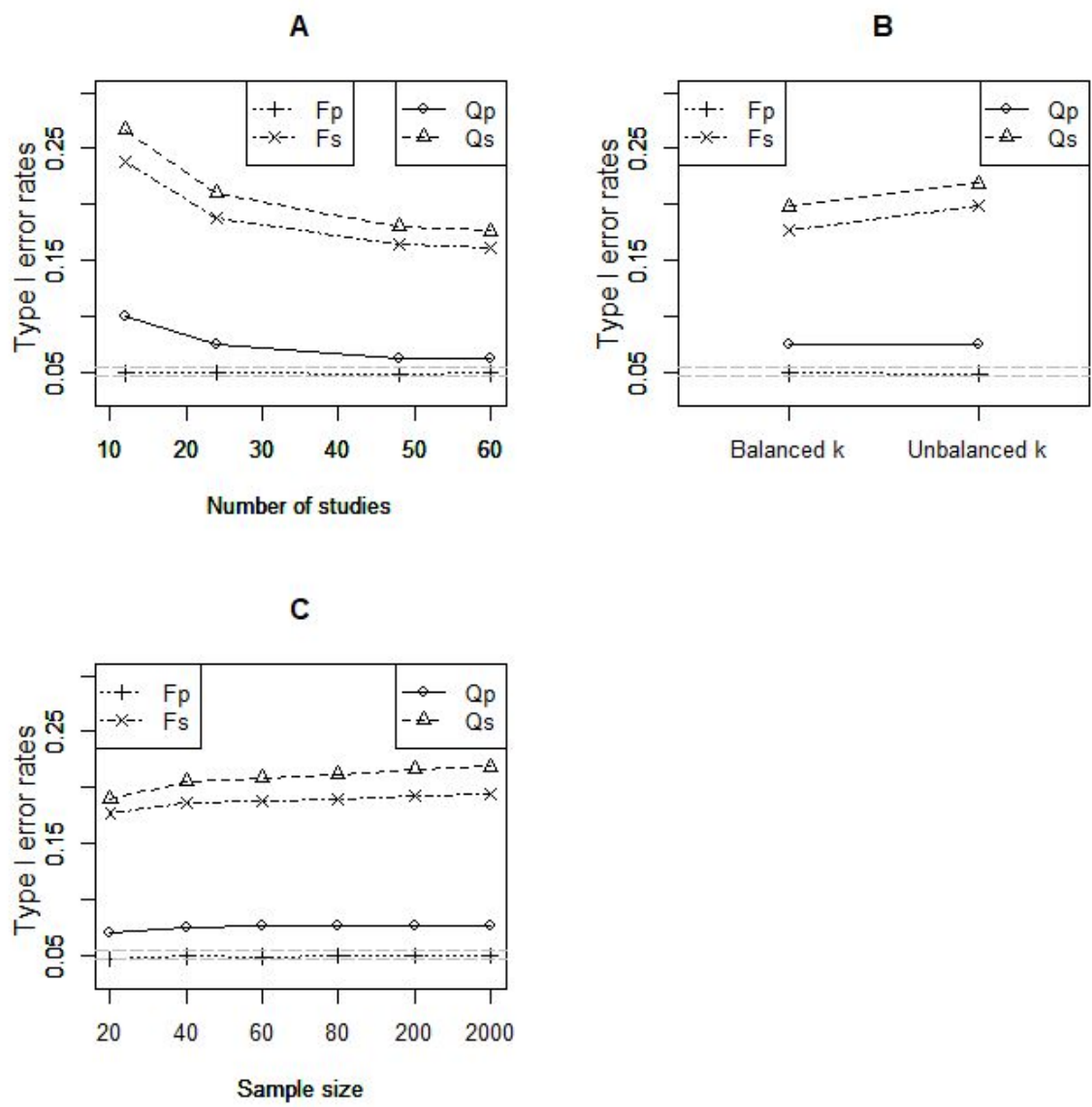
Supplementary figure 1. Average Type I error rates in scenarios with homoscedastic residual between-studies variances across the two categories of the moderator using the DerSimonian and Laird estimator.



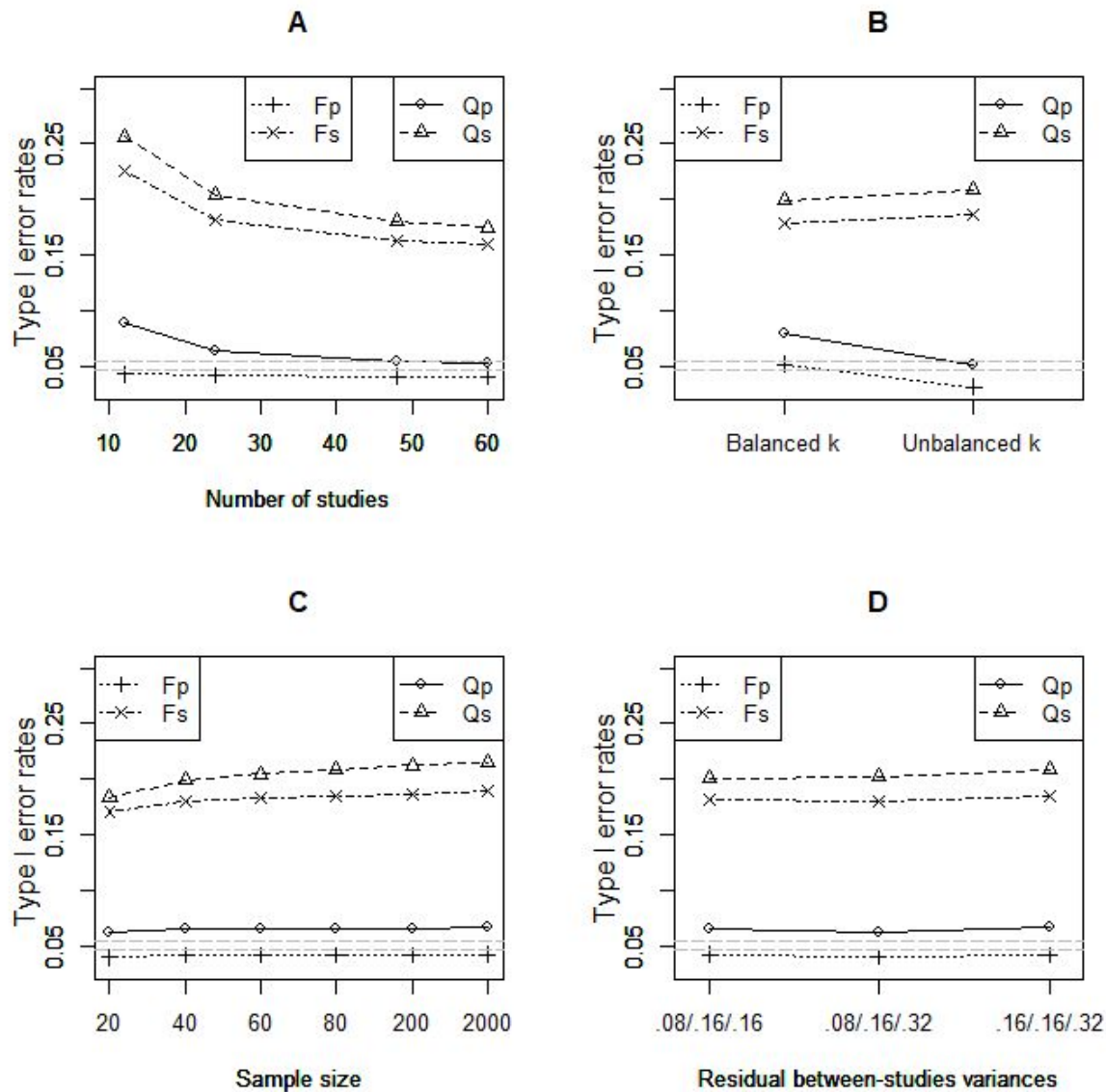
Supplementary figure 2. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the two categories of the moderator and smaller variance in the smaller category using the DerSimonian and Laird estimator.



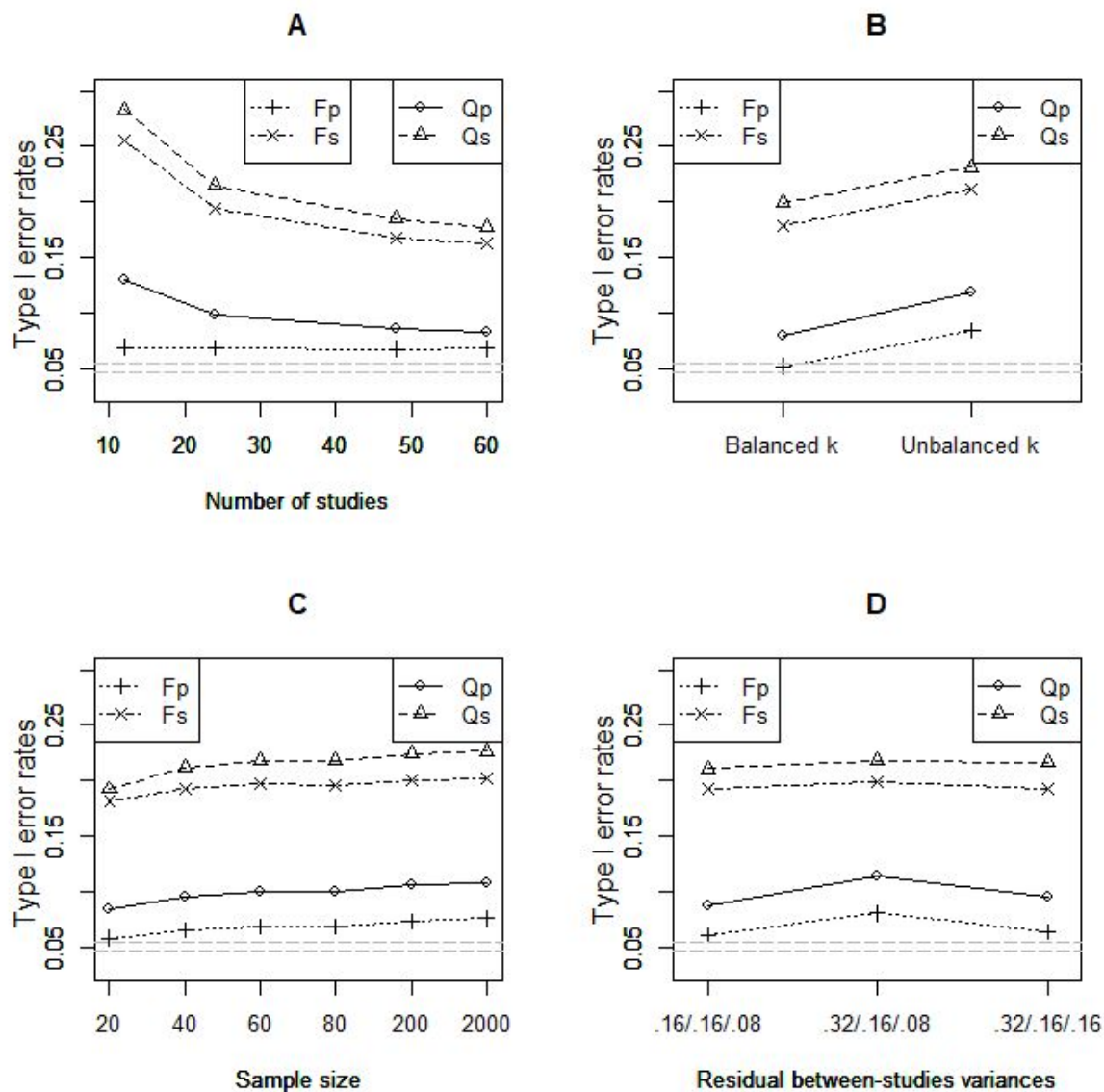
Supplementary figure 3. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the two categories of the moderator and larger variance in the smaller category using the DerSimonian and Laird estimator.



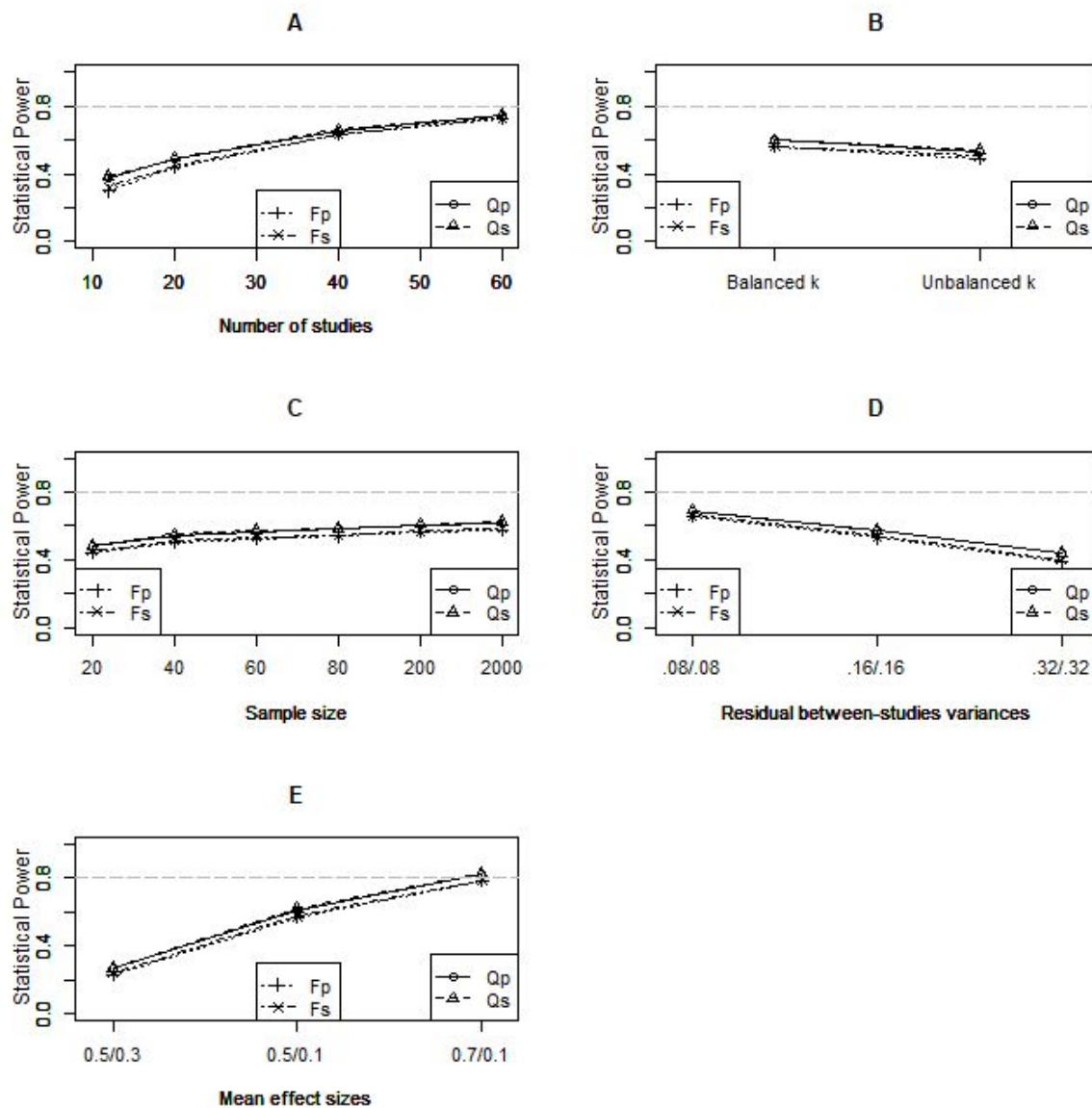
Supplementary figure 4. Average Type I error rates in scenarios with homoscedastic residual between-studies variances across the three categories of the moderator using the DerSimonian and Laird estimator.



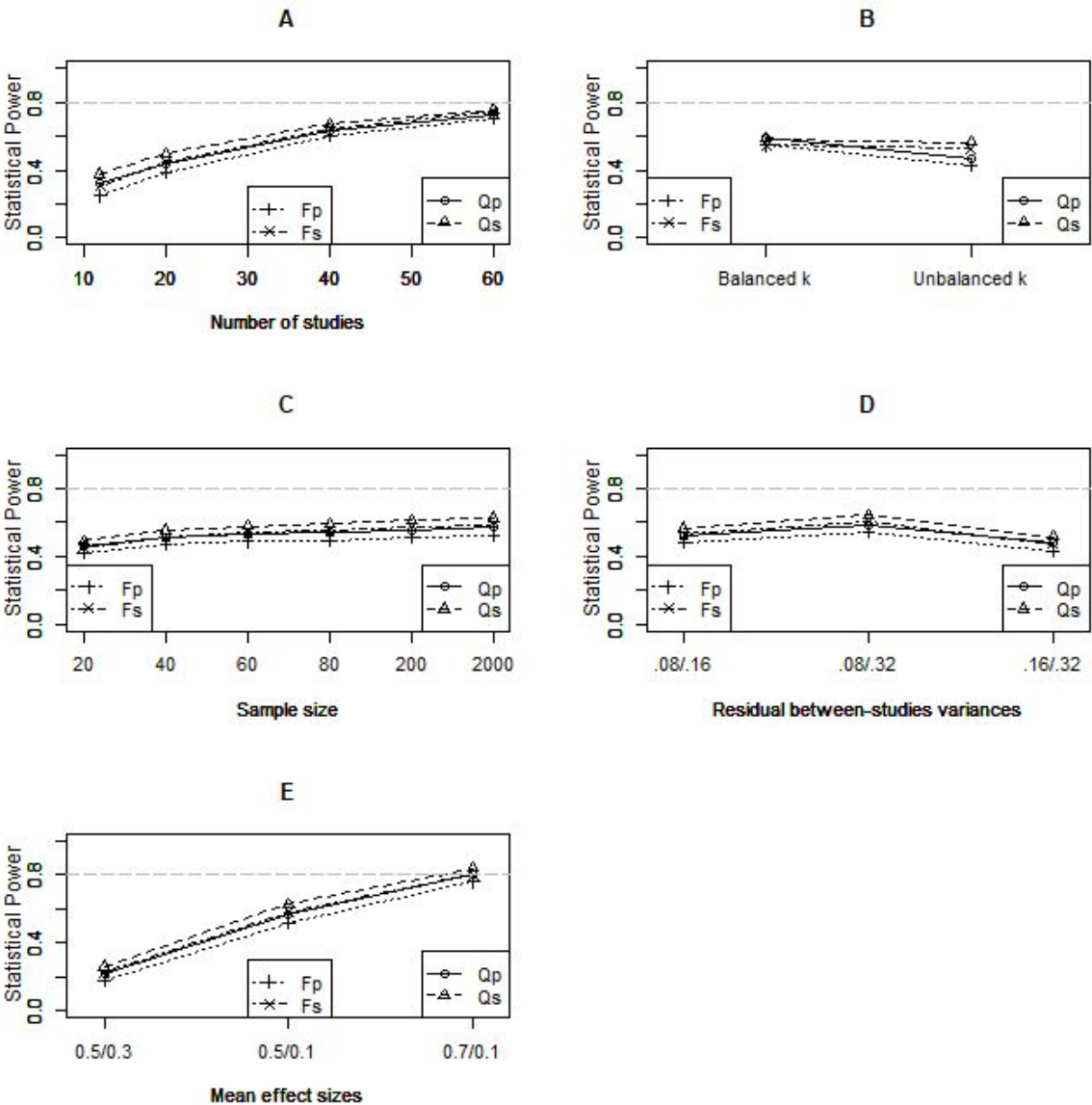
Supplementary figure 5. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the three categories of the moderator and smaller variance in the smaller category using the DerSimonian and Laird estimator.



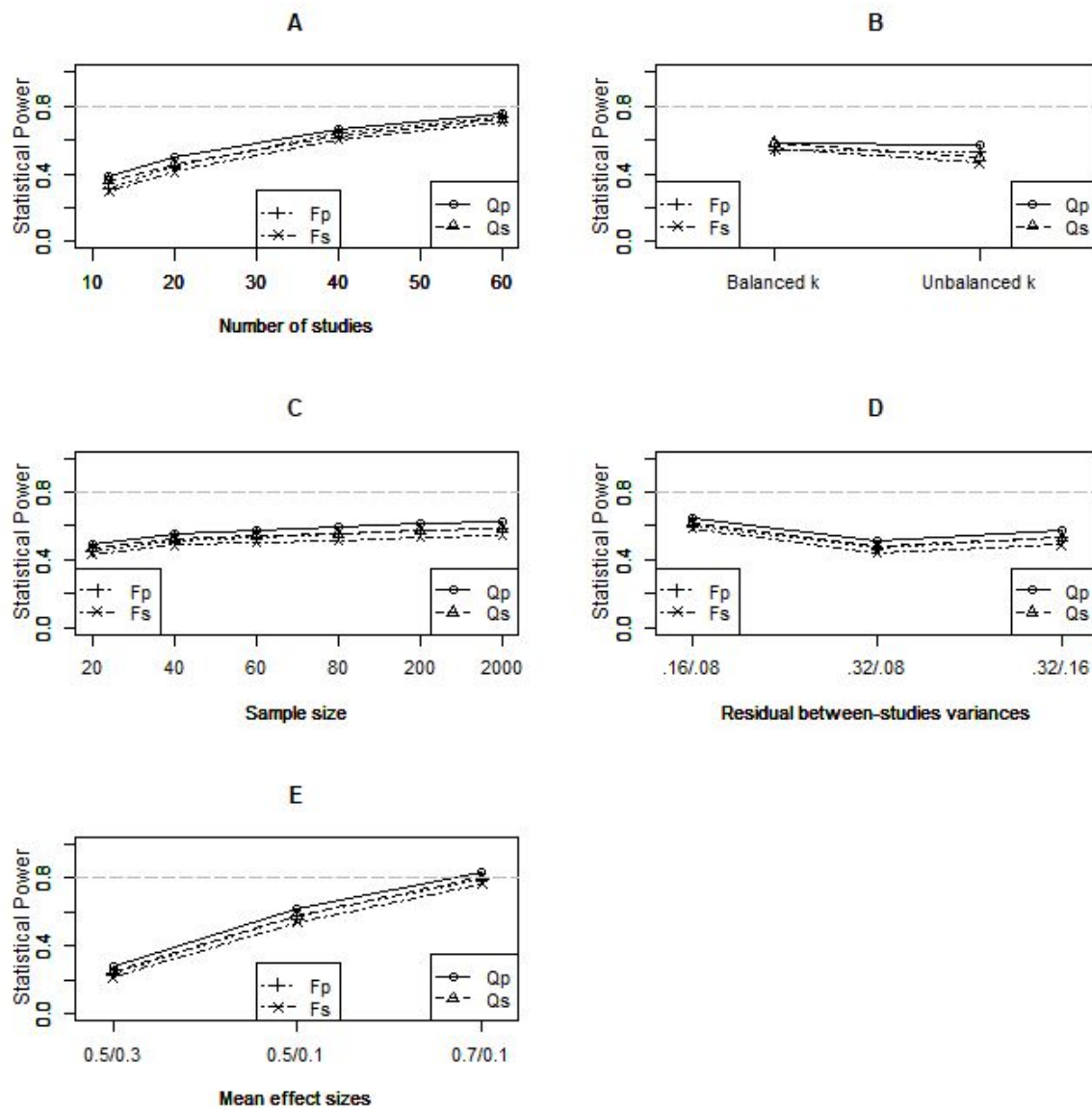
Supplementary figure 6. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the three categories of the moderator and larger variance in the smaller category using the DerSimonian and Laird estimator.



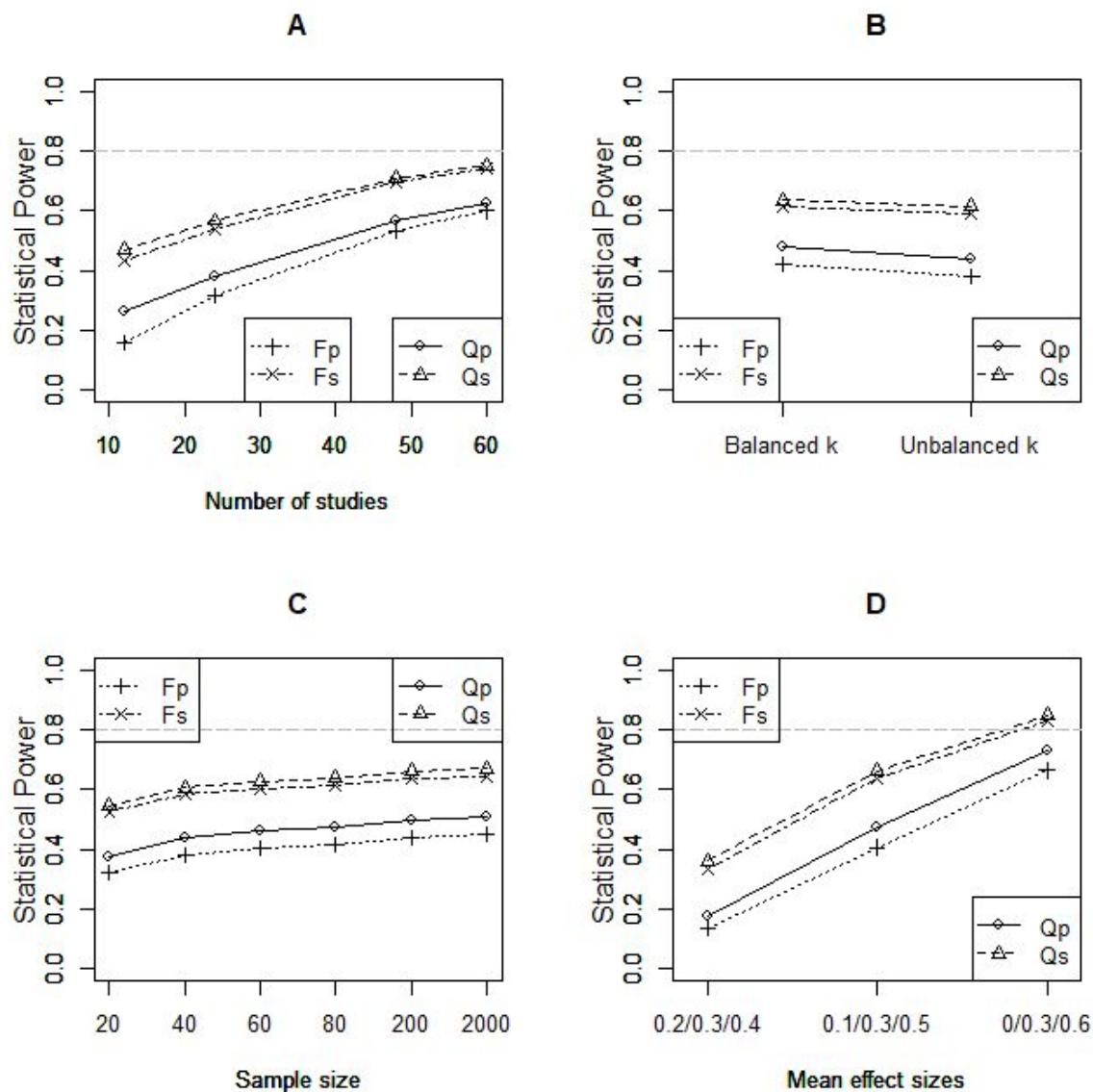
Supplementary figure 7. Average power rates in scenarios with homoscedastic residual between-studies variances across the two categories of the moderator using the DerSimonian and Laird estimator.



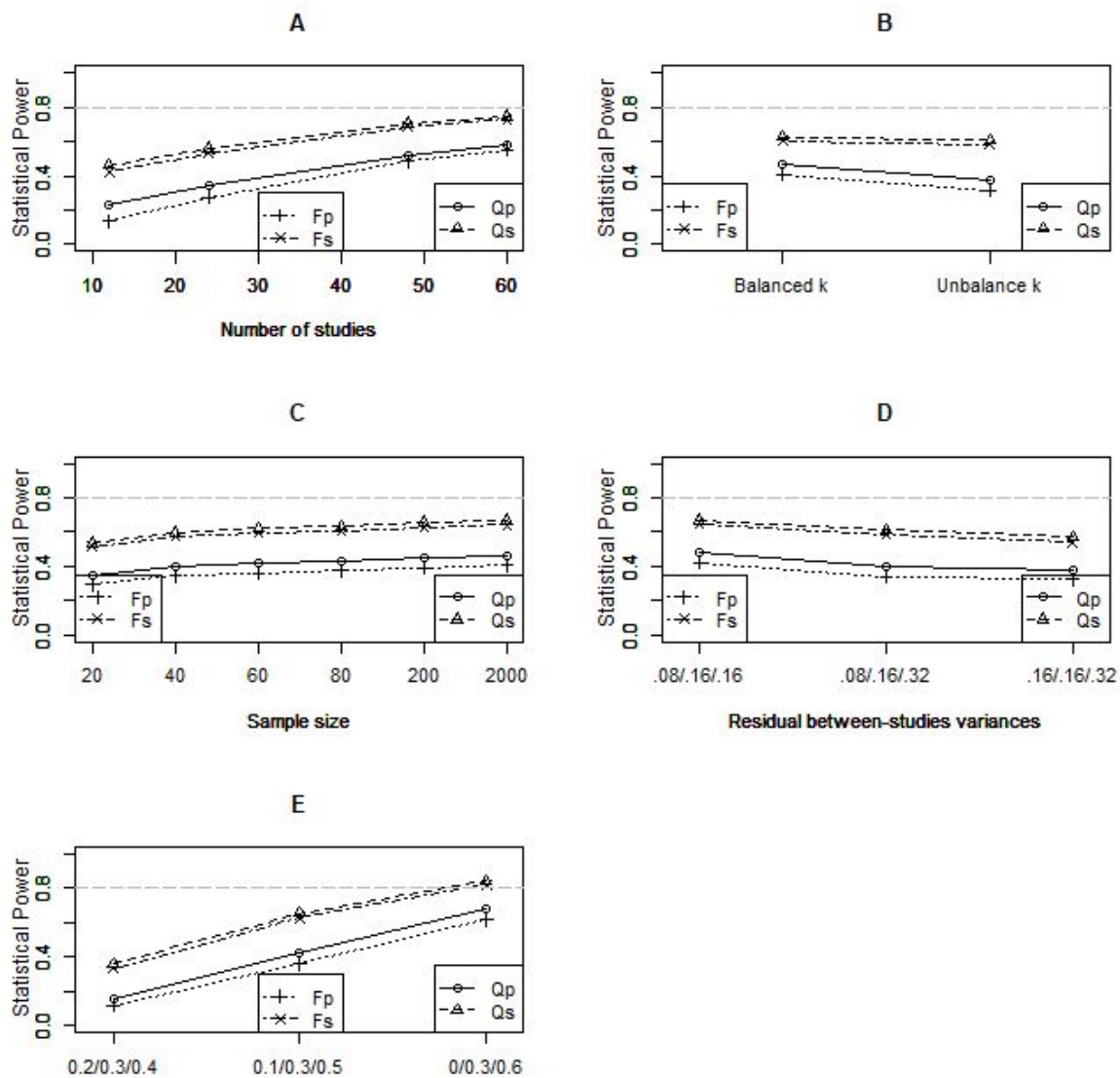
Supplementary figure 8. Average power rates in scenarios with heteroscedastic residual between-studies variances across the two categories of the moderator and smaller variance in the smaller category using the DerSimonian and Laird estimator.



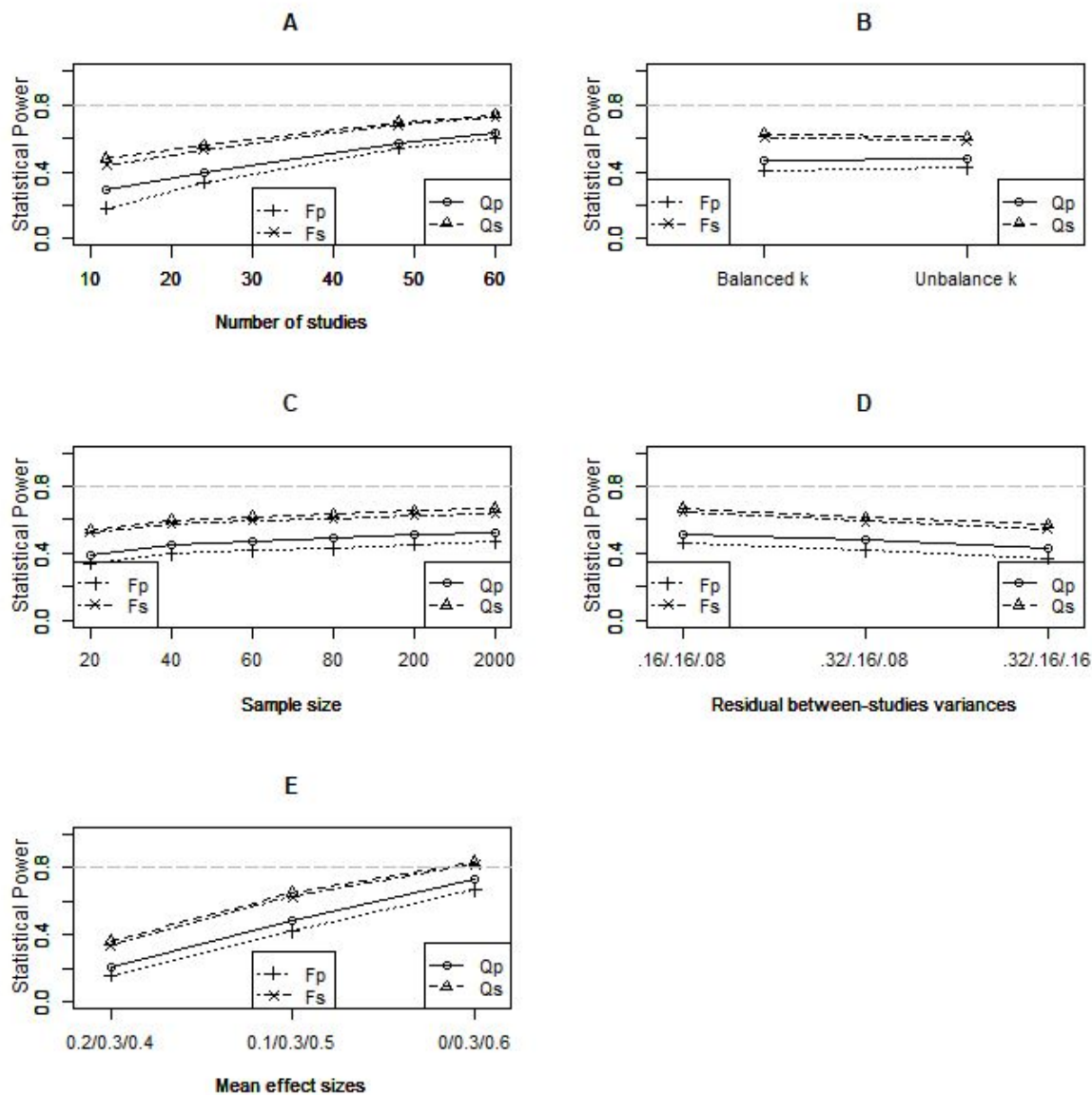
Supplementary figure 9. Average power rates in scenarios with heteroscedastic residual between-studies variances across the two categories of the moderator and larger variance in the smaller category using the DerSimonian and Laird estimator.



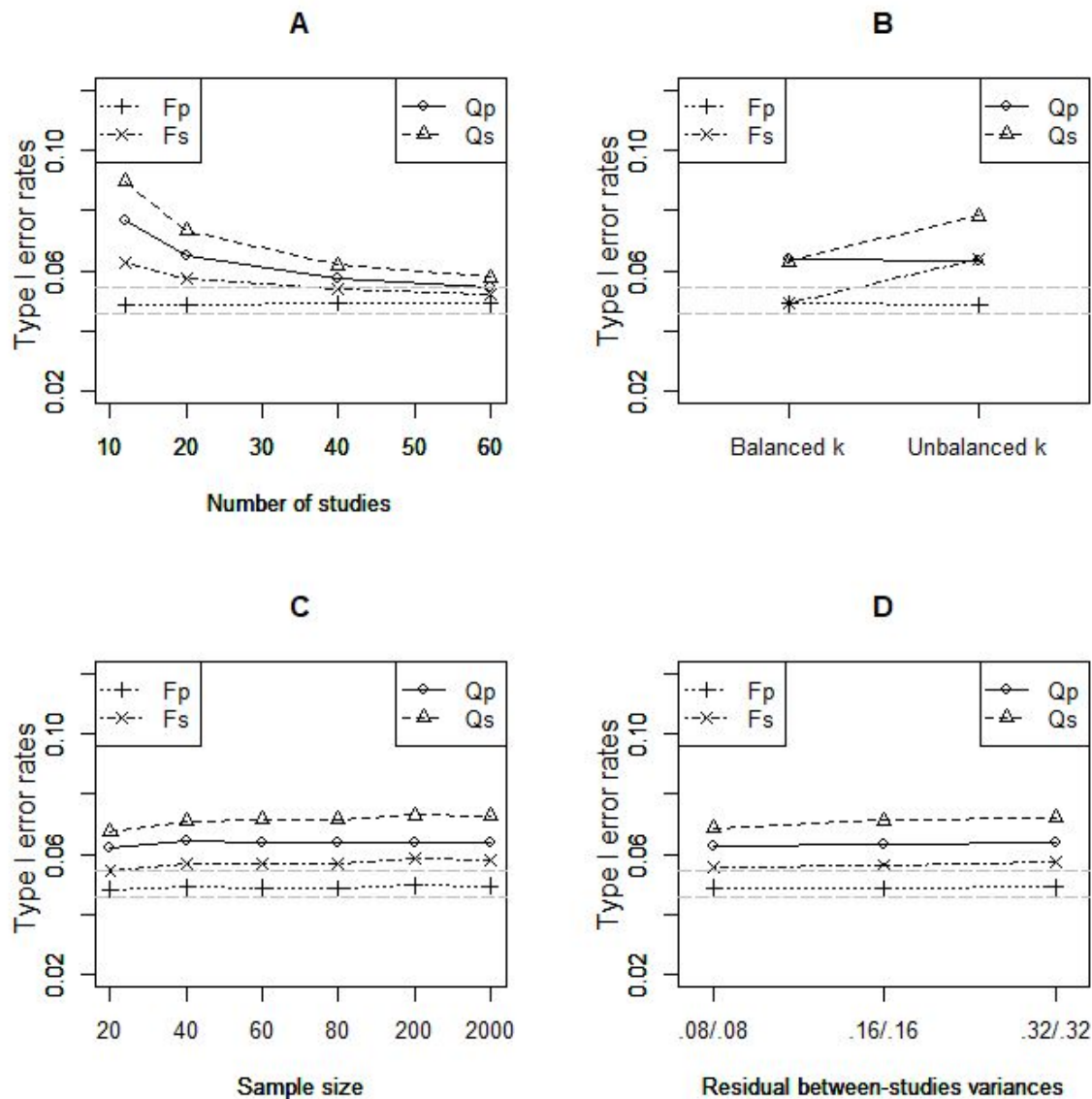
Supplementary figure 10. Average power rates in scenarios with homoscedastic residual between-studies variances across the three categories of the moderator using the DerSimonian and Laird estimator.



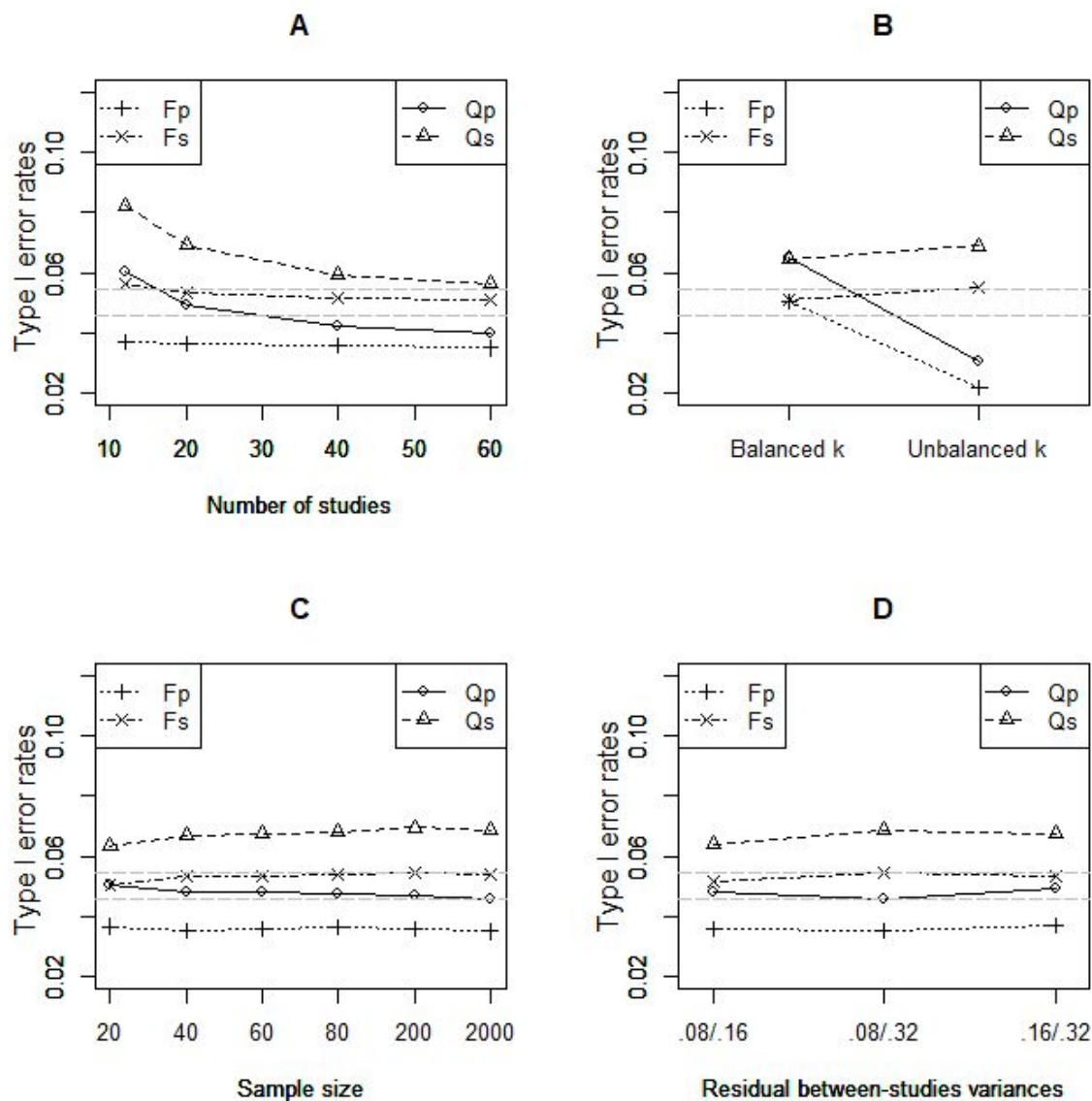
Supplementary figure 11. Average power rates in scenarios with heteroscedastic residual between-studies variances across the three categories of the moderator and smaller variance in the smaller category using the DerSimonian and Laird estimator.



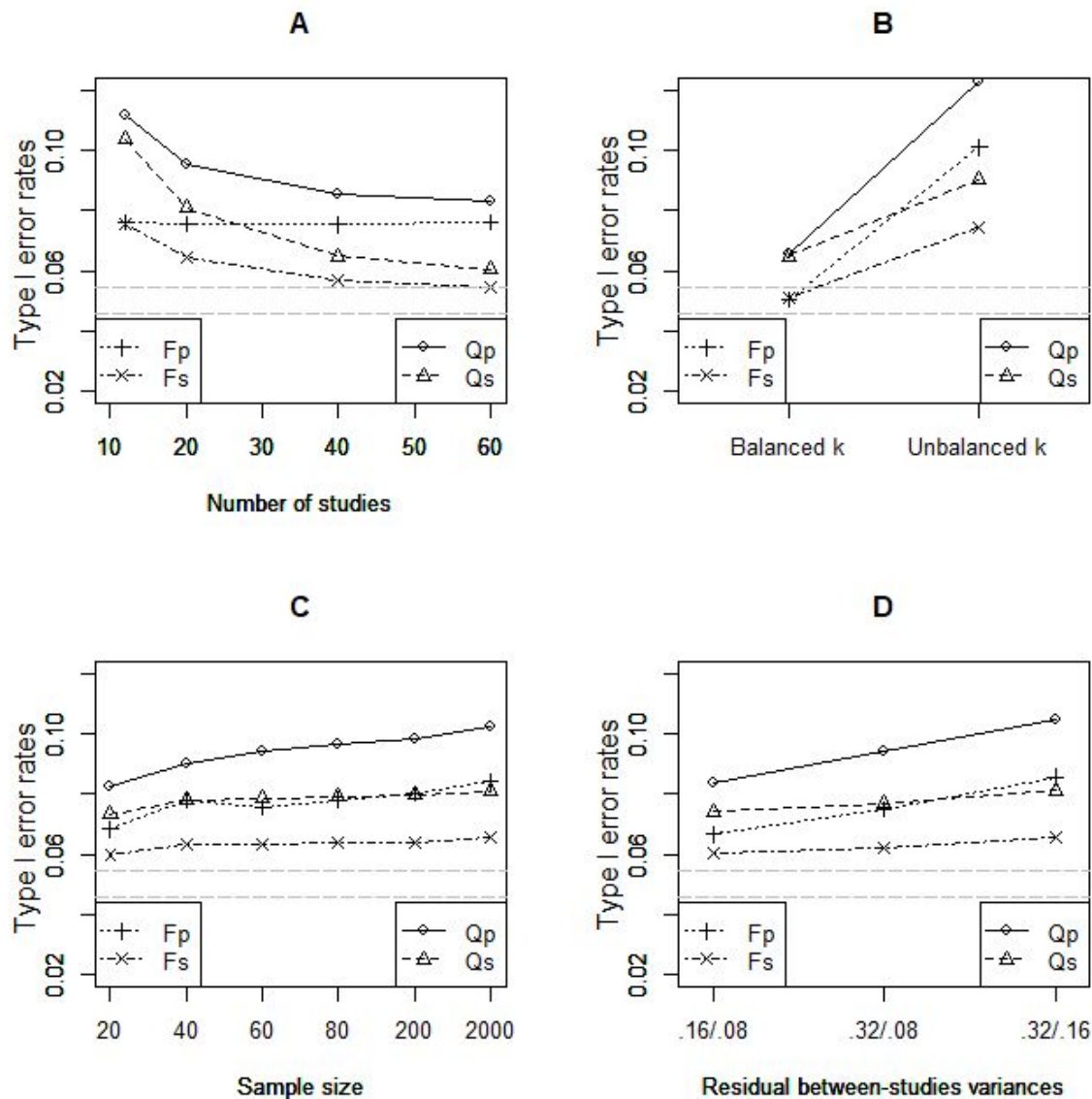
Supplementary figure 12. Average power rates in scenarios with heteroscedastic residual between-studies variances across the three categories of the moderator and larger variance in the smaller category using the DerSimonian and Laird estimator.



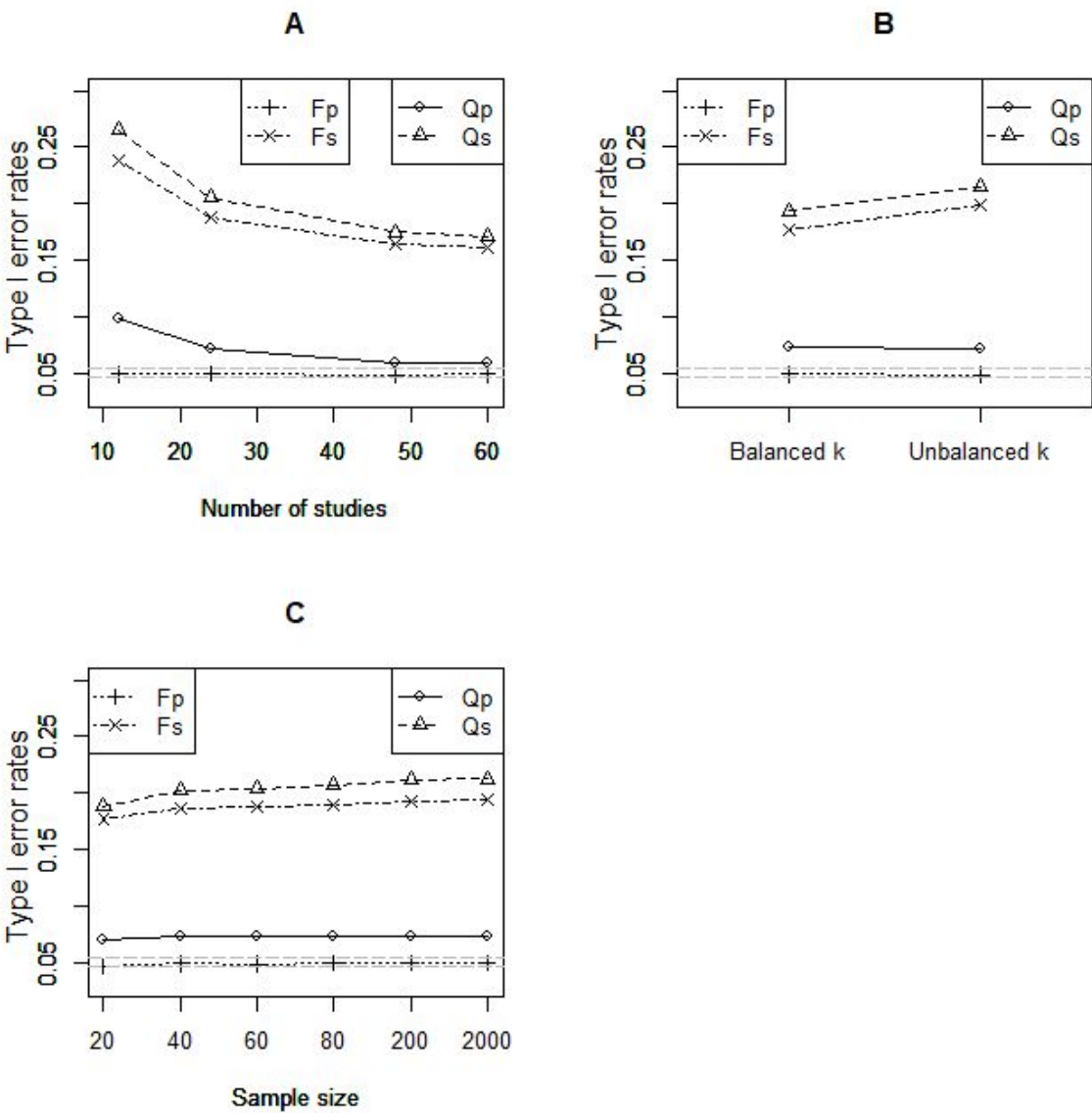
Supplementary figure 13. Average Type I error rates in scenarios with homoscedastic residual between-studies variances across the two categories of the moderator using the restricted maximum likelihood estimator.



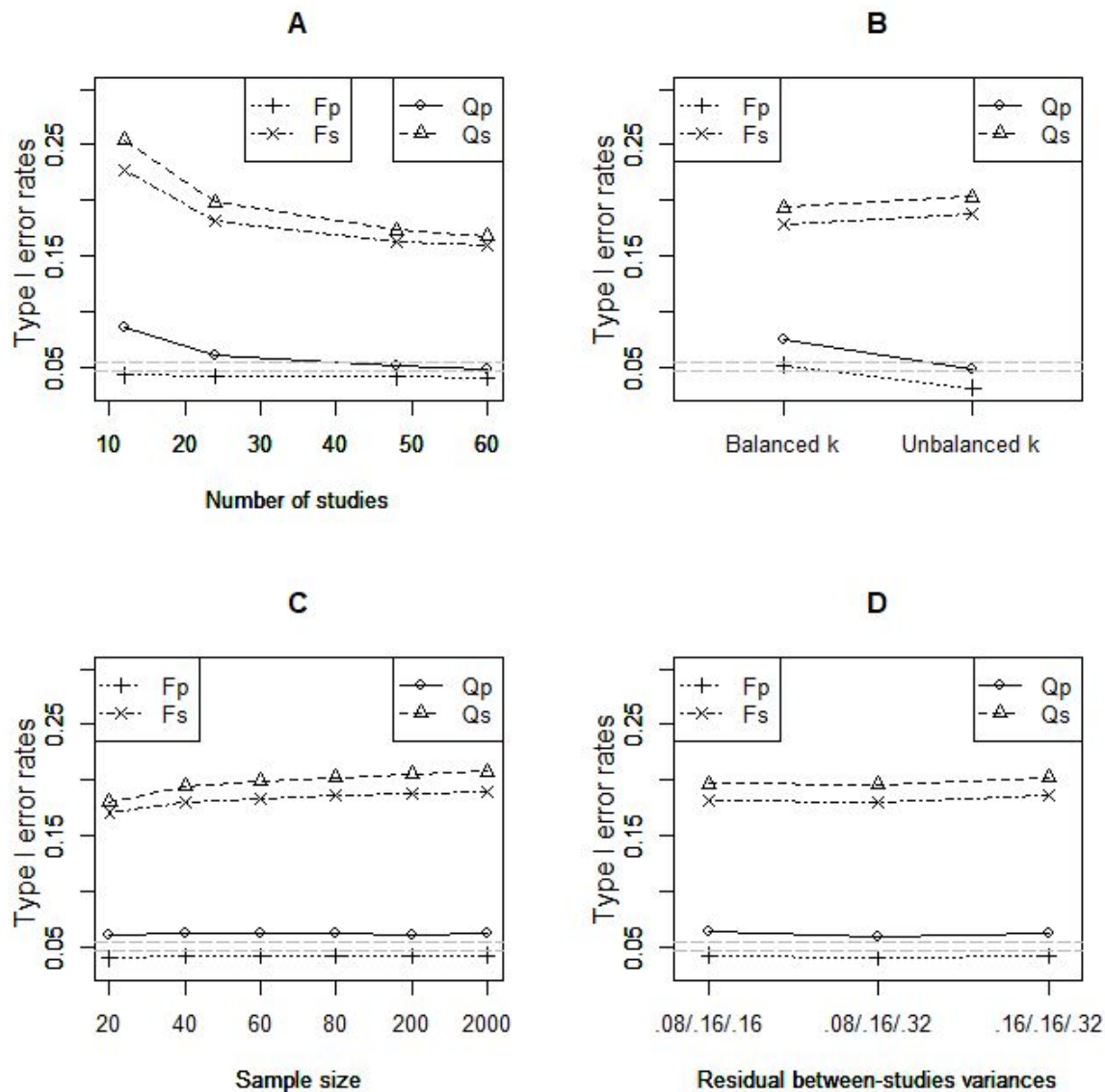
Supplementary figure 14. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the two categories of the moderator and smaller variance in the smaller category using the restricted maximum likelihood estimator.



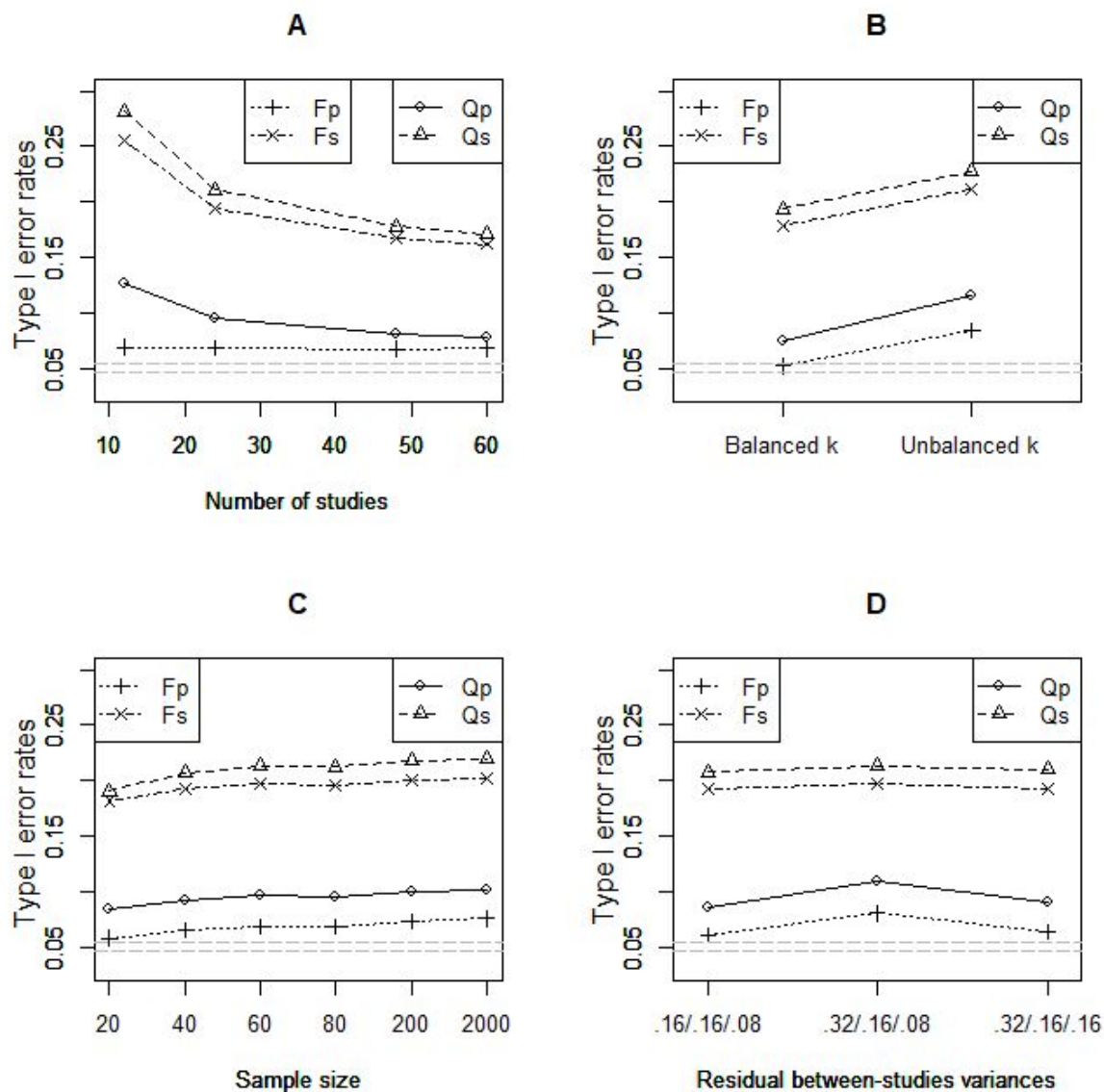
Supplementary figure 15. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the two categories of the moderator and larger variance in the smaller category using the restricted maximum likelihood estimator.



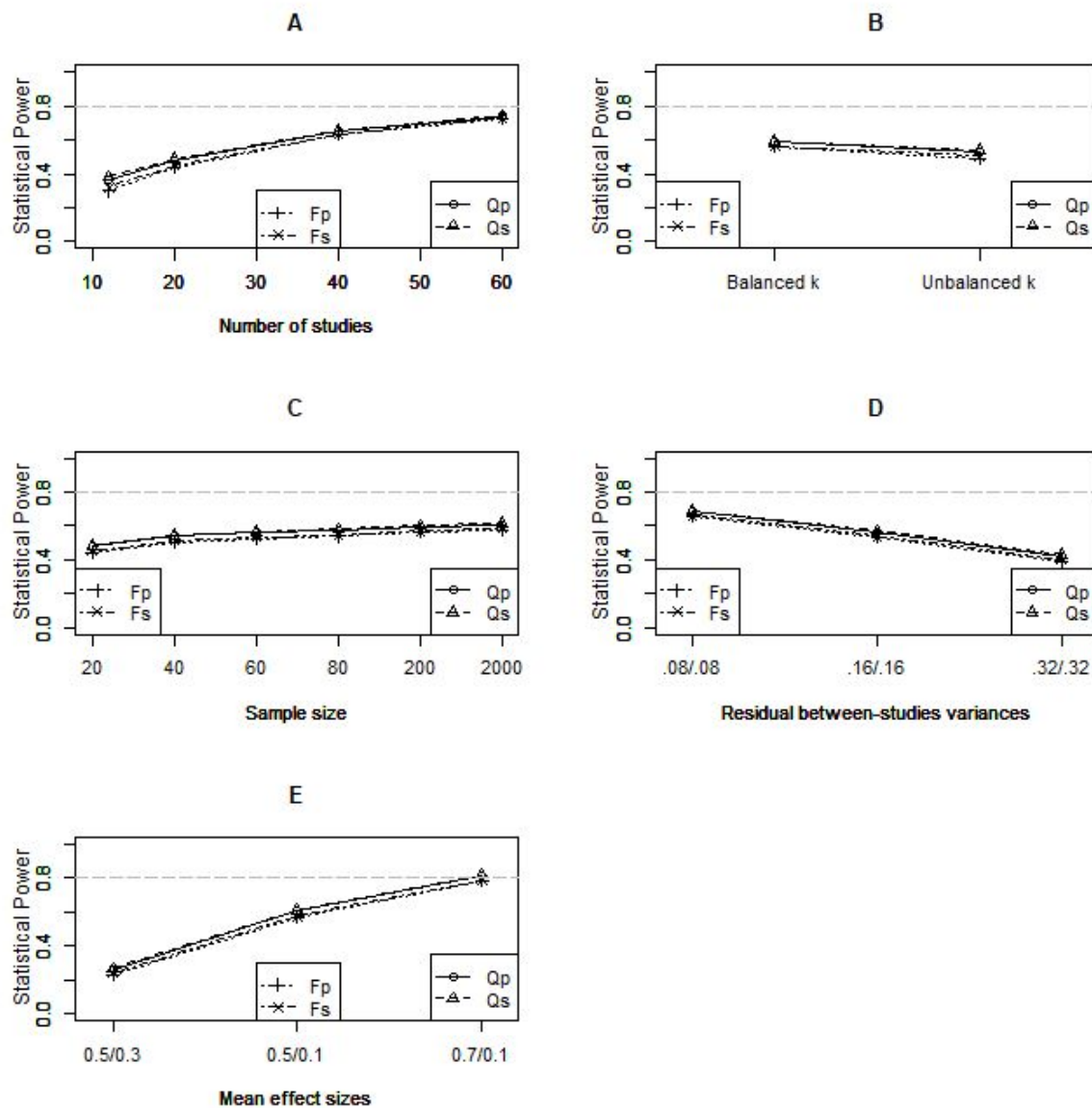
Supplementary figure 16. Average Type I error rates in scenarios with homoscedastic residual between-studies variances across the three categories of the moderator using the restricted maximum likelihood estimator.



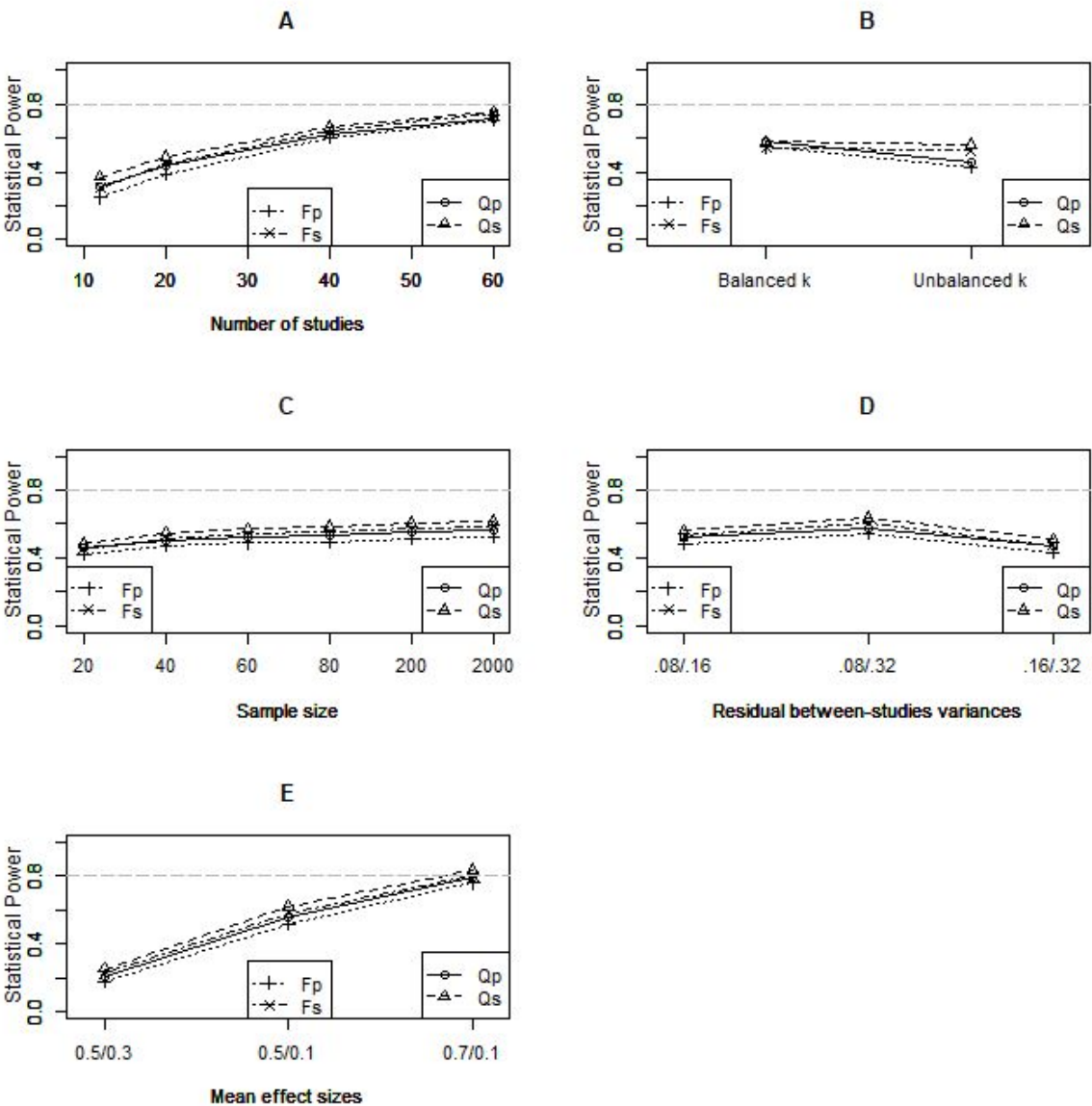
Supplementary figure 17. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the three categories of the moderator and smaller variance in the smaller category using the restricted maximum likelihood estimator.



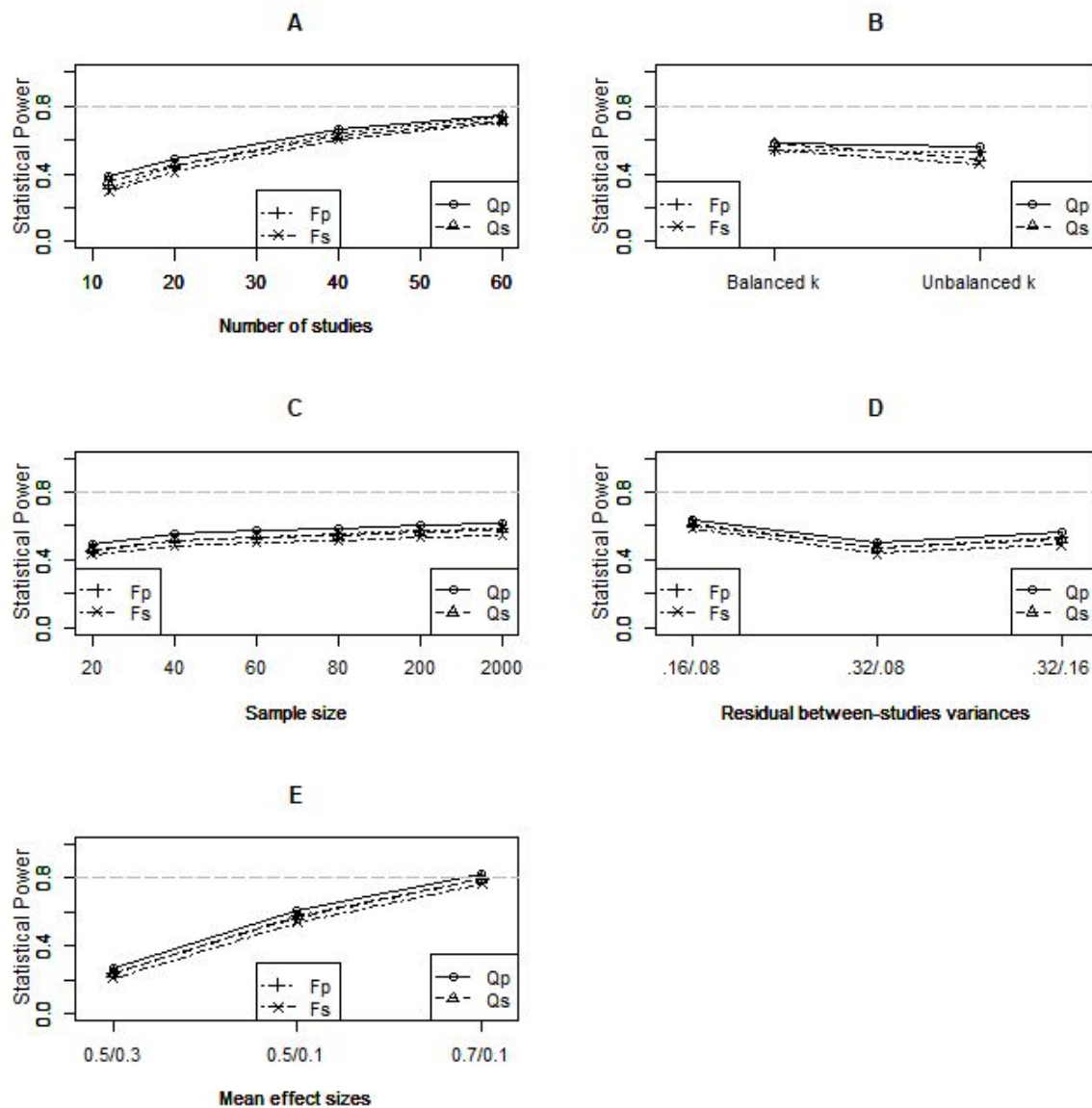
Supplementary figure 18. Average Type I error rates in scenarios with heteroscedastic residual between-studies variances across the three categories of the moderator and larger variance in the smaller category using the restricted maximum likelihood estimator.



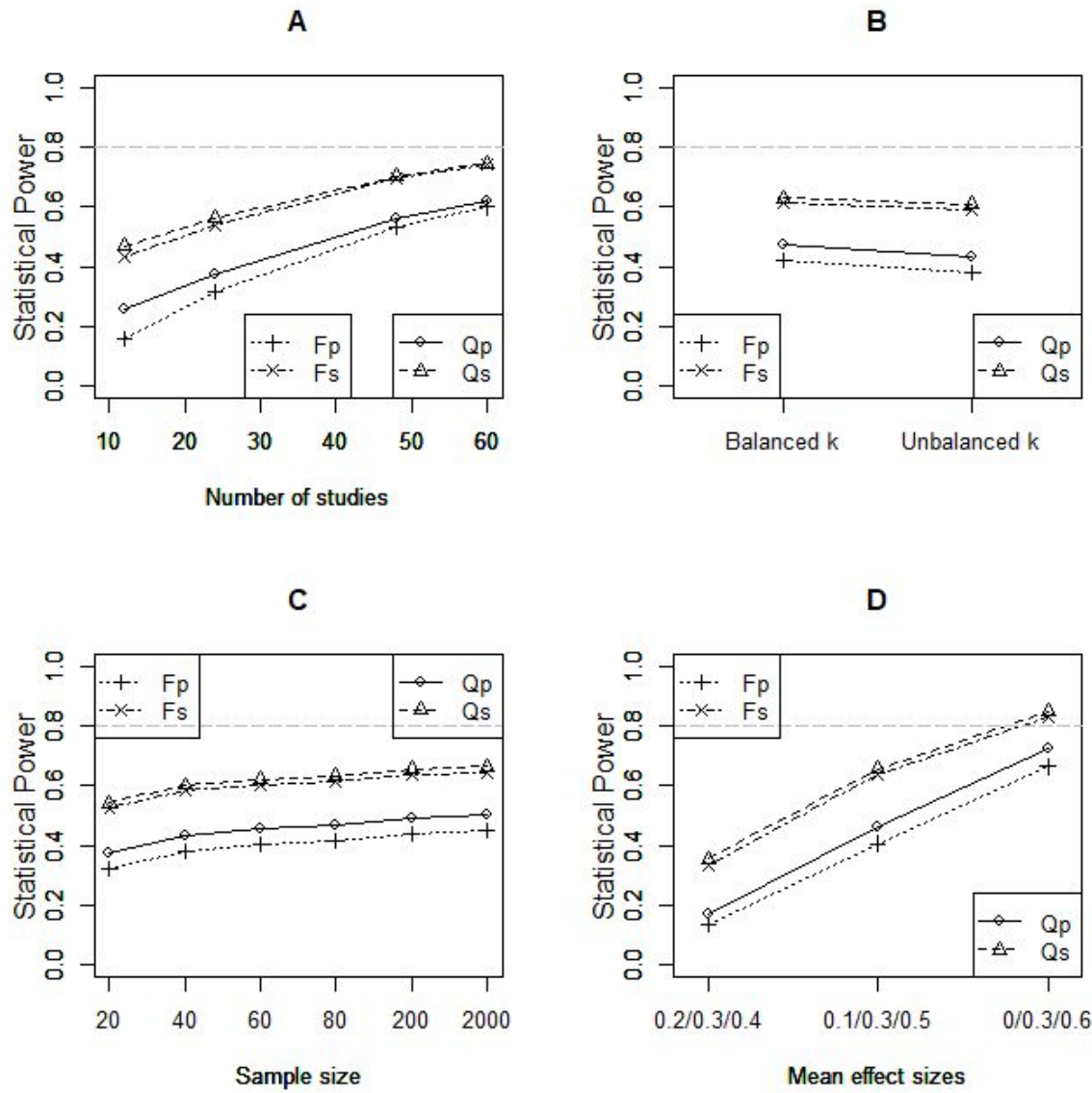
Supplementary figure 19. Average power rates in scenarios with homoscedastic residual between-studies variances across the two categories of the moderator using the restricted maximum likelihood estimator.



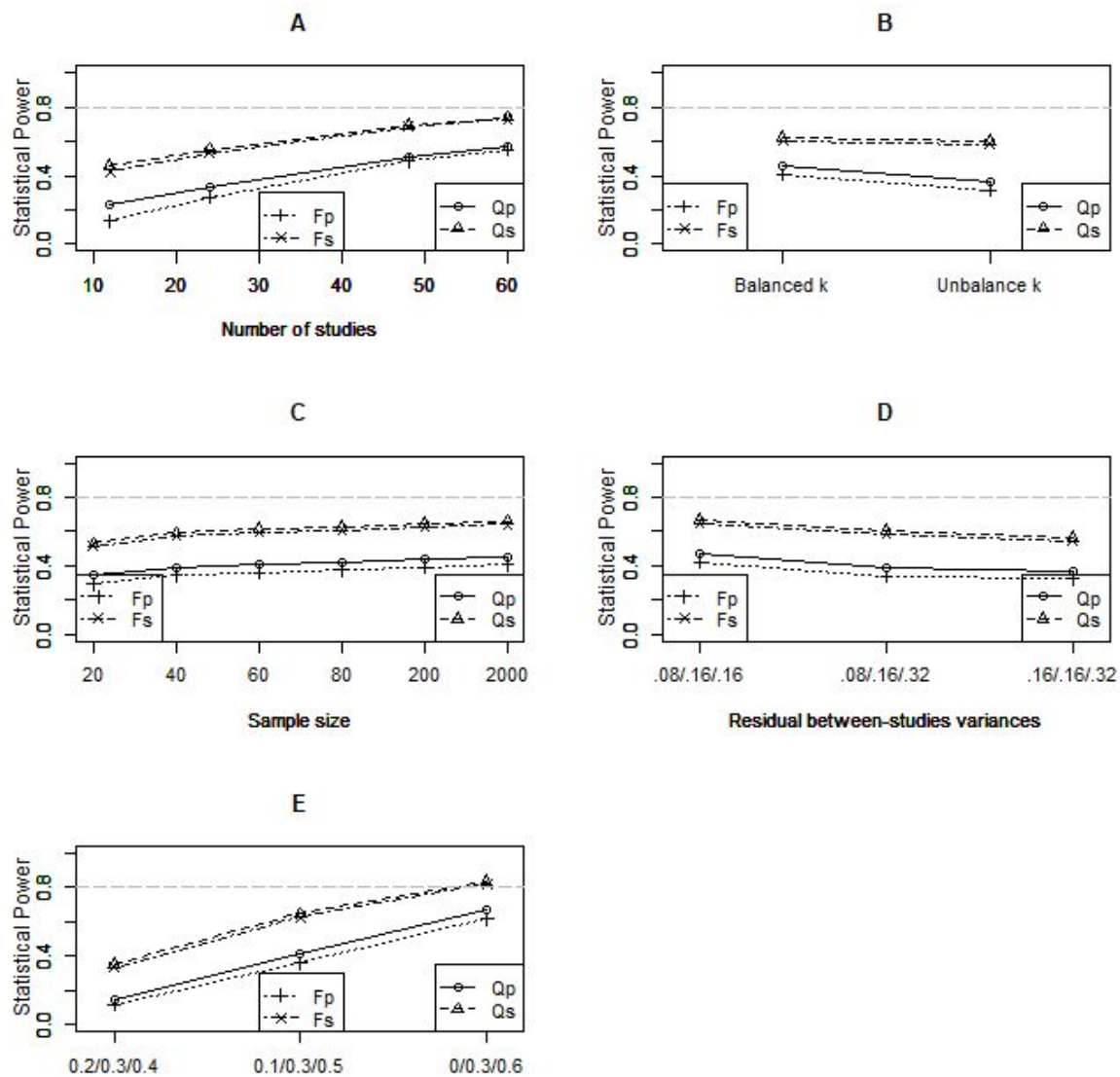
Supplementary figure 20. Average power rates in scenarios with heteroscedastic residual between-studies variances across the two categories of the moderator and smaller variance in the smaller category using the restricted maximum likelihood estimator.



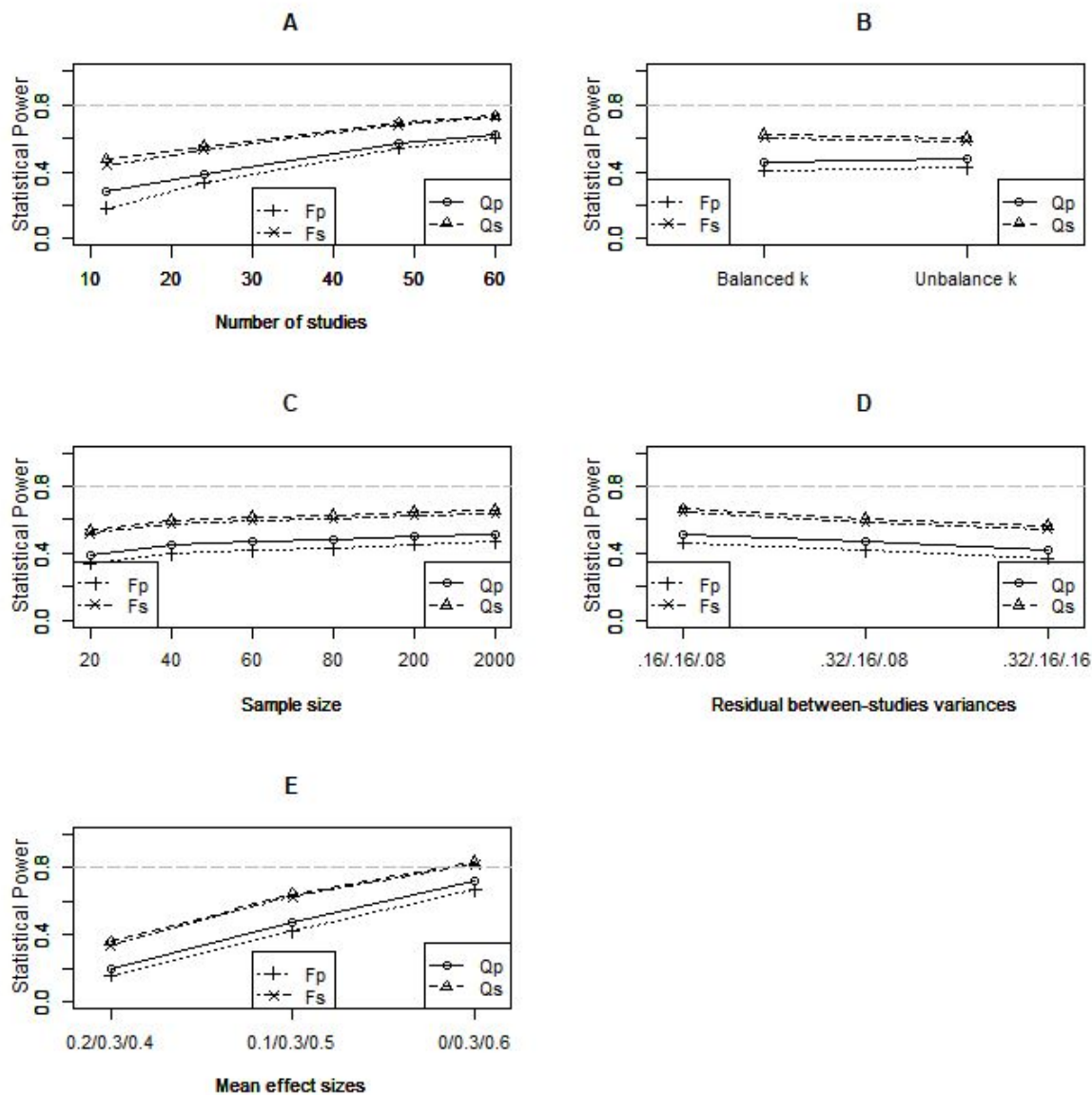
Supplementary figure 21. Average power rates in scenarios with heteroscedastic residual between-studies variances across the two categories of the moderator and larger variance in the smaller category using the restricted maximum likelihood estimator.



Supplementary figure 22. Average power rates in scenarios with homoscedastic residual between-studies variances across the three categories of the moderator using the restricted maximum likelihood estimator.



Supplementary figure 23. Average power rates in scenarios with heteroscedastic residual between-studies variances across the three categories of the moderator and smaller variance in the smaller category using the restricted maximum likelihood estimator.



Supplementary figure 24. Average power rates in scenarios with heteroscedastic residual between-studies variances across the three categories of the moderator and larger variance in the smaller category using the restricted maximum likelihood estimator.