A Welch-type test for comparing K linear contrasts under heteroscedasticity with applications to meta-analysis

E. Kulinskaya¹, E. Knight², and H.Gao¹

RESEARCH REPORT 2002-1
HRDSU,
UNIVERSITY OF HERTFORDSHIRE
AUGUST 2002

Summary

A test for comparing I linear contrasts with heteroscedasticity, both across components of each contrast and between the contrasts, is developed under assumption of normality. The test is based on the weighted sums of squares. This is an extension of methods for weighted one-way ANOVA developed in Welch (1951) [21] under the null, and in Kulinskaya et al (2000) [12] under the alternatives. We provide very accurate approximations to the null distribution and to the distribution under alternatives. The quality of these approximations is studied by simulation. The main application is the I by 2 layout which is widespread in meta-analysis. Our methods allow the homogeneity of effect sizes across I studies to be tested, without the assumption of equal variances in the treatment and the control groups.

¹ Elena Kulinskaya is a Reader, and Haiyan Gao is a Research Assistant at the University of Hertfordshire, Hatfield AL10 9AB, UK.

²Emily Knight is a statistician at the Department of Paediatrics, University of Cambridge, Box 116, Level 8, Addenbrooke's Hospital Hills Road, Cambridge. CB2 2QQ. UK

³ The authors thank the Department of Health, Directorate of Health and Social Care, Midlands and East of England, for support of this research, contract number RCC33039

1 INTRODUCTION AND SUMMARY

In this paper we intend, under the assumption of normality, to develop a test for comparing several independent linear contrasts under heteroscedasticity, both across components of each contrast and between the contrasts. We deal with heteroscedasticity by using tests based on the weighted sums of squares. Methods for the case of one-way ANOVA were developed in [21] under the null, and in [12] under the alternatives.

The main application is an I by 2 layout, though developed here theory applies to a general I by K layout. Consider a study where two independent samples (for example observations in treatment and control groups) are stratified into I levels of an external factor. This results in I pairs of samples of n_{1i} observations in the control and n_{2i} observations in the treatment group. These observations are from $2I$ different normal populations whose true means are m_{1i} and m_{2i} , and true variances are σ_{1i} and σ_{2i} in the control and treatment groups respectively, $i = 1, \dots, I$. Parameters of interest are the treatment effects $\mu_i = m_{2i} - m_{1i}$ at each level. These are estimated by the differences in sample means for two independent samples. The variance of $\hat{\mu}_i$ may be estimated by the sum of squared standard errors: $\hat{v}_i = s_{1i}^2/n_{1i} + s_{2i}^2/n_{2i}$. Its distribution is a linear combination of chi-square distributions. We need to test the null hypothesis that the population treatment effects are equal without imposing the condition that their variances are necessarily equal. The problem differs from Welch (1951) in that the variance estimates do not have the chi-square distribution. Welch's algebra is modified to include this case. We also provide an approximation to the distribution of the test statistic under alternatives. Extensive simulations show that the theoretical derivations work. A summary of the simulation results is included.

1.1 Notation and preliminaries

We assume there are I independent samples of size n_i , $i = 1, \dots, I$ which are to be compared in regards to a particular parameter denoted by μ_i . For simplicity, this parameter can be thought of as a population mean. In an application to meta analysis considered later, this parameter is a 'raw' effect size, e.g. a difference of two means. To perform the comparison, an estimate of the parameter of interest is obtained from each sample. These estimates

$\hat{\mu}_i, i = 1, \dots, I$ are assumed to be normally distributed with mean μ_i and variance $v_i = \sigma_i^2/n_i$.

Denote the overall sample size and mean by $N = \sum_{i=1}^I n_i$ and $\hat{\mu} = (\sum_{i=1}^I n_i \hat{\mu}_i)/N$. When the group variances are equal $\sigma_1 = \sigma_2 = \dots = \sigma_I$, the joint variance can be estimated by pooled sample variance and the classical F -test is appropriate for testing the null hypothesis of the equality of parameters μ_i .

However, when the variances are unequal it is well known that this test can be misleading, see e.g. Miller (1997)[15], pp. 89–91.

Approximations to the true distribution of the unweighted F -statistic or other statistics based on its numerator q can be found in Box (1954)[1], Brown and Forsythe (1974)[2], Mehrotra (1997)[14]. For further discussion see [11]. Earlier and more involved approximations were obtained in Tan (1982) [18]. We prefer to use the weighted ANOVA approach introduced in Welch (1951)[21] and James (1951) [7] because it accounts for the accuracy with which group means (or other effect sizes) can be estimated.

This approach to overcoming the problem of different variances introduces weights w_i , proportional to inverse variances of $\hat{\mu}_i$, and defines $\hat{\mu}_w$ to be the weighted estimate of the centre of the entire population:

$$\hat{\mu}_w = \sum_{i=1}^I w_i \hat{\mu}_i / \sum_{i=1}^I w_i . \quad (1)$$

Then the variability from group to group is explained by the weighted sum of squares:

$$q_w \equiv \sum_{i=1}^I w_i (\hat{\mu}_i - \hat{\mu}_w)^2 . \quad (2)$$

The distribution of the statistic $q_{\hat{w}}$, obtained by substitution of estimated weights $\hat{w}_i$ for w_i , is of primary interest in this paper. $q_{\hat{w}}$ estimates the population non-centrality parameter $\lambda = \sum_{i=1}^I w_i (\mu_i - \bar{\mu}_w)^2$, where $\bar{\mu}_w$ is the population counterpart of $\hat{\mu}_w$ (1). Since $\lambda = O(N)$, an appropriate measure of effect size would be λ/N . Another measure of effect size, the population weighted coefficient of determination ρ and its estimate $\hat{\rho}$ based on $q_{\hat{w}}$ was introduced in Kulinskaya, Staudte and Berezner (2000) [12]. For inverse variance weights $\rho = 1/(1 + N/\lambda)$. We use ρ to conveniently measure the distance from the null hypothesis.

Welch (1951)[21] and James (1951) [7] studied the weighted sum of squares given in (2) with weights $w_i \equiv n_i/\sigma_i^2$. They assumed that the group variances σ_i^2 can be estimated by sample variances $s_i^2, i = 1, \dots, I$, and that these estimates are independent of $\hat{\mu}_i$ and are chi-square distributed. The weights w_i were estimated by $\hat{w}_i = n_i/s_i^2$. For very large sample sizes the null distribution of this statistic can be approximated by the chi-square distribution with $I - 1$ d.f. James (1951)[7] and Welch (1951)[21] found an improved approximation to the null distribution of this statistic. The approaches of Welch and James are asymptotically equivalent but based on different approximations. An approximation to the non-null distribution of this statistic was found in Kulinskaya, Staudte and Berezner (2000)[12]. In some further work James generalised his test to a general linear model [8], and a much simpler derivation of an approximate null distribution based on the first two moments is given by Johansen (1980) [10]. James [9] also derived a confidence interval for the weighted overall mean $\hat{\mu}_w$ in the case of different within-sample variances but equal means. Tan and Tabatai (1986) [20] and Tan (1987) [19] used James-type procedures to test for the parallelism of several straight lines and in the ANCOVA model with one factor, and one covariate, respectively. Modifications to include robust estimates of means and variances were introduced and studied by simulation in numerous publications by Wilcox and co-authors, see [23] and [11].

1.2 Application to meta-analysis

One of the wide-spread problems of meta-analysis is testing for heterogeneity of effect sizes. Given I studies of sizes n_i with respective observed effect sizes $\hat{\theta}_i, i = 1, \dots, I$, the null hypothesis $H_0 : \theta_i = \theta, i = 1, \dots, I$ for some unknown θ is tested against the omnibus alternative that H_0 is false. θ is used in this subsection to denote a general effect size, not necessarily based on means, instead of μ in the previous discussion.

Two approaches to modelling this problem are fixed effects and random effects models, with the majority of publications dealing with the former, see Hedges and Olkin (1985) [6], Whitehead and Whitehead (1991) [22], and Fleiss (1993) [5], among others. We will concentrate on the fixed effects approach from now on. It is assumed that asymptotically $\hat{\theta}_i$ is normally distributed $N(\theta_i, w_i^{-1})$. Then a large sample test of homogeneity is based on the fact that the weighted sum of squares $Q = \sum w_i(\hat{\theta}_i - \hat{\theta})^2$, where

$\hat{\theta} = \sum w_i \hat{\theta}_i / \sum w_i$, has asymptotically a Chi-square distribution with $I - 1$ d.f. This test was introduced by Cochran(1954) [3]. Note that this is exactly the same q_w (2) statistic as was investigated by James (1951)[7] and Welch (1951) [21]. The large sample approach is illustrated in Fleiss (1993) [5] for Cohen's (1988) d [4], the relative risk and odds-ratio, as well as the Mantel-Haenszel estimate of the odds ratio, and Fisher's Z-transformed product-moment correlation. See Whitehead and Whitehead (1991) [22] for examples of this approach with the log-odds ratio, and with the log-hazard ratio in survival analysis. A recent paper by Hedges and Pigott (2001) [13] advocates using noncentral $\chi^2_{I-1}(\lambda)$ distribution for calculations of power and sample sizes.

Methods of Welch [21] and James [7] are not directly applicable to meta-analysis because some of their assumptions may be violated. Very often, estimates of mean and variance for an effect size are not independent. This is true for all so called standardised effect sizes, where the parameter of interest is divided by its standard deviation or some other measure of variation. Another underlying assumption, that of normality, is only true asymptotically, if at all. Still, there exists a large area of meta-analysis for which the Welch test is both very relevant and applicable without much technical modification. This is when the effect of interest is a raw non-standardised parameter which is estimated from a general linear model. Even without assuming normality of the underlying distribution of the raw data, linear model estimates such as sample means achieve normality rather quickly, and their variances can be estimated independently from sample variances. On the other hand the modest sample sizes in some studies make the James/Welch correction non-negligible.

As an example of this problem in meta-analysis we examine the data from Mumford et al (1984) [16]. The problem under consideration is the effect of psychological intervention on reducing hospital length of stay (in days). Data is available from 8 studies ranging from 8 to 50 patients in each group. The data is shown in Table 1.

	Psychotherapy			Control			Effect size	Weights	
Study	n_1	$\bar{X}_1$	sd_1	n_2	$\bar{X}_2$	sd_2	$\bar{X}_1 - \bar{X}_2$	pooled	unpooled
1	13	5.00	4.70	13	6.50	3.80	-1.5	0.356	0.356
2	30	4.90	1.71	50	6.10	2.30	-1.2	4.251	4.920
3	35	22.50	3.44	35	24.90	10.65	-2.4	0.279	0.279
4	20	12.50	1.47	20	12.30	1.66	0.2	4.068	4.068
5	10	3.37	0.92	10	3.19	0.79	0.18	6.800	6.800
6	13	4.90	1.10	14	5.50	0.90	-0.6	6.727	6.625
7	9	10.56	1.13	9	12.78	2.05	-2.22	1.643	1.643
8	8	6.50	0.76	8	7.38	1.41	-0.88	3.118	3.118

Table 1: Mumford data

Note the modest sample sizes, which make the use of asymptotic results dubious, and also quite different standard deviations between treatment and control groups in studies 3, 7 and 8. However, the sample sizes in treatment and control groups are equal except for studies 2 and 6. In fact the weights based on the pooled variances and on the separate estimates of variance coincide for 6 studies out of 8. An asymptotic analysis of this (famous) data set based on the standard chi-square approximation is provided in Shadish and Haddock (1994), p.273 [17].

In this particular study we are interested in days of hospitalisation. The weighted estimate of combined raw effect magnitude (based on differences of means weighted by inverse pooled variances) is $T = -.54$ days, with a standard error of .19, with a confidence interval of $(-.91, -.16)$, and $p = .0048$, i.e. the effect is significantly different from zero. But the homogeneity test $Q = 13.92$ with 7 df results in $p = .053$, suggesting that the mean differences might differ across studies. This analysis is an example of the juxtaposition of two erroneous techniques: pooling the unequal variances within the studies, and using the chi-square approximation in weighted ANOVA which is valid only for large sample sizes. We shall return to this example in Section 3.1.

2 Approximations to the distribution of the weighted sum of squares

In this section we assume the substitution of $\hat{w}_i$ for w_i has been made in q_w and denote the resulting statistic by $q_{\hat{w}}$. We also assume that the $\hat{w}_i$ are asymptotically independent from the $\hat{\mu}_i$, which are, in their turn, asymptotically normal $N(\mu_i, w_i^{-1})$.

We assume that our weight estimates $\hat{w}_i$ are asymptotically unbiased and $\sqrt{n_i}$ -consistent estimates of the weights w_i . We do not make any assumptions about the distribution of the $\hat{w}_i$ apart from the existence of their central moments M_j , and their order $M_1 = O(1), M_j = O(n), j \geq 2$.

It is shown in the Appendix that the first 3 cumulants of $q_{\hat{w}}$ are approximated by

$$\begin{aligned}\kappa_1 &= I - 1 + \lambda + A_0 + B_0, \\ \kappa_2 &= 2(I - 1 + 2\lambda + 2(A_0 + 2B_0) + 3P_{220}/2 + 3P_{212} + P_{204}/2), \\ \kappa_3 &= 8(I - 1 + 3\lambda + 3(A_0 + 3B_0) + 9P_{220}/2 + 27P_{212}/2 + 3P_{204} \\ &\quad + P_{306}/8).\end{aligned}\tag{3}$$

Coefficients A_0, B_0 and P_{ijs} are given in (11) in the Appendix for a general case of any random weights. They are calculated from the moments of sample weights $\hat{w}_i$ and the distances of row means μ_i from the grand mean. These coefficients have to be derived for each particular distribution of sample weights. For the case of inverse sample variance weights the above moments were obtained by Welch (1951) under H_0 , and by Kulinskaya et al (2001) [12] under the alternatives. The case of inverse variance weights with a sum of chi-squares distribution is explored in the next section. The resulting coefficients are given by (8).

Under H_0 , only the terms $I - 1$, and $A_0 = P_{220}$ are non-zero in the above equations. We can use the approach of Welch (1951) to obtain an approximation to the null distribution, which is valid to the order of N^{-1} . More specifically, we propose to approximate the null distribution of the $q_{\hat{w}}$ by the $cF_{I-1, \hat{f}_2}$ distribution with

$$c = \kappa_1 \frac{(\hat{f}_2 - 2)}{\hat{f}_2}, \quad \hat{f}_2 = \frac{2(I - 3)\kappa_1^2 + 4(I - 1)\kappa_2}{(I - 1)\kappa_2 - 2\kappa_1^2}.\tag{4}$$

This approximation is obtained exactly as in section 4 of the Welch (1951) paper, by equating the first two moments of the relevant F distribution to κ_1 and κ_2 .

To approximate the distribution of $q_{\hat{w}}$ under alternatives we tried shifted F distribution, but it does not work. We propose instead the use of the $c\chi_f^2 + d$ distribution. The constants c, d , and f are to be chosen so that the first three cumulants of $c\chi_f^2 + d$ agree with cumulants $\kappa_1 - \kappa_3$ for q_w given in Equation (3). The appropriate values of c, d and f are

$$\begin{aligned} c &= \kappa_3/4\kappa_2 \\ d &= \kappa_1 - 2\kappa_2^2/\kappa_3, \\ f &= 8\kappa_2^3/\kappa_3^2. \end{aligned} \tag{5}$$

This approximation can also be used for the null distribution of $q_{\hat{w}}$. The quality of this approximation for the null distribution is compared to that of the above F -approximation and is found to be inferior.

3 Calculation of central moments for an inverse variance estimator having a sum of chi-squares distribution

Throughout this section we consider just one-way layout with K independent samples from the i th row (or study) out of I . Its subscript i is skipped in this section for convenience. Its overall mean μ is estimated by $\hat{\mu}_n$. The variance $\text{Var}(\hat{\mu}_n)$ is denoted by v_n . We assume that its variance estimator $\hat{v}_n$ is distributed as $\sum c_k \chi_{f_k}^2$. Our main application is when $\hat{v}_n$ is an estimator of variance for a contrast $T_n = \sum a_k \bar{x}_k$. We assume that the variances σ_k^2 differ between K samples of sizes n_k , with $n = \sum n_k$, and $\hat{v}_n = \sum a_k^2 s_k^2/n_k$ estimates the variance of T_n which is $v_n = \sum a_k^2 \sigma_k^2/n_k$. We have $\hat{v}_n = \sum a_k^2 (f_k s_k^2/\sigma_k^2)(\sigma_k^2/f_k n_k) = \sum a_k^2 (\sigma_k^2/f_k n_k) \chi_{f_k}^2$, i.e. here $c_k = a_k^2 \sigma_k^2/f_k n_k$.

Define a weight $w_n = v_n^{-1}$ and its sample estimate $\hat{w}_n = \hat{v}_n^{-1}$. We can obtain the moments of $\hat{w}_n$ using the delta method. We use the Taylor series for x^{-t} :

$$x^{-t} = x_0^{-t} + \sum_{j=1}^{\infty} (-1)^j \binom{t+j-1}{j} (x - x_0)^j x_0^{-t-j}.$$

Substituting $\hat{v}_n$ for x , and v_n for x_0 , and taking expectation at both sides, we obtain

$$E[(\frac{\hat{v}_n}{v_n})^{-t}] = E(\hat{w}_n/w_n)^t = 1 + \sum_{j=1}^{\infty} (-1)^j \binom{t+j-1}{j} E[(\frac{\hat{v}_n - v_n}{v_n})^j]. \quad (6)$$

For $\hat{v}_n = \sum c_k \chi_{f_k}^2$ we have

$$\begin{aligned} E[\hat{v}_n] &= v_n = \sum c_k f_k \\ E[\hat{v}_n - v_n]^2 &= m_2 = 2 \sum c_k^2 f_k, \\ E[\hat{v}_n - v_n]^3 &= m_3 = 8 \sum c_k^3 f_k, \\ E[\hat{v}_n - v_n]^4 &= m_4 + 3m_2^2 = 48 \sum c_k^4 f_k + 3m_2^2, \end{aligned}$$

where $m_j, j = 2, 3, 4$ are the cumulants of $\hat{v}_n$. For a contrast T_n these moments are:

$$v_n = \sum a_k^2 \sigma_k^2 / n_k; \quad m_2 = 2 \sum a_k^4 \sigma_k^4 / f_k n_k^2, \quad m_3 = 8 \sum a_k^6 \sigma_k^6 / f_k^2 n_k^3. \quad (7)$$

Note that $m_j = O(n^{1-2j})$, $v_n = O(1/n)$, and $m_j/v_n^j = O(n^{1-j})$. It follows that $E[(\frac{\hat{v}_n - v_n}{v_n})^j] = O(n^{-1})$ for $j = 2$ and $O(n^{-2})$ for $j = 3$ and $j = 4$. Next we calculate the moments $E(\hat{w}_n/w_n)^t$ for $t = 1, 2, 3$ from (6) as

$$E(\hat{w}_n/w_n) = 1 + m_2/v_n^2 - m_3/v_n^3 + (m_4 + 3m_2^2)/v_n^4 + O(1/n^3).$$

Similarly,

$$E(\hat{w}_n/w_n)^2 = 1 + 3m_2/v_n^2 - 4m_3/v_n^3 + 5(m_4 + 3m_2^2)/v_n^4 + O(1/n^3),$$

and

$$E(\hat{w}_n/w_n)^3 = 1 + 6m_2/v_n^2 - 10m_3/v_n^3 + 15(m_4 + 3m_2^2)/v_n^4 + O(1/n^3).$$

Finally, we obtain the central moments as

$$M_j = E(\hat{w}_n - w_n)^j = w_n^j E[(\hat{w}_n/w_n) - 1]^j.$$

In particular

$$M_1 = E(\hat{w}_n - w_n) = w_n[m_2/v_n^2 + O(1/n^2)] = m_2 w_n^3 + O(1/n),$$

$$M_2 = E(\hat{w}_n - w_n)^2 = w_n^2[m_2/v_n^2 + O(1/n^2)] = m_2 w_n^4 + O(1),$$

$$M_3 = E(\hat{w}_n - w_n)^3 = w_n^3[9m_2^2/v_n^4 - m_3/v_n^3 + o(1/n^2)] = 9m_2^2w_n^7 - m_3w_n^6 + O(1).$$

Now we can calculate coefficients (11) introduced in the Appendix as follows:

$$\begin{aligned} P_{1js} &= \sum m_{2t}w_t^{3-j}(1 - w_t/\sum w_k)^j(\mu_t - \bar{\mu})^s, \\ P_{2js} &= \sum m_{2t}w_t^{4-j}(1 - w_t/\sum w_k)^j(\mu_t - \bar{\mu})^s, \\ P_{3js} &= \sum (6m_{2t}^2w_t - m_{3t})w_t^{6-j}(1 - w_t/\sum w_k)^j(\mu_t - \bar{\mu})^s, \end{aligned} \quad (8)$$

and $A_0 = P_{220}$, $B_0 = P_{212}$.

3.1 Analysis of data by Mumford et al.(1984)

Now we are able to reanalyse the data by Mumford et al. (1984) which was introduced in Section 1.2, and was reproduced in Table 1. First we have recalculated the value of $Q = 14.204$ using weights based on unequal variances (given in the last column of Table 1). Using the standard chi-square approximation now results in $p = 0.048$, i.e. the studies are significantly different at the 0.05 level. We have recalculated the p-value using the techniques from the previous section. The only values we need to calculate the moments (3) under the null hypothesis are: $I - 1 = 7$ and $A_0 = P_{220} = 0.593$. The null distribution is approximated by cF_{7,f_2} distribution with $c = 7.409$ and $f_2 = 82.47$, see (4). Finally we obtain $p=0.077$, so the mean differences are not significantly different at the 0.05 level.

Now we can use the distribution under alternatives to calculate the power of the above test at 5% significance level (for observed sample means and variances). The 5% critical value for cF_{7,f_2} distribution with the parameters c and f_2 given above is 15.73. The extra values needed to calculate the moments (3) under alternatives are $B_0 = P_{212} = 1.283$, $P_{204} = 5.337$, $P_{306} = 11.855$. The parameter λ in (3) is estimated by $Q = 14.204$. The values c, d and f needed for the chi-square approximation (5) are $c = 2.044$, $d = -0.95987$, and $f = 11.762$. The power is calculated as 1 minus the value of the approximating $c\chi_f^2 + d$ cumulative distribution function at 15.73, and it equals 76%. This is not too low a value of power given the widespread requirement of 80% power in medical applications. This confirms our conclusion of no differences across studies.

Differences in results between us and previous analyses of Mumford data may not seem to be considerable (p-value changed from .053 to .077). And there were two reasons for this difference: usage of different vs pooled variances and the improved Welch approximation. We decided to further explore this data to see both effects separately, and to study the quality of Welch approximation for realistic sample sizes. The Mumford data is too balanced to show much difference between methods based on pooled or different variances. That is why we decided to distort the variances, sample sizes and means in following ways. Case 0 corresponds to original Mumford data. The standard deviation for the 5th study was changed to .46 in the following 3 simulations. Case 1: sample sizes are 70 and 10 for the fifth study; Case 2: sample sizes are 10 and 70 for the fifth study; Case 3: sample sizes are 60 and 10 for the 3rd study, and 70 and 10 for the fifth study. For the distribution under alternatives we changed all raw effect sizes μ_i to -.55 (average effect size) apart from study 5 where we put $\mu_5 = .6$. We have studied by simulation quality of 6 approximations for the null case: standard chi-square, our chi-square approximation based on the moments (3), and the Welch F-approximation, each in two variations: with different and pooled variances. For the alternatives we studied just the first four approximations (no F-approximation there). Standard chi-square means noncentral $\chi^2(\lambda)$ distribution. Results for 95% nominal level are summarised in the table 2 below. Full methodology of simulations is described in the next section.

Mumford data and its variations	Standard χ^2		χ^2 with moments (3)		F-distribution	
	different variances	pooled variances	different variances	pooled variances	different variances	pooled variances
null distribution						
case 0	0.914	0.919	0.945	0.944	0.944	0.944
case 1	0.915	0.890	0.947	0.921	0.947	0.920
case 2	0.916	0.936	0.945	0.956	0.944	0.956
case 3	0.901	0.732	0.943	0.779	0.942	0.777
alternatives						
case 0	0.894	0.897	0.943	0.939		
case 1	0.889	0.866	0.945	0.901		
case 2	0.907	0.956	0.946	0.971		
case 3	0.882	0.802	0.943	0.846		

Table 2: Empirical $\hat{\alpha}^*$ levels at 95% nominal level for 3 approximations to the null distribu-

tion of $q_{\hat{w}}$, and for two approximations to its distribution under alternatives. Sample sizes and variances are variations of data by Mumford et al., 1984.

What is the quality of power calculations associated with the two approximations given in Table 2? For the approximation based on the moments (3) with unpooled weights the quality is very good, with absolute error being uniformly less than 0.001. The difference in various approximations is the most striking at the right tail, the values given below are for $\alpha = .995$. For the most unbalanced case 3 and using separate variances, the power of the approximation (3) is 23.3% with absolute error of 0.0005. Standard non-central chi-square approximation has theoretical power of 36.6% vs empirical power of 28.4%. With pooled weights theoretical power of the approximation (3) is 46.5% vs empirical level of 56.5%. Chi-square approximation gives 68.5 vs 60.2%.

This example shows the dangers of using pooled variances and the importance of improved approximations to distributions of statistics of interest in meta-analysis.

4 COMPARISON OF APPROXIMATIONS

In this section we first explain how we are going to explore and compare various approximations to the unknown distribution of $q_{\hat{w}}$. In Section 4.1 we summarize the combinations of sample sizes, means and variances which we used to simulate the data. In Section 4.2 we explore the quality of the Chi-square approximation to the non-null distribution of $q_{\hat{w}}$. Then we explain in Section 4.3 how a useful approximation can lead to a reliable confidence interval for the mean or median of $q_{\hat{w}}$ or for λ . In Section 4.4 we explore and compare different approximations to the null distribution. We also study the more realistic case of estimating the parameter A_0 . Finally in Section 4.5 we study how the confidence intervals introduced in Section 4.3 perform under the null.

4.1 Summary of parameters and sample sizes in simulations.

In the simulations we wanted to explore the effects of differences in means, differences in variances, and sample sizes on the quality of different approximations to the unknown distribution of $q_{\hat{w}}$. We have chosen 3 different sets of means (m_1 to m_3), to be at varying distances from the null; 3 sets of variances (v_1 to v_3), at varying distances from homogeneity; and 3 by 2 combinations of sample sizes (n_1 to n_9), varying from ‘small’ (under 100) to ‘large’ (up to 1500). The second dimension of the grid for sample sizes concerned differently balanced subsamples within rows, from equal sizes to very unbalanced. We performed 81 simulations for the non-null case, and 27+9 simulations for the null case. The extra 9 simulations for the null case were run with equal variances $\sigma_{i1}^2 = \sigma_{i2}^2 = 1$ for $i = 1, 2, 3$. This set of variances is denoted by v_4 in Tables 6 to 8. Simulations were run in the following combinations:

1. Means varied over the following 3 sets of values:

$$m_1 = \begin{vmatrix} 1 & 3 \\ 2 & 12 \\ 3 & 18 \end{vmatrix}, \quad m_2 = \begin{vmatrix} 1 & 3 \\ 2 & 6 \\ 3 & 9 \end{vmatrix}, \quad m_3 = \begin{vmatrix} 1 & 1.5 \\ 2 & 3 \\ 3 & 5 \end{vmatrix}.$$

This choice of means resulted in effect sizes for the test of homogeneity as given by population determination coefficient ρ^2 varying from ‘small’ (under 0.11) to ‘medium’ (0.27 to 0.45) to ‘large’ (above 0.79). For the null case means were kept constant at $\mu_i \equiv 1$.

2. Population variances σ_{ij}^2 varied over the following 3 (plus v_4 for the null distribution) sets of values:

$$v_1 = \begin{vmatrix} 0.5 & 1 \\ 0.5 & 1 \\ 0.5 & 1 \end{vmatrix}, \quad v_2 = \begin{vmatrix} 0.5 & 1 \\ 0.5 & 1 \\ 1 & 1 \end{vmatrix}, \quad v_3 = \begin{vmatrix} 0.5 & 1 \\ 1 & 1 \\ 1 & 1 \end{vmatrix}, \quad v_4 = \begin{vmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{vmatrix}.$$

3. Sample sizes varied over the following 3 by 3 grid:

$$\begin{aligned}
n_1 &= \begin{vmatrix} 25 & 25 \\ 38 & 38 \\ 50 & 50 \end{vmatrix}, \quad n_2 = \begin{vmatrix} 10 & 40 \\ 38 & 38 \\ 80 & 20 \end{vmatrix}, \quad n_3 = \begin{vmatrix} 40 & 10 \\ 60 & 15 \\ 80 & 20 \end{vmatrix}, \quad (\text{'small'}) \\
n_4 &= \begin{vmatrix} 50 & 50 \\ 100 & 100 \\ 150 & 150 \end{vmatrix}, \quad n_5 = \begin{vmatrix} 20 & 80 \\ 100 & 100 \\ 240 & 60 \end{vmatrix}, \quad n_6 = \begin{vmatrix} 80 & 20 \\ 160 & 40 \\ 240 & 60 \end{vmatrix}, \quad (\text{'medium'}) \\
n_7 &= \begin{vmatrix} 250 & 250 \\ 500 & 500 \\ 750 & 750 \end{vmatrix}, \quad n_8 = \begin{vmatrix} 100 & 400 \\ 500 & 500 \\ 1200 & 300 \end{vmatrix}, \quad n_9 = \begin{vmatrix} 400 & 100 \\ 800 & 200 \\ 1200 & 300 \end{vmatrix}. \quad (\text{'large'})
\end{aligned}$$

For each combination $\{n_i m_j v_k\}$ we generated $T = 50,000$ samples from the corresponding normal populations, and for each sample calculated $q_{\hat{w}}$. The weights for the t th value of $q_{\hat{w}}$ are based on the t th set of sample variances. Let X_t , $t = 1, \dots, T$ represent the T sample values of $q_{\hat{w}}$. The T values of $q_{\hat{w}}$ were ordered to obtain the order statistics $X_{(1)}, \dots, X_{(T)}$ and corresponding empirical quantiles $\hat{k}_\alpha \equiv X_{(\alpha T)}$.

4.2 Results for the distribution under alternatives

Our Chi-squared approximations are of the form $c\chi_f^2 + d$, where c , d and f are based on the known moments (3). Let F^* denote the distribution of $c\chi_f^2 + d$, and regard it as an approximation to F , the unknown distribution of $q_{\hat{w}}$. Let $k_\alpha^* = F^{*-1}(\alpha)$ denote the known quantile of our approximating distribution. We are interested in the empirical levels $\hat{\alpha}^* \equiv \#\{X_i \leq k_\alpha^*\}/T$ which are empirical estimates of the actual levels for the Chi-squared approximation $c\chi_f^2 + d$ being studied. We calculated the errors $\hat{\alpha}^* - \alpha$ and the relative errors $\epsilon(\hat{\alpha}^*) \equiv (\hat{\alpha}^* - \alpha)/\min\{\alpha, 1 - \alpha\}$, for varying values of α . These relative errors for $\alpha = 0.05, 0.95, 0.025$, and 0.975 are summarised in Table 3.

At the left tail relative errors $\epsilon(\hat{\alpha}^*)$ are usually positive, i.e. $\hat{\alpha}^*$ is larger than the nominal level α , though this is sometimes not true for the larger sample sizes. The relative errors increase with effect size, and decrease with total sample size. Unbalanced sample sizes have larger relative errors than balanced sample sizes (errors for n_1 less than for n_2 , and for n_2 they are less

than for n_3 , etc.). This relationship almost disappears at large sample sizes, where even the smallest sample size is big enough for the approximation to work. For 5% nominal level, maximum relative error varies from 125% at $\{n_3m_1v_2\}$ to 63% at $\{n_6m_1v_3\}$ to 17% at $\{n_9m_1v_2\}$. There is no obvious relationship with the variances.

At the right tail relative errors $\epsilon(\hat{\alpha}^*)$ are usually negative, i.e. $\hat{\alpha}^*$ is smaller than the nominal level α , though this is sometimes not true for the larger sample sizes. There is no evident relationship of relative errors with effect sizes, or variances. The errors are smaller than on the left tail, and decrease with total sample size. Unbalanced sample sizes still have larger relative errors than balanced sample sizes. For 95% nominal level maximum relative error varies from -31% at $\{n_3m_1v_1\}$ to -13% at $\{n_6m_1v_1\}$ to -5% at $\{n_9m_1v_3\}$.

The approximation evidently has somewhat thinner tails than the true distribution. What is its ability to estimate power? To answer this question directly, we compared our approximation to the empirical power at the null critical values k_α^* obtained using the F-approximation (4). The approximate power is $P^* = 1 - F^*(k_\alpha^*)$ for the Chi-squared approximation $c\chi_f^2 + d$ being studied. The empirical power is $\hat{P} = \#\{X_i \geq k_\alpha^*\}/T$ which is an empirical estimate of the actual power at k_α^* . The errors are $P^* - \hat{P}$. These values were calculated for each of our simulations at varying values of α . We discovered that our alternatives were too far from the null and resulted in a power of 1 except for simulations at ‘small’ effect sizes and small sample sizes. These results are summarised in Table 4. The quality of power estimates is very high for these relatively small sample sizes. This is explained by good overall quality of our approximation on the right tail. High quality of power estimation for alternatives close to the null can be deduced from the comparison of performance at the right tail for the chi-square approximation based on moments (3) and the F-approximation under the null given in Table 6.

n	m	v	ρ^2	$\epsilon(\widehat{0.05}^*)$	$\epsilon(\widehat{0.95}^*)$	$\epsilon(\widehat{0.025}^*)$	$\epsilon(\widehat{0.975}^*)$
n1	m1	v1	0.89	26.32	-8.68	44.88	-13.76
n1	m1	v2	0.88	21.60	-6.00	38.64	-6.32
n1	m1	v3	0.88	27.96	-6.68	48.96	-3.60
n1	m2	v1	0.45	4.32	-3.84	13.04	-12.00
n1	m2	v2	0.42	8.44	-3.44	10.32	-5.20
n1	m2	v3	0.42	4.88	-5.96	13.68	-8.48
n1	m3	v1	0.11	1.96	-1.00	2.32	-0.96
n1	m3	v2	0.10	-1.44	3.44	-0.08	3.12
n1	m3	v3	0.10	-0.52	0.52	3.04	1.60
n2	m1	v1	0.85	45.60	-13.24	76.64	-14.88
n2	m1	v2	0.85	42.56	-9.76	74.40	-9.76
n2	m1	v3	0.85	39.16	-6.28	71.60	-6.40
n2	m2	v1	0.35	15.32	-9.92	25.28	-18.16
n2	m2	v2	0.34	12.12	-9.76	22.24	-15.84
n2	m2	v3	0.34	10.76	-10.40	16.48	-16.80
n2	m3	v1	0.08	5.68	-4.52	9.76	-9.52
n2	m3	v2	0.07	3.88	-5.28	9.36	-12.00
n2	m3	v3	0.07	1.60	-3.40	5.44	-12.24
n3	m1	v1	0.82	119.48	-30.72	237.04	-39.44
n3	m1	v2	0.81	124.88	-24.68	252.48	-27.28
n3	m1	v3	0.81	123.24	-22.96	247.68	-29.68
n3	m2	v1	0.31	37.68	-23.28	67.36	-39.92
n3	m2	v2	0.30	35.00	-18.80	66.24	-27.44
n3	m2	v3	0.29	39.56	-18.76	72.16	-29.04
n3	m3	v1	0.06	0.80	-4.68	2.16	-12.40
n3	m3	v2	0.06	2.04	-5.56	0.88	-11.84
n3	m3	v3	0.06	3.36	-5.44	1.04	-140.48
n4	m1	v1	0.88	15.32	-4.12	26.64	-4.18
n4	m1	v2	0.87	15.36	-3.72	25.44	-2.96
n4	m1	v3	0.87	15.64	-2.08	29.36	-1.68
n4	m2	v1	0.43	1.84	-8.16	3.04	-10.32
n4	m2	v2	0.40	2.28	-3.36	6.32	-7.68
n4	m2	v3	0.39	5.96	-2.92	7.92	-6.48
n4	m3	v1	0.11	0.60	1.12	-0.88	-1.36
n4	m3	v2	0.10	0.48	-0.16	2.16	-0.40
n4	m3	v3	0.09	-3.00	-2.64	-1.44	-0.72
n5	m1	v1	0.83	21.52	-3.12	34.16	-1.28
n5	m1	v2	0.83	19.52	-2.44	33.28	-4.00
n5	m1	v3	0.83	28.24	-3.04	47.68	3.12

Table 3: Relative errors $\epsilon(\hat{\alpha}^*)$ for the empirical percentages $\hat{\alpha}^* = 0.05; 0.95; 0.025; 0.975$ of the chi-squared approximation to $q_{\hat{w}}$ based on cumulants (3). Note that $\epsilon(\hat{\alpha}^*)$ are shown as percentages. Continued next page.

n	m	v	ρ^2	$\epsilon(\widehat{0.05^*})$	$\epsilon(\widehat{0.95^*})$	$\epsilon(\widehat{0.025^*})$	$\epsilon(\widehat{0.975^*})$
n5	m2	v1	0.33	8.92	-5.80	13.20	-10.40
n5	m2	v2	0.32	4.44	-6.32	10.24	-14.56
n5	m2	v3	0.32	8.48	-5.84	14.32	-6.08
n5	m3	v1	0.08	3.72	0.04	5.20	0.08
n5	m3	v2	0.07	-1.08	0.40	0.40	-2.56
n5	m3	v3	0.07	3.56	-2.44	2.64	-1.12
n6	m1	v1	0.79	60.64	-12.52	116.16	-15.20
n6	m1	v2	0.79	57.68	-10.40	108.24	-8.08
n6	m1	v3	0.79	63.40	-9.40	119.28	-10.24
n6	m2	v1	0.28	22.12	-7.72	40.48	-14.96
n6	m2	v2	0.27	21.60	-12.04	35.92	-16.56
n6	m2	v3	0.27	22.20	-10.96	44.00	-15.76
n6	m3	v1	0.06	1.68	-4.64	2.48	-6.32
n6	m3	v2	0.06	1.80	-2.84	4.72	-3.92
n6	m3	v3	0.06	6.24	-0.40	10.88	-3.60
n7	m1	v1	0.88	1.04	0.76	3.12	-0.16
n7	m1	v2	0.87	8.28	-1.20	10.16	2.48
n7	m1	v3	0.87	4.12	-0.56	8.00	2.48
n7	m2	v1	0.43	3.72	-0.48	7.28	1.84
n7	m2	v2	0.40	-1.24	0.64	-1.68	3.76
n7	m2	v3	0.39	3.80	0.64	7.84	-0.40
n7	m3	v1	0.11	-0.04	-0.64	0.08	-2.40
n7	m3	v2	0.10	0.24	-1.96	3.44	-1.96
n7	m3	v3	0.09	0.44	-0.08	4.24	-0.40
n8	m1	v1	0.83	5.80	1.24	10.80	1.44
n8	m1	v2	0.83	-2.00	-4.12	-0.16	-4.16
n8	m1	v3	0.83	7.56	0.72	16.40	3.12
n8	m2	v1	0.33	0.64	0.28	0.48	-0.56
n8	m2	v2	0.32	0.72	-1.56	1.76	-3.76
n8	m2	v3	0.32	3.48	-0.52	3.28	1.36
n8	m3	v1	0.08	-1.16	-1.52	2.24	-4.32
n8	m3	v2	0.07	0.56	-2.84	0.48	-2.80
n8	m3	v3	0.07	-1.44	-3.24	-2.80	-3.60
n9	m1	v1	0.79	9.96	1.80	18.64	1.20
n9	m1	v2	0.79	16.84	-0.36	25.68	2.40
n9	m1	v3	0.79	11.52	-4.72	20.80	-7.28
n9	m2	v1	0.28	3.56	-1.60	3.56	0.08
n9	m2	v2	0.27	6.04	-1.60	9.60	-1.44
n9	m2	v3	0.27	4.92	-3.88	9.92	-5.92
n9	m3	v1	0.06	-0.04	0.44	3.12	1.68
n9	m3	v2	0.06	-0.04	-3.76	2.72	-0.16
n9	m3	v3	0.06	-1.44	2.96	0.08	5.60

Table 3 (continued)

n	m	v	α	power $\hat{P}$	power P^*	$P^* - \hat{P}$
n1	m3	v1	0.99	0.9916	0.9920	0.0004
n1	m3	v2	0.99	0.9785	0.9781	-0.0004
n1	m3	v3	0.99	0.9751	0.9759	0.0008
n2	m3	v1	0.99	0.8976	0.8992	0.0016
n2	m3	v2	0.99	0.8760	0.8765	0.0004
n2	m3	v3	0.99	0.8736	0.8732	-0.0004
n3	m3	v1	0.99	0.7955	0.7960	0.0005
n3	m3	v2	0.99	0.7686	0.7698	0.0012
n3	m3	v3	0.99	0.7689	0.7675	-0.0014
n1	m3	v1	0.95	0.9987	0.9988	0.0001
n1	m3	v2	0.95	0.9958	0.9959	0.0001
n1	m3	v3	0.95	0.9955	0.9954	-0.0001
n2	m3	v1	0.95	0.9723	0.9747	0.0024
n2	m3	v2	0.95	0.9638	0.9658	0.0019
n2	m3	v3	0.95	0.9630	0.9643	0.0013
n3	m3	v1	0.95	0.9339	0.9336	-0.0002
n3	m3	v2	0.95	0.9202	0.9207	0.0005
n3	m3	v3	0.95	0.9169	0.9192	0.0022
n1	m3	v1	0.90	0.9995	0.9996	0.0001
n1	m3	v2	0.90	0.9984	0.9985	0.0000
n1	m3	v3	0.90	0.9983	0.9983	-0.0001
n2	m3	v1	0.90	0.9873	0.9897	0.0024
n2	m3	v2	0.90	0.9833	0.9850	0.0017
n2	m3	v3	0.90	0.9827	0.9843	0.0016
n3	m3	v1	0.90	0.9675	0.9683	0.0007
n3	m3	v2	0.90	0.9595	0.9606	0.0012
n3	m3	v3	0.90	0.9576	0.9597	0.0021

Table 4: Power estimates P^* based on the chi-squared approximation to $q_{\hat{w}}$ with cumulants (3), their empirical counterparts $\hat{P}$, and absolute errors for alpha-levels 0.90, 0.95, and 0.99. Critical values were estimated from the F-approximation (4).

4.3 Confidence intervals for λ , and the mean or median of $q_{\hat{w}}$.

We have demonstrated how the non-null distribution of $q_{\hat{w}}$ can be approximated by that of $c\chi_f^2 + d$, where d, c and f are based on the formulae (3) and provided the parameters μ_i, v_i are known. In practice the parameters are unknown and can be estimated by the corresponding sample moments $\hat{\mu}_i, \hat{v}_i$, leading to $\tilde{d}, \tilde{c}, \tilde{f}$ and an approximating $\tilde{c}\chi_{\tilde{f}}^2 + \tilde{d}$. Taking the α and $1 - \alpha$

quantiles of this estimated approximating distribution $\tilde{c}\chi_{\tilde{f}}^2 + \tilde{d}$ we obtain a confidence interval for the central $1-2\alpha$ proportion of the distribution of $q_{\hat{w}}$. We wanted to know if such an interval (based only on the data) happens to provide a $1-2\alpha$ confidence interval for the mean or median of $q_{\hat{w}}$. Similarly, a confidence interval for λ may be constructed as the above interval shifted by $I - 1 + \hat{A}_0 + \hat{B}_0$, see the expression for κ_1 in (3). This interval can easily be transformed into a confidence interval for the effect size measures based on λ , such as λ/N and ρ^2 .

For each generated sample (out of 50,000 samples) we calculated the above described confidence interval. The true mean of $q_{\hat{w}}$ is well approximated by formulae (3), and the true median by that of $k_{0.5}^*$. Therefore we merely counted the number of times the interval covered the ‘unknown’ mean and median and λ and recorded the actual coverage proportions. These empirical coverage probabilities for λ for 90%, 95%, and 99% nominal confidence levels are summarised in Table 5.

The coverage probabilities are usually smaller than the nominal levels for high effect sizes, i.e. far from the null, and higher than the nominal for low effect sizes, i.e. near the null hypothesis. The relative errors of these coverage probabilities decrease with total sample size. Unbalanced sample sizes result in larger relative errors than balanced sample sizes. For 95% nominal level and small sample sizes coverage varies from 89.4% at $\{n_3m_1v_2\}$ to 96.1% at $\{n_3m_3v_2\}$. This improves to coverage from 92.0% at $\{n_6m_1v_2\}$ to 95.5% at $\{n_3m_3v_2\}$, and from 94.2% at $\{n_9m_1v_3\}$ to 95.2% at $\{n_9m_3v_3\}$. There is no obvious relationship with the variances.

Overall, these coverage probabilities are only slightly different than the nominal levels, which suggests that we can now find reasonably accurate confidence intervals for λ and for the effect size measures based on λ .

n	m	v	ρ^2	90%	95%	99%
n1	m1	v1	0.89	0.891	0.939	0.978
n1	m1	v2	0.88	0.894	0.940	0.979
n1	m1	v3	0.88	0.892	0.938	0.979
n1	m2	v1	0.45	0.903	0.948	0.986
n1	m2	v2	0.42	0.902	0.949	0.987
n1	m2	v3	0.42	0.901	0.948	0.988
n1	m3	v1	0.11	0.909	0.957	0.991
n1	m3	v2	0.10	0.914	0.958	0.992
n1	m3	v3	0.10	0.912	0.957	0.991
n2	m1	v1	0.85	0.876	0.923	0.968
n2	m1	v2	0.85	0.875	0.921	0.966
n2	m1	v3	0.85	0.876	0.925	0.968
n2	m2	v1	0.35	0.898	0.943	0.981
n2	m2	v2	0.34	0.897	0.942	0.982
n2	m2	v3	0.34	0.898	0.942	0.981
n2	m3	v1	0.08	0.915	0.956	0.989
n2	m3	v2	0.07	0.915	0.957	0.989
n2	m3	v3	0.07	0.916	0.957	0.990
n3	m1	v1	0.82	0.853	0.900	0.943
n3	m1	v2	0.81	0.846	0.894	0.939
n3	m1	v3	0.81	0.850	0.897	0.940
n3	m2	v1	0.31	0.888	0.931	0.970
n3	m2	v2	0.30	0.887	0.932	0.972
n3	m2	v3	0.29	0.887	0.932	0.970
n3	m3	v1	0.06	0.924	0.961	0.989
n3	m3	v2	0.06	0.923	0.961	0.989
n3	m3	v3	0.06	0.921	0.960	0.990
n4	m1	v1	0.88	0.895	0.943	0.984
n4	m1	v2	0.87	0.894	0.943	0.983
n4	m1	v3	0.87	0.895	0.942	0.983
n4	m2	v1	0.43	0.899	0.947	0.988
n4	m2	v2	0.40	0.901	0.948	0.987
n4	m2	v3	0.39	0.900	0.948	0.987
n4	m3	v1	0.11	0.905	0.953	0.991
n4	m3	v2	0.10	0.905	0.952	0.990
n4	m3	v3	0.09	0.905	0.953	0.991
n5	m1	v1	0.83	0.887	0.937	0.978
n5	m1	v2	0.83	0.886	0.935	0.976
n5	m1	v3	0.83	0.885	0.934	0.977

Table 5: Empirical coverage probabilities of the noncentrality parameter λ at 90%, 95%, and 99% nominal levels based on the Chi-squared approximation to $q_{\hat{w}}$ with cumulants (3). Continued next page.

n	m	v	ρ^2	90%	95%	99%
n5	m2	v1	0.33	0.898	0.945	0.985
n5	m2	v2	0.32	0.897	0.945	0.984
n5	m2	v3	0.32	0.896	0.944	0.985
n5	m3	v1	0.08	0.907	0.954	0.990
n5	m3	v2	0.07	0.909	0.954	0.990
n5	m3	v3	0.07	0.906	0.953	0.991
n6	m1	v1	0.79	0.875	0.922	0.964
n6	m1	v2	0.79	0.873	0.920	0.963
n6	m1	v3	0.79	0.873	0.921	0.962
n6	m2	v1	0.28	0.893	0.939	0.978
n6	m2	v2	0.27	0.891	0.938	0.977
n6	m2	v3	0.27	0.890	0.937	0.977
n6	m3	v1	0.06	0.910	0.955	0.990
n6	m3	v2	0.06	0.910	0.955	0.990
n6	m3	v3	0.06	0.909	0.955	0.990
n7	m1	v1	0.88	0.900	0.950	0.989
n7	m1	v2	0.87	0.897	0.948	0.989
n7	m1	v3	0.87	0.899	0.948	0.989
n7	m2	v1	0.43	0.899	0.949	0.990
n7	m2	v2	0.40	0.901	0.952	0.989
n7	m2	v3	0.39	0.899	0.948	0.990
n7	m3	v1	0.11	0.901	0.950	0.990
n7	m3	v2	0.10	0.900	0.950	0.990
n7	m3	v3	0.09	0.901	0.950	0.989
n8	m1	v1	0.83	0.897	0.948	0.988
n8	m1	v2	0.83	0.898	0.947	0.987
n8	m1	v3	0.83	0.896	0.946	0.987
n8	m2	v1	0.33	0.901	0.950	0.989
n8	m2	v2	0.32	0.900	0.948	0.989
n8	m2	v3	0.32	0.899	0.950	0.989
n8	m3	v1	0.08	0.901	0.950	0.990
n8	m3	v2	0.07	0.900	0.951	0.990
n8	m3	v3	0.07	0.900	0.951	0.989
n9	m1	v1	0.79	0.897	0.947	0.985
n9	m1	v2	0.79	0.893	0.944	0.985
n9	m1	v3	0.79	0.893	0.942	0.984
n9	m2	v1	0.28	0.899	0.947	0.988
n9	m2	v2	0.27	0.898	0.947	0.987
n9	m2	v3	0.27	0.896	0.946	0.987
n9	m3	v1	0.06	0.903	0.951	0.990
n9	m3	v2	0.06	0.901	0.951	0.990
n9	m3	v3	0.06	0.905	0.952	0.990

Table 5 (continued)

4.4 Results for the null distribution

We were interested in comparing our approximation to that by F-distribution introduced in Welch(1951)[21], and to the standard χ^2 -approximation under the null hypothesis. Principles of simulating the null distribution for $q_{\hat{w}}$ were the same as above, but some details differed by necessity. In the first part of the simulation (assuming equal means and known variances when calculating the moments (3)) only the terms with $I - 1$ and $A_0 = P_{220}$ are non-zero in (3). Relative errors $\epsilon(\hat{\alpha}^*) \equiv (\hat{\alpha}^* - \alpha)/\min\{\alpha, 1 - \alpha\}$, at the nominal levels $\alpha = 0.05$ and 0.95 for the three approximations to the null distribution of $q_{\hat{w}}$ are summarised in Table 6.

F-approximation proved to be very accurate at both tails. The standard chi-square approximation is rather accurate at the left tail, and not very good at the right tail. At the right tail its relative errors $\epsilon(\hat{\alpha}^*)$ are usually negative, i.e. $\hat{\alpha}^*$ is smaller than the nominal level α . That means that it will underestimate the p-value even for large sample sizes. The errors are the worst at the unequal sample sizes (n_3 , n_6 , and n_9). Unequal variances also don't help.

The chi-square approximation based on the moments (3) is of about the same quality as the F-approximation on the right tail. That means that it can be used as well as the F-approximation for calculation of critical values or the p-values when $p > 0.5$. On the other hand, this approximation consistently overestimates the nominal level α at the left tail. At the left tail its relative errors $\epsilon(\hat{\alpha}^*)$ are positive, i.e. $\hat{\alpha}^*$ is larger than the nominal level α . The relative errors almost do not depend on the total sample size. Unbalanced sample sizes have larger relative errors than balanced sample sizes. There is no obvious relationship with the variances.

n	v	Standard χ^2		F-distribution		χ^2 with moments (3)	
		$\epsilon(\widehat{0.05}^*)$	$\epsilon(\widehat{0.95}^*)$	$\epsilon(\widehat{0.05}^*)$	$\epsilon(\widehat{0.95}^*)$	$\epsilon(\widehat{0.05}^*)$	$\epsilon(\widehat{0.95}^*)$
n1	v1	-1.56	-19.96	0.76	-1.16	48.24	1.08
n1	v2	-0.04	-19.68	1.60	-0.96	51.32	1.24
n1	v3	-1.36	-13.08	0.16	4.16	47.32	5.88
n1	v4	-3.72	-19.76	-2.28	-2.96	47.04	-1.20
n2	v1	-2.00	-49.00	3.12	-4.44	60.24	0.00
n2	v2	-3.40	-40.52	2.72	-0.16	60.68	4.32
n2	v3	-5.92	-38.48	-1.44	-0.20	60.40	4.48
n2	v4	-2.68	-53.24	4.24	-2.64	57.64	2.48
n3	v1	-1.44	-81.04	8.72	-11.00	40.28	-4.88
n3	v2	-2.24	-78.60	7.04	-10.32	42.44	-3.00
n3	v3	-3.64	-76.32	4.88	-11.84	45.20	-6.24
n3	v4	-0.16	-63.44	7.36	-7.40	55.76	-1.16
n4	v1	0.64	-10.24	1.48	-0.80	33.36	0.64
n4	v2	0.84	-8.76	1.48	0.16	32.84	1.12
n4	v3	0.00	-11.28	0.68	-2.72	30.84	-1.24
n4	v4	-2.64	-7.12	-2.08	1.36	27.76	2.88
n5	v1	0.56	-23.60	2.56	-2.92	56.60	-0.16
n5	v2	-1.20	-24.44	0.72	-3.20	50.84	-0.40
n5	v3	-1.40	-19.96	0.48	-0.72	51.20	1.32
n5	v4	-2.20	-24.48	0.48	-0.36	59.08	2.36
n6	v1	-1.44	-35.52	2.40	-1.20	63.32	2.60
n6	v2	-4.12	-33.44	-0.48	-1.08	63.20	1.76
n6	v3	-0.48	-34.44	2.64	-2.28	60.88	1.04
n6	v4	-4.84	-29.56	-2.00	-0.44	59.60	2.96
n7	v1	-0.92	-3.92	-0.76	0.88	16.64	1.64
n7	v2	-0.44	-5.60	-0.12	-1.08	16.00	-0.40
n7	v3	-0.88	-3.12	-0.56	1.72	17.24	2.20
n7	v4	-1.96	-4.96	-1.72	0.04	13.24	0.52
n8	v1	-3.20	-12.68	-2.48	-1.28	30.64	0.12
n8	v2	-0.04	-9.84	0.48	-0.36	31.92	1.16
n8	v3	-1.80	-9.20	-0.84	-0.20	32.80	1.28
n8	v4	-1.16	-13.92	-0.48	-1.20	38.08	1.00
n9	v1	0.92	-15.52	2.36	0.44	50.12	2.32
n9	v2	-0.72	-17.00	0.96	-0.84	46.76	1.60
n9	v3	0.00	-17.48	1.76	-0.60	46.96	1.64
n9	v4	-2.00	-15.16	-0.96	-2.36	42.40	-0.52

Table 6: Relative errors $\epsilon(\hat{\alpha}^*)$ for the empirical percentages $\hat{\alpha}^* = 0.05$ and 0.95 of the 3 approximations to the null distribution of $q_{\hat{w}}$. Note that $\epsilon(\hat{\alpha}^*)$ are shown as percentages.

In practice, the parameters v_i which are used for calculating the correction term $A_0 = P_{220}$ for the Welch approximation by F-distribution and for our chi-square approximation based on the moments (3) are unknown and can be estimated by $\hat{v}_i$. This will result in an estimated value of $A_0 = P_{220}$. In the second part of the simulation under the null hypothesis we wanted to see how this will affect significance levels and critical values for the null distribution. We have used the same formulae for ‘null’ moments, but with estimated $A_0 = P_{220}$. Note that we did not include estimated parameters B_0 and P_{ijs} for $s \neq 0$ although these ‘skipped’ estimated parameters are non-zero, since their theoretical counterparts are zero under the null. We can compare empirical levels $\hat{\alpha}^*$ to nominal levels α , exactly as was described in subsection 4.2. The only difference is that now we have 50,000 of these values and their errors. To summarise this data we are using descriptive statistics. Both approximations perform equally well on the right tail, with the F-approximation being much better on the left tail. Minimum, median, upper quartile and maximum values for relative errors $\epsilon(\hat{\alpha}^*)$ at the nominal levels $\alpha = 0.05$ and 0.95 for the F- approximation are given in Table 7.

n	v	relative errors at 5% level				relative errors at 95% level			
		min	med	75%	max	min	med	75%	max
n1	v1	0.08	0.76	1.20	2.28	-9.08	0.92	1.44	4.24
n1	v2	1.00	1.68	1.96	3.16	-9.00	0.64	1.72	4.80
n1	v3	-0.16	0.16	0.40	1.56	-12.40	-4.36	-3.84	-1.72
n1	v4	-2.76	-2.16	-2.00	-0.84	-6.56	2.32	2.76	4.92
n2	v1	-0.44	2.80	4.48	12.48	-27.00	5.96	12.64	27.52
n2	v2	-1.84	1.80	4.08	12.04	-33.72	1.92	8.76	22.60
n2	v3	-4.48	-1.80	-0.60	6.80	-28.04	1.64	7.40	20.24
n2	v4	-1.20	4.00	5.60	12.88	-22.76	5.08	14.64	35.96
n3	v1	1.16	8.52	10.40	17.76	-14.28	12.16	19.88	50.40
n3	v2	0.12	6.68	8.80	16.76	-16.60	11.40	18.80	47.72
n3	v3	-1.48	4.68	6.24	15.44	-19.96	12.96	20.96	51.80
n3	v4	1.52	7.00	8.72	17.60	-22.20	9.40	17.48	40.76
n4	v1	1.44	1.48	1.56	2.12	-3.76	0.76	1.20	2.76
n4	v2	1.32	1.48	1.56	1.88	-4.08	-0.20	0.00	1.52
n4	v3	0.60	0.72	0.72	0.96	-1.44	2.60	3.20	4.04
n4	v4	-2.12	-2.04	-2.04	-1.80	-5.16	-1.64	-1.36	-0.64
n5	v1	1.52	2.52	2.80	5.32	-10.28	3.24	5.64	12.40
n5	v2	-0.28	0.68	1.04	2.80	-10.56	3.68	6.76	13.24
n5	v3	-0.44	0.36	0.76	2.92	-14.80	1.12	2.76	9.60
n5	v4	-1.56	0.44	0.76	2.36	-11.60	0.76	4.00	15.08
n6	v1	-0.04	2.40	2.76	4.60	-10.96	1.44	3.88	17.08
n6	v2	-2.44	-0.52	-0.20	1.76	-9.56	1.28	3.96	15.04
n6	v3	0.80	2.56	3.04	4.80	-10.72	2.68	5.44	20.44
n6	v4	-3.92	-2.00	-1.52	1.16	-13.64	0.76	3.72	16.36
n7	v1	-0.76	-0.76	-0.72	-0.52	-2.76	-0.92	-0.72	-0.28
n7	v2	-0.20	-0.12	-0.12	0.00	-0.08	1.08	1.16	1.40
n7	v3	-0.64	-0.56	-0.56	-0.36	-2.72	-1.72	-1.64	-1.32
n7	v4	-1.72	-1.72	-1.72	-1.52	-1.20	-0.08	-0.04	0.36
n8	v1	-2.96	-2.48	-2.40	-6.00	-4.64	1.40	2.72	5.80
n8	v2	0.20	0.48	0.52	1.24	-6.32	0.56	1.16	4.24
n8	v3	-1.24	-0.88	-0.72	-0.08	-5.76	0.20	0.84	4.00
n8	v4	-0.96	-0.48	-0.32	0.16	-5.20	1.36	2.24	7.80
n9	v1	2.00	2.36	2.48	2.92	-4.96	-0.36	0.24	4.16
n9	v2	0.40	0.96	1.32	2.12	-5.04	0.88	1.68	5.96
n9	v3	1.08	1.76	1.88	2.24	-4.64	0.64	1.76	6.44
n9	v4	-1.48	-0.96	-0.80	-0.12	-3.12	2.40	3.48	6.60

Table 7: Descriptive statistics for relative errors $\epsilon(\hat{\alpha})$ at 5% and 95% nominal levels based on the F -approximation to the null distribution of $q_{\hat{w}}$. Note that $\epsilon(\hat{\alpha}^*)$ are shown as percentages.

4.5 Confidence intervals for λ , and the mean or median of $q_{\hat{w}}$ when the null hypothesis holds

In the third part of the simulation for the null case (coverage probabilities) we have used the full formulae (3) estimating all parameters from the sample moments $\hat{\mu}_i$, $\hat{v}_i$. This difference in calculating moments from the previous subsection is due to the fact that when looking at coverage for λ we do not know whether the null hypothesis holds. We have calculated empirical coverage probabilities of the true value of $\lambda = 0$ and of the true mean and median of $q_{\hat{w}}$ only for our approximation based on (3), since the other two approximations are valid only under the null. These empirical coverage probabilities for 90% and 95% nominal confidence levels are given in Table 8. It can be seen that nominal confidence levels grossly underestimate the true coverage probabilities under the null.

Substitution of sample estimates of means and variances into the formulae (3) results in estimating λ by $q_{\hat{w}}$. We used this biased estimate of λ without much problem for the simulation under alternatives. In the hope of improving coverage we tried to estimate λ by its unbiased estimate $\hat{\lambda}_0 = q_{\hat{w}} - (I - 1 + \hat{A}_0 + \hat{B}_0)$ in (3). This may result in a negative value of $\hat{\lambda}$ and negative estimate of second moment κ_2 in (3). As a compromise we tried next to use an estimate $\hat{\lambda}_+ = \max(q_{\hat{w}} - (I - 1 + \hat{A}_0 + \hat{B}_0), 0)$. This estimate is still positively biased, but less so than $\hat{\lambda} = q_{\hat{w}}$. When we were calculating a confidence interval and met $\hat{\lambda}_+ = 0$ we also put $P_{ijs} = 0$ for $s \neq 0$, i.e. used the null distribution with estimated A_0 . The coverage probabilities did not improve and we do not reproduce them here. We do not recommend the usage of $\hat{\lambda}_+$. We believe that the over-large coverage probabilities are mostly due to the wrong shape of the chi-square distribution on the left tail, see Tables 6 and 7 in the previous Subsection.

n	v	90%			95%		
		mean	median	λ	mean	median	λ
n1	v1	0.976	0.949	0.983	0.989	0.977	0.994
n1	v2	0.975	0.949	0.983	0.989	0.977	0.994
n1	v3	0.976	0.952	0.983	0.990	0.978	0.994
n1	v4	0.975	0.948	0.983	0.989	0.977	0.994
n2	v1	0.970	0.939	0.985	0.985	0.969	0.994
n2	v2	0.973	0.942	0.986	0.987	0.972	0.994
n2	v3	0.971	0.943	0.984	0.985	0.971	0.993
n2	v4	0.969	0.938	0.984	0.984	0.967	0.992
n3	v1	0.964	0.932	0.987	0.980	0.962	0.994
n3	v2	0.965	0.932	0.987	0.981	0.963	0.994
n3	v3	0.964	0.932	0.984	0.980	0.962	0.993
n3	v4	0.967	0.935	0.985	0.982	0.966	0.993
n4	v1	0.978	0.954	0.982	0.992	0.982	0.994
n4	v2	0.978	0.955	0.982	0.991	0.982	0.993
n4	v3	0.977	0.953	0.982	0.991	0.981	0.993
n4	v4	0.979	0.956	0.982	0.992	0.982	0.994
n5	v1	0.975	0.947	0.983	0.988	0.976	0.993
n5	v2	0.975	0.947	0.983	0.990	0.977	0.994
n5	v3	0.977	0.949	0.984	0.991	0.979	0.994
n5	v4	0.974	0.946	0.983	0.988	0.975	0.994
n6	v1	0.973	0.943	0.985	0.987	0.972	0.994
n6	v2	0.973	0.943	0.986	0.988	0.973	0.994
n6	v3	0.973	0.942	0.985	0.988	0.974	0.994
n6	v4	0.973	0.945	0.984	0.987	0.974	0.994
n7	v1	0.980	0.958	0.982	0.993	0.984	0.994
n7	v2	0.980	0.957	0.981	0.993	0.984	0.994
n7	v3	0.980	0.958	0.982	0.993	0.985	0.994
n7	v4	0.980	0.958	0.982	0.993	0.985	0.994
n8	v1	0.978	0.954	0.982	0.991	0.981	0.994
n8	v2	0.978	0.954	0.982	0.992	0.981	0.993
n8	v3	0.978	0.955	0.982	0.992	0.982	0.994
n8	v4	0.977	0.953	0.982	0.991	0.980	0.994
n9	v1	0.977	0.950	0.983	0.990	0.979	0.993
n9	v2	0.976	0.950	0.982	0.990	0.978	0.993
n9	v3	0.976	0.951	0.983	0.991	0.979	0.994
n9	v4	0.975	0.951	0.982	0.990	0.979	0.993

Table 8: Empirical coverage probabilities for the mean and median of $q_{\hat{w}}$ and for the noncentrality parameter λ under the null hypothesis at 90% and 95% nominal levels based on the Chi-squared approximation to $q_{\hat{w}}$ with cumulants (3).

5 Discussion

On the basis of our simulations we can recommend the F-approximation to the null distribution of $q_{\hat{w}}$, and the chi-square approximation based on the moments (3) for the distribution under alternatives. These approximations work well for moderate to large sample sizes. We want to emphasize that normality is only required for the estimates of the population parameters of interest. We also allow for different variances in each cell.

Our simulation results demonstrate how misleading routine use of the standard chi-square approximation to the distribution of $q_{\hat{w}}$ can be. The test for homogeneity of effect sizes which is based on this approximation and is used in meta-analysis is liberal even for large sample sizes. This was illustrated on the Mumford et al (1984) [16] data. Table 6 shows how big the relative error of this test in determining the p-value can be. From Table 7 it follows that even the worst errors which can be obtained by using our improved F-approximation to the null distribution of $q_{\hat{w}}$ are smaller than the errors from the standard chi-square approximation. We hope that the techniques developed in this paper will become a part of the standard portfolio of methods for meta-analysis.

Under the alternatives the chi-square approximation based on the moments (3) leads to reliable power estimates or confidence intervals for the effect sizes. The quality of this approximation deteriorates under the null. This results in overly wide confidence intervals for the effect sizes but does not affect the quality of power estimates. A unified approach to obtaining the distribution of $q_{\hat{w}}$, perhaps based on the saddlepoint approximation using the moment generating function derived here, would be a subject of further research. Another direction of further research is to incorporate robust estimates of effect sizes.

6 Appendix: Derivation of the cumulants of the weighted sum of squares explained q_w under the alternative hypothesis

In this section we assume that the $\hat{w}_i$ are asymptotically independent from the $\hat{\mu}_i$, which are, in their turn, asymptotically normal $N(\mu_i, \sigma_i^2/n_i)$. We do not make any assumptions about the distribution of $\hat{w}_i$ apart from the existence of their central moments M_k , and their order $M_1 = O(1)$, $M_k = O(n)$, $k \geq 2$.

Welch (1951) used a symbolic extension to the order $O((n_t - 1)^{-1})$ of the moment generating function $M(u)$ for $q_{\hat{w}}$, the weighted $SS_{explained}$, to derive an approximation to the distribution of $q_{\hat{w}}$ under the null hypothesis of equal means. We formally carry out the same technique under the alternative hypothesis to obtain expressions for the first 4 cumulants of $q_{\hat{w}}$; these cumulants are the basis for a noncentral Chi-squared approximation to the distribution of $q_{\hat{w}}$. Throughout this derivation we let $\bar{\mu}$ represent $\bar{\mu}_w = \sum_{i=1}^I w_i \mu_i / \sum_{i=1}^I w_i$, the weighted mean of the group means, and $\hat{\mu}$ is its sample counterpart. We obtain

$$M(u) = E_1 E_2 \exp[u \sum \hat{w}_t (\hat{\mu}_i - \hat{\mu})^2],$$

where E_2 denotes averaging over the joint distribution of the $\hat{w}_i$, and E_1 denotes averaging over the joint distribution of the $\hat{\mu}_i$. We proceed only to the first order of $1/n_t$ in this expansion. Welsh (1951) used $w_i = n_i/\sigma_i^2$ and proceeded to the first order under the null hypothesis. Kulinskaya et al (2000) proceeded to the first order for the same weights under the general alternatives.

After some algebra we obtain

$$M(u) = \Psi \left[1 + u \sum_t \left\{ M_{1t} - \frac{M_{2t}}{\sum w_t} \right\} v_t^2 + \frac{u^2}{2} \sum_t M_{2t} v_t^4 + \frac{1}{6} u^3 \sum_t (M_{3t} - 3M_{1s} M_{2t}) v_t^6 \right], \quad (9)$$

where

$$\begin{aligned} \Psi &= \int (2\pi)^{-I/2} (w_1 w_2 \cdots w_I)^{1/2} \\ &\times \exp \left[-\frac{1}{2} \sum_t w_t (y_t - \mu_t)^2 + u \sum_t w_t (y_t - \bar{y})^2 + \sum_t (y_t - \bar{y}) \delta_t \right] dy_1 \cdots dy_I. \end{aligned}$$

and δ_t denotes differentiation with respect to v_t with v_t put equal to zero after differentiation has been carried out. The above equations contain (13)

and (14) of Welch (1951) as special cases. After integration

$$\begin{aligned}\Psi = & (1-2u)^{-\frac{I-1}{2}} \exp[-\frac{1}{2} \sum w_t (\mu_t - \bar{\mu})^2] \\ & \times \exp \left[\frac{1}{2(1-2u)} \left\{ \sum \frac{\delta_t^2}{w_t} - \frac{(\sum \delta_t)^2}{\sum w_t} + \sum w_t (\mu_t - \bar{\mu})^2 + 2 \sum \delta_t (\mu_t - \bar{\mu}) \right\} \right].\end{aligned}$$

Rearranging terms we obtain

$$\Psi = C \times \exp \left[\frac{1}{2(1-2u)} \left\{ \sum \frac{\delta_t^2}{w_t} - \frac{(\sum \delta_t)^2}{\sum w_t} + 2 \sum \delta_t (\mu_t - \bar{\mu}) \right\} \right], \quad (10)$$

where

$$C = (1-2u)^{-\frac{I-1}{2}} \exp \left[\frac{u}{1-2u} \sum w_t (\mu_t - \bar{\mu})^2 \right].$$

Expanding Ψ in powers of δ_t and neglecting all the cross-products (terms with $\delta_s \delta_t$ for $t \neq s$) and odd powers we obtain

$$\begin{aligned}\Psi = & C \left[1 + \sum \left\{ \frac{1}{2(1-2u)} \left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right) \delta_t^2 + \frac{1}{2^3(1-2u)^2} \left[\left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right)^2 \delta_t^4 + 4 \delta_t^2 (\mu_t - \bar{\mu})^2 \right] \right. \right. \\ & + \frac{1}{3!2^3(1-2u)^3} \left[12 (\mu_t - \bar{\mu})^2 \left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right) \delta_t^4 + \left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right)^3 \delta_t^6 \right] \\ & + \frac{1}{4!2^4(1-2u)^4} \left[16 (\mu_t - \bar{\mu})^4 \delta_t^4 \right. \\ & + 24 \left[\left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right)^2 \delta_t^6 (\mu_t - \bar{\mu})^2 \right] \\ & + \frac{80}{5!2^5(1-2u)^5} \left[(\mu_t - \bar{\mu})^4 \delta_t^6 \left(\frac{1}{w_t} - \frac{1}{\sum w_s} \right) \right] \\ & \left. \left. + \frac{1}{6!(1-2u)^6} [(\mu_t - \bar{\mu})^6 \delta_t^6] \right\} \right].\end{aligned}$$

It follows that

$$\begin{aligned}\Psi(v_t^2) &= C \left[\frac{1}{1-2u} \left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right) + \frac{1}{(1-2u)^2} (\mu_t - \bar{\mu})^2 \right], \\ \Psi(v_t^4) &= C \left[\frac{3}{(1-2u)^2} \left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right)^2 + \frac{6}{(1-2u)^3} (\mu_t - \bar{\mu})^2 \left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right) + \frac{1}{(1-2u)^4} (\mu_t - \bar{\mu})^4 \right], \\ \Psi(v_t^6) &= C \left[\frac{15}{(1-2u)^3} \left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right)^3 + \frac{45}{(1-2u)^4} (\mu_t - \bar{\mu})^2 \left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right)^2 \right. \\ & \quad \left. + \frac{15}{(1-2u)^5} (\mu_t - \bar{\mu})^4 \left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right) + \frac{1}{(1-2u)^6} (\mu_t - \bar{\mu})^6 \right].\end{aligned}$$

To facilitate algebra we introduce coefficients A_0 and B_0 , and polynomials P_{ijs} as

$$\begin{aligned}A_0 &= \sum_t (M_{1t} - M_{2t} / \sum w_t) \left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right), \\ B_0 &= \sum_t (M_{1t} - M_{2t} / \sum w_t) (\mu_t - \bar{\mu}_t)^2, \\ P_{ijs} &= \sum_t M_{it} \left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right)^j (\mu_t - \bar{\mu}_t)^s, \quad i = 1, 2; \\ P_{3js} &= \sum_t (M_{3t} - 3M_{1t}M_{2t}) \left(\frac{1}{w_t} - \frac{1}{\sum w_t} \right)^j (\mu_t - \bar{\mu}_t)^s.\end{aligned} \quad (11)$$

Assuming that the weights $w_i = O(n_i)$ we have the following orders: $P_{1js} = O(n^{-j})$, and $P_{ijs} = O(n^{1-j})$, for $i = 2, 3$. Substituting the expressions for $\Psi(v_t^2)$, $\Psi(v_t^4)$, and $\Psi(v_t^6)$ into $M(u)$ given by (9) we obtain

$$\begin{aligned} M(u) = C[1 + & \frac{u}{(1-2u)}A_0 + \frac{u}{(1-2u)^2}B_0 \\ & + \frac{3u^2}{2(1-2u)^2}P_{220} + \frac{3u^2}{(1-2u)^3}P_{212} + \frac{u^2}{2(1-2u)^4}P_{204} \\ & + \frac{u^3}{6}[\frac{15}{(1-2u)^3}P_{330} + \frac{45}{(1-2u)^4}P_{322} + \frac{15}{(1-2u)^5}P_{314} \\ & + \frac{1}{(1-2u)^6}P_{306}]. \end{aligned} \quad (12)$$

This and further formulae for a particular case of inverse sample variance weights were derived earlier in Kulinskaya et al(2001). To simplify further calculations we define the following family of functions $\phi_0 = \ln(1-2u)$, $\phi_i = (1-2u)^{-i}$, $i = 1, 2, \dots$. We have the following relationships for the derivatives:

$$\begin{aligned} \phi_0^{(s)} &= -2^s(s-1)!\phi_s, \\ \phi_I^{(s)} &= 2^s((I+s-1)!/(s-1)!) \phi_{I+s}. \end{aligned} \quad (13)$$

Using the above notation, the cumulant generating function is

$$\begin{aligned} K(u) &= -\frac{I-1}{2}\phi_0 + \frac{1}{2}(\phi_1 - 1)(\lambda + A_0) + \frac{1}{2}(\phi_2 - \phi_1)B_0 \\ &+ \frac{3}{8}(1 - 2\phi_1 + \phi_2)P_{220} + \frac{3}{4}(\phi_1 - 2\phi_2 + \phi_3)P_{212} \\ &+ \frac{1}{8}(\phi_2 - 2\phi_3 + \phi_4)P_{204} + \frac{15}{48}(\phi_3 - 3\phi_2 + 3\phi_1 - 1)P_{330} \\ &+ \frac{45}{48}(\phi_4 - 3\phi_3 + 3\phi_2 - \phi_1)P_{322} + \frac{15}{48}(\phi_5 - 3\phi_4 + 3\phi_3 - \phi_2)P_{314} \\ &+ \frac{1}{48}(\phi_6 - 3\phi_5 + 3\phi_4 - \phi_3)P_{306}. \end{aligned} \quad (14)$$

We also need the following constants:

$$\begin{aligned} C_0 &= -\frac{I-1}{2}, \\ C_1 &= [\lambda + A_0 - B_0 - \frac{3}{2}P_{220} + \frac{3}{2}P_{212} + \frac{45}{24}P_{330} - \frac{45}{24}P_{322}]/2, \\ C_2 &= [B_0 + \frac{3}{4}P_{220} - 3P_{212} + \frac{1}{4}P_{204} - \frac{45}{24}P_{330} + \frac{135}{24}P_{322} - \frac{15}{24}P_{314}]/2, \\ C_3 &= [3P_{212} - P_{204} + \frac{15}{12}P_{330} - \frac{135}{12}P_{322} + \frac{45}{12}P_{314} - \frac{1}{12}P_{306}]/4, \\ C_4 &= [\frac{45}{12}P_{322} + \frac{1}{2}P_{204} - \frac{45}{12}P_{314} + \frac{3}{12}P_{306}]/4, \\ C_5 &= [15P_{314} - 3P_{306}]/48, \\ C_6 &= [P_{306}]/48. \end{aligned} \quad (15)$$

In the above notation we have

$$K(u) = const + C_0\phi_0 + \sum_{j=1}^6 C_j\phi_j. \quad (16)$$

From the form of derivatives given in equation (13) we obtain the s -th cumulant as

$$\begin{aligned} \kappa_s = & 2^s (s-1)! [-C_0 + sC_1 + s(s+1)C_2 + \frac{s(s+1)(s+2)}{2}C_3 + \frac{s(s+1)(s+2)(s+3)}{3!}C_4 \\ & + \frac{s(s+1)(s+2)(s+3)(s+4)}{4!}C_5 + \frac{s(s+1)(s+2)(s+3)(s+4)(s+5)}{5!}C_6]. \end{aligned}$$

In particular, we can obtain the first four cumulants as given below:

$$\begin{aligned} \kappa_1 &= 2(-C_0 + C_1 + 2C_2 + 3C_3 + 4C_4 + 5C_5 + 6C_6), \\ \kappa_2 &= 4(-C_0 + 2C_1 + 6C_2 + 12C_3 + 20C_4 + 30C_5 + 42C_6), \\ \kappa_3 &= 16(-C_0 + 3C_1 + 12C_2 + 30C_3 + 60C_4 + 105C_5 + 168C_6), \\ \kappa_4 &= 96(-C_0 + 4C_1 + 20C_2 + 60C_3 + 140C_4 + 280C_5 + 504C_6). \end{aligned} \tag{17}$$

Explicitly,

$$\begin{aligned} \kappa_1 &= I - 1 + \lambda + A_0 + B_0, \\ \kappa_2 &= 2(I - 1 + 2\lambda + 2(A_0 + 2B_0) + 3P_{220}/2 + 3P_{212} + P_{204}/2), \\ \kappa_3 &= 8(I - 1 + 3\lambda + 3(A_0 + 3B_0) + 9P_{220}/2 + 27P_{212}/2 + 3P_{204} \\ &\quad + P_{306}/8) + 15P_{330}/8 + 15 \cdot 183P_{322}/8 + 15P_{314}/8. \end{aligned} \tag{18}$$

Note that in expression (18) the term with P_{330} is of the order $O(n^{-2})$, P_{322} is of the order $O(n^{-1})$, and P_{314} is of the order $O(1)$, whereas P_{306} is of the order $O(n)$. So far the first three of the above terms continued to be carried over, but they are skipped in the expression for κ_3 in (3) above. At the preliminary stage of simulations we discovered the superiority of the resulting approximation.

Under the null hypothesis only the terms with A_0 , and P_{ij0} are non-zero, and the F-approximation similar to one used by Welch (1951) should be used.

References

- [1] G.E.P. Box. Some theorems on quadratic forms applied in the study of analysis of variance problems. I. effect of inequality of variance in the one-way classification. *Annals of Mathematical Statistics*, 25:290–302, 1954.
- [2] M.B. Brown and A.B. Forsythe. The small-sample behavior of some statistics which test the equality of several means. *Technometrics*, 16:129–132, 1974.

- [3] W.G. Cochran. The combination of estimates from different experiments. *Biometrics*, 10:101–129, 1954.
- [4] J. Cohen. *Statistical power analysis for the behavioural sciences, 2nd Edition*. L. Erlbaum Associates, Hillsdale, N.J., 1988.
- [5] J.L. Fleiss. The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, 2:121–145, 1993.
- [6] L.V. Hedges and I. Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, London, 1985.
- [7] G.S. James. The comparison of several groups of observations when the ratios of the population variances are unknown. *Biometrika*, 38:324–329, 1951.
- [8] G.S. James. Tests of linear hypotheses in univariate and multivariate analysis when the ratios of the population variances are unknown. *Biometrika*, 41:19–43, 1954.
- [9] G.S. James. On the accuracy of weighted means and ratios. *Biometrika*, 43:304–321, 1956.
- [10] S. Johansen. The Welch-James approximation to the distribution of the residual sum of squares in a weighted linear regression. *Biometrika*, 67:85–92, 1980.
- [11] H.J. Keselman and R. R. Wilcox. The ‘improved’ Brown and Forsythe test for mean equality: Some things can’t be fixed. *Communications in Statistics- Simulation and Computation*, 28:687–698, 1999.
- [12] E. Kulinskaya, R.G. Staudte, and A. Berezner. Confidence intervals for R^2 and power calculations for unequal means in a one-way ANOVA weighted for unequal variances. Technical Report 2000-16, La Trobe University, 2000.
- [13] Hedges L.V. and Pigott T.D. The power of statistical tests in meta-analysis. *Psychological Methods*, 6:203–217, 2001.

- [14] D. V. Mehrotra. Improving the Brown-Forsythe solution to the generalized Behrens-Fisher problem. *Communications in Statistics. Simulation and Computation*, 26:1139–1145, 1997.
- [15] R.G. Miller, Jr. *Beyond ANOVA: Basics of Applied Statistics*. Chapman and Hall, London, 1997.
- [16] E. Mumford, H.J. Schlesinger, C. Glass, G.V. and Patric, and T. Cuerdon. A new look at evidence about reduced cost of medical utilization following mental health treatment. *American Journal of Psychiatry*, 141:1145–1158, 1984.
- [17] W.R. Shadish and C.K. Haddock. *Combining estimates of effect size*. The Handbook of Research Synthesis. Russel Sage Foundation, New York, 1994.
- [18] W.Y. Tan. On approximating probability distributions of a statistic of Brown-Forsythe from normal and nonnormal universes. *J.Statist. Comput. Simulation*, 16:35–55, 1982.
- [19] W.Y. Tan. On analysis of covariance under heteroscedasticity and robustness with respect to nonnormality. *Biom.J.*, 29:3–9, 1987.
- [20] W.Y. Tan and M.A. Tabatabai. Harmonic mean approach to regression analysis under heteroscedasticity and nonnormality. *Biom.J.*, 28:815–824, 1986.
- [21] B.L. Welch. On the comparison of several mean values: an alternative approach. *Biometrika*, 38:330–336, 1951.
- [22] A. Whitehead and J. Whitehead. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, 10:1665–1677, 1991.
- [23] R.R. Wilcox. *Introduction to robust estimation and hypothesis testing*. Academic Press, San Diego, CA, 1997.