Meta-analysis of ratios of sample variances

Luke A. Prendergast*† and Robert G. Staudte

When conducting a meta-analysis of standardized mean differences (SMDs), it is common to use Cohen’s $d$, or its variants, that require equal variances in the two arms of each study. While interpretation of these SMDs is simple, this alone should not be used as a justification for assuming equal variances. Until now, researchers have either used an $F$-test for each individual study or perhaps even conveniently ignored such tools altogether. In this paper, we propose a meta-analysis of ratios of sample variances to assess whether the equality of variances assumptions is justified prior to a meta-analysis of SMDs. Quantile–quantile plots, an omnibus test for equal variances or an overall meta-estimate of the ratio of variances can all be used to formally justify the use of less common methods when evidence of unequal variances is found. The methods in this paper are simple to implement and the validity of the approaches are reinforced by simulation studies and an application to a real data set. Copyright © 2016 John Wiley & Sons, Ltd.

Keywords: $F$-test; normalization; variance stabilization; maximum likelihood estimation; test for unequal variances

1. Introduction

The $F$-distribution (see, for e.g., Chapters 27 and 30 of [1]) is encountered in many applications of statistics including the testing of equality of variances, analysis of variance (ANOVA), and more broadly a test for the overall model in linear regression analyses. A problem that may be of interest to many is how best to combine the evidence from several independent $F$-distributed random variables to improve power. For example, suppose that we have $K$ independent studies reporting $F$-statistics (or summary measures that can be used to calculate an $F$-statistic) for the testing of the same hypotheses. Rather than relying on $K$ individual hypothesis tests, we seek to combine these $F$-statistics to obtain a single measure that can be used to test the hypotheses with increased power or to obtain an improved estimation of an associated parameter.

Of particular interest here is a meta-analysis for standardized mean differences (SMDs). In each study, there are two populations with means $\mu_1$ and $\mu_2$ and the same variance $\sigma^2$, and the population SMD is $\delta = (\mu_1 - \mu_2)/\sigma$. An estimate of $\delta$ is Cohen’s $d$ [2], denoted $d = (\bar{x}_1 - \bar{x}_2)/s_p$, where $\bar{x}_1$ and $\bar{x}_2$ are sample means estimates for $n_1$ and $n_2$ sampled observations from each of the populations, and $s_p^2$ is the pooled sample variance estimate of $\sigma^2$.

Cohen [2] suggested a ‘rule of thumb’ for the SMD where 0.2 is considered small, 0.5 medium, and 0.8 large, and values of $d$ from studies with data collected on different measurement scales can be readily compared. It is important here to ensure that the assumption of equal variances is justified, or at least that there is a lack of evidence to suggest that the population variances are not equal, and one possibility is to conduct an $F$-test for equality of variances [1, p.323] in each study. Our interest is on the availability of more than one study and on whether the assumption of equal variances is justified in general.

While many tests can be reasonably robust to even moderate departures from equal variances (e.g., ANOVA), this is an entirely different scenario. Here, the equal variances assumption is used to obtain a simple-to-interpret measure, $\delta$, or its variations. If the variances are not equal, then the estimate $s_p$.
may not be a good measure of the overall variance, and the resulting estimate \(d\) may be misleading. We emphasize this point in the next section.

In Section 2, we will provide a motivating example before discussing the transformation of \(F\)-statistics in Section 3. Maximum likelihood estimators are discussed in Section 4 where both fixed effect and random effects models are considered. Simulations are provided in Section 5 that assess the performance of estimators of the ratio of variances. In Section 6, we reconsider the example in Section 2 and carry out the meta-analysis for the ratios of sample variances. Concluding remarks are provided in Section 7.

2. A motivating example

Thakkinstian et al. [3] reported on 15 studies with the data shown for 13 of these given in Table I. Because two of the 15 studies were not used in their meta-analysis of SMDs, which we consider shortly, we have not included these studies in our table.

There are three groups for each of the 13 studies, namely, BB, Bb, and bb, which refer to genotypes of individuals studied. The estimate is the mean spinal bone mass density (BMD) for premenopausal women, and the question is whether there is a difference in mean BMD with respect to genotype. In a meta-analysis of SMDs, Thakkinstian et al. [3] compare the combined group of Bb with bb, which has the additional benefit of alleviating some small sample sizes that may undermine the assumed normality of the estimated SMDs that is required for the meta-analysis (for more on meta-analysis of SMDs, including normality of the estimates, see, for e.g., Borenstein et al. [4]). However, some small sizes are still present for the BB group. The summary statistics for this combined group are also included in Table I and is our focus here.

For comparison of the variances between the BB and Bb/bb groups, we take the ratio of the estimated variances giving, to two decimal places,

\[
0.79, 0.01, 0.64, 1.46, 1.81, 1.32, 0.34, 1.33, 2.16, 1.80, 1.01, 1.17, 1.11
\]

and we refer to these values as \(f_1, \ldots, f_{13}\), respectively. In the scientific literature, it is common to assume that BMD is normally distributed (see, for e.g., [5]). Under this assumption of normality of the underlying populations, if the true variances are equal between the two groups then the distribution of the \(i\)th ratio is \(F_{n_{1i} - 1, n_{2i} - 1}\) (\(i = 1, \ldots, 13\)) where \(n_{1i}\) and \(n_{2i}\) are the sample sizes for each of the BB and Bb/bb groups data in the \(i\)th study. We reject the possibility of equal variances in a test of variance equality from the \(i\)th study if \(f_i\) is either too small or too large. Later, we provide the \(p\)-values for each of the 13 tests:

\[
0.838, 0.141, 0.360, 0.377, 0.010, 0.090, 0.130, 0.049, 0.000, 0.009, 0.911, 0.617, 0.714
\]

| Study | n | Mean | SD | n | Mean | SD | n | Mean | SD | n | Mean | SD |
|-------|---|------|----|---|------|----|---|------|----|---|------|----|
| BB    |   |      |    | Bb |      |    | bb|      |    |   |      |    |
| BB    |   |      |    | Bb |      |    | bb|      |    |   |      |    |
| BB    |   |      |    | Bb |      |    | bb|      |    |   |      |    |
| BB    |   |      |    | Bb |      |    | bb|      |    |   |      |    |
| BB    |   |      |    | Bb |      |    | bb|      |    |   |      |    |
| BB    |   |      |    | Bb |      |    | bb|      |    |   |      |    |
| BB    |   |      |    | Bb |      |    | bb|      |    |   |      |    |
| BB    |   |      |    | Bb |      |    | bb|      |    |   |      |    |
| BB    |   |      |    | Bb |      |    | bb|      |    |   |      |    |
| BB    |   |      |    | Bb |      |    | bb|      |    |   |      |    |
| BB    |   |      |    | Bb |      |    | bb|      |    |   |      |    |
| BB    |   |      |    | Bb |      |    | bb|      |    |   |      |    |

Table I. Data from Table 2 of [3] (excluding two studies that were not part of their meta-analysis).

Sample sizes (n), means and the standard deviations (SD) for three genotype groups and including the combined grouping of Bb and bb (Bb/Bb).
with the fifth, sixth, eighth, ninth, and tenth studies providing some evidence of unequal variances. Clearly, the smallest of the \( p \)-values is very small, but this may suggest that a trait specific to this study has resulted in unequal variances. A question one should ask in such a situation is whether a meta-analysis of the SMDs is sufficiently robust in regard to departures from equal variances. In this particular example, it appears that the data across studies is measured on the same scale. Consequently, the raw mean difference could be used instead of the SMD, which is often used not only for interpretative reasons but also to combine results from studies for which data has been measured on different scales. When studies do collect data on different scales, another option is to consider the ratio of means, or to be precise, the log ratio of means [6] or to use Glass’s effect [7] which uses the variance estimate from one group only.

In Table II, we provide simulated coverage probabilities (1000 trials) for interval estimators for Cohen’s \( d \) (SMD), the raw mean difference, and the log ratio of means. Sample sizes were chosen equal to those for the comparison between the BB and Bb/bb groups where data was sampled from normal distributions with equal means (both 1.0) and \( \rho = \sigma_1^2/\sigma_2^2 \) shown in the table. The results were obtained using the metafor R package [8] with a restricted maximum likelihood estimate for the variance of an assumed random effect. As can be seen in the table, while the coverage probabilities for the mean difference and log ratio of means are typically close to nominal (0.95), the coverage for the SMD is sensitive to \( \rho \) with close to nominal coverage only for two choices of \( \rho \), one of which is for equal variances. This single example highlights some dangers with assuming equal variances, and a more exhaustive search for examples will no doubt yield even more evidence. It should be pointed out that while the use of log ratio of means can be used when study variables are on a different scale, in practice it is very common to use Cohen’s \( d \).

The aforementioned simulation reported in Table II considered only the case when both means are equal. When they are not equal, interpretation of Cohen’s \( d \) becomes difficult because the estimate is highly dependent on the sample sizes when the variances are not equal. As an example, we consider a total sample size of \( 200 = n_1 + n_2 \) and focus on the estimated Cohen’s \( d \) when \( n_1 = 10, 50, 75, 100, 125, 150, \) and 190. For each choice of \( n_1 \), we generate \( n_1 \) values from \( N(1.1, 0.12^2) \), and \( n_2 = 100 - n_1 \) values from the \( N(1.0, 0.2^2) \) distribution. This is repeated 10,000 times, and in Table III we report the average and standard deviations of the estimates. We can see from the reported means that, on average, the estimated \( d \) varies from moderate to large depending on the sample sizes. Additionally, the reported standard deviations also suggest that it would not be unusual to observe small to very large estimates again depending on the sample size allocations. As this example highlights, interpretation of Cohen’s \( d \), when the group standard deviations are not equal, can be problematic.

Returning to our example data, what we would like to know is whether these results provide persistent evidence of unequal variances in general and as a consequence mean that one should use SMDs with caution, if at all. Rather than leave this as a subjective problem, we provide a more formal approach to answering this question, which may benefit researchers by providing at least some formal justification for a lack of evidence or as a means to move away from the assumption of equal variances and adapt the meta-analysis accordingly. Of course one option is to not use Cohen’s \( d \) altogether. However, given its widespread use and interpretation, this is unlikely to occur.

| Table II. Simulated coverage probabilities for interval estimators of the SMD, MD, and MR for varying choices of \( \rho \). |
|---|
| Est. | 0.15^2/0.15^2 | 0.15^2/0.2^2 | 0.1^2/0.2^2 | 0.1^2/0.3^2 | 0.2^2/0.15^2 | 0.2^2/0.2^2 | 0.3^2/0.1^2 |
| SMD | 0.956 | 0.981 | 0.998 | 1.00 | 0.955 | 0.924 | 0.916 |
| MD | 0.961 | 0.954 | 0.955 | 0.957 | 0.968 | 0.959 | 0.942 |
| MR | 0.962 | 0.953 | 0.957 | 0.963 | 0.966 | 0.957 | 0.935 |

SMD, standardized mean difference; MD, mean difference; MR, mean ratio.

| Table III. Mean and standard deviations (from 10,000 iterations) of simulated Cohen’s \( d \) for \( n_1 \) observations generated from \( N(1.1, 0.12^2) \), \( n_2 = 200 - n_1 \) from the \( N(1.0, 0.2^2) \) distribution. |
|---|
| \( n_1 \) | 10 | 50 | 75 | 100 | 125 | 150 | 190 |
| Mean \( d \) | 0.505 | 0.544 | 0.576 | 0.606 | 0.649 | 0.694 | 0.849 |
| SD \( d \) | 0.210 | 0.131 | 0.134 | 0.145 | 0.169 | 0.212 | 0.514 |

SD, standardized deviation.
3. Variance stabilization of $F$ statistics

An obvious way to query whether the assumption of equal variances is justified across $K$ studies is to consider the ratio of the estimated variances. However, with sample sizes often varying greatly among studies, interpreting the estimated ratios (both collectively and individually) is not straightforward. Our first focus will therefore be on re-scaling the estimated ratios of variances so that they are all on an easily interpreted scale.

Variance stabilization transformations seek to both normalize and scale an estimate; in this case, we are seeking an approximate variance of one. The benefits can include improved coverage of confidence intervals and a transformed test statistic that is easy to interpret. Variance stabilization for commonly encountered estimates are still leading to improved inference, as is evidenced by recent research and in the meta-analysis setting goes back at least to [9]. Recent contributions include, for example, Kulinskaya et al. [10], who use it systematically in every chapter; Kulinskaya et al. [11] and Prendergast and Staudte [12], who consider variance stabilization of the difference in binomial proportions; Rücker et al. [13], who advocate variance stabilize transformed effects in the two-sample binomial setting; [14] who do the same for the one-sample binomial; Malloy et al. [15, 16], who variance stabilize t-distributed effects; and Jackson et al. [17], who variance stabilize an untruncated estimate of the between-study variance. For more discussion on some of the advantages of variance stabilization see Morgenthaler and Staudte [18]. In this section, we describe variance stabilization of ratios of estimated variances, which is assumed to be $F$-distributed when the data has been sampled from two independent normal distributions. Simulations will follow in Section 5.

Given $S \sim F_{v_1,v_2}$, the goal is to find a variance stabilizing transformation (VST) $T = T(S)$ for which $T \sim N(0, 1)$ approximately. For a single study such a transformation is not necessary because probabilities and percentiles from the $F$ distribution are easily obtained. However, one small advantage in the single study setting is that the transformed statistic is immediately interpretable because of the distribution being free of the degrees of freedom. This small advantage for a single study leads to a much bigger advantage for multiple studies. In what follows, we explore four possible VSTs that may be used to achieve such advantages.

3.1. Some possible transformations

Let $\Phi^{-1}$ denote the inverse cumulative distribution function (inverse cdf) for the standard normal distribution, and let $F$ denote the cdf for the $F_{v_1,v_2}$ distribution. Then, as pointed out by a referee, a possible normalizing transformation is

$$T_0(S) = \Phi^{-1}[F(S)]$$ (1)

where $T_0(S) \sim N(0, 1)$ when $S \sim F_{v_1,v_2}$. Note that the distribution of $T_0(S)$ is free of $v_1$ and $v_2$ so that several transformed $F$-statistics can easily be interpreted and compared even when their degrees of freedom differ.

While the transformation $T_0$ can be obtained computationally, other simple transformations also exist and which may offer some additional advantages. One such transformation is the VST defined

$$T_1(S) = \frac{1}{\sqrt{c_1}} \ln\left(\frac{S}{\text{E}[S]}\right) + \frac{\sqrt{c_1}}{2}$$ (2)

where $c_1 = 2(v_1 + v_2 - 2)/(v_1(v_2 - 4))$, and $\text{E}[S] = v_2/(v_2 - 2)$ is the mean for the $F_{v_1,v_2}$ distribution for $v_2 > 2$. When $S \sim F_{v_1,v_2}$, $T_1(S)$ has approximate mean zero and approximate variance one. In some situations, this VST can offer several advantages especially when considering the ratio of sample variances. Firstly, we shall see later in Section 3.4 that simple closed-form expressions for estimators of assumed fixed ratio of variances obtained from several studies can be found and which agree with a maximum likelihood estimator derived in the following section. Secondly, even when the true ratio of variances is not equal to one (i.e., when it is not true that $S \sim F_{v_1,v_2}$ because $\rho \neq 1$), the approximate variance of $T_1(S)$ remains unchanged so that it is a VST; technical details for this VST are provided in Appendix A.1.

In our search for further normalizing and variance stabilizing transformations, we arrived at three other possibilities. While the advantages of $T_0$ and $T_1$ are enough to justify our focus from this point onwards,
we give a brief description of these transformations in Appendix A.2 because they may be useful in further research. We also provide some simulation details comparing the performance of all of the VSTs, including $T_1$ in Appendix A.3.

### 3.2. Quantile–quantile plots to assess violations of the assumption of equal variances

Returning to our meta analysis of SMD’s where we want a meta-analysis of variance ratios, assume that we have $K$ independent $F$-distributed random variables denoted $S_1, \ldots, S_K$ with potentially different degrees of freedom (denoted $\nu_{1i}$ and $\nu_{2i}$ for the $k$th study) and population variances $\sigma^2_{1i}$ and $\sigma^2_{2i}$. When $\sigma^2_{1i} \neq \sigma^2_{2i}$, then the $S_i$'s are distributed as $(\sigma^2_{1i}/\sigma^2_{2i}) F_{\nu_{1i},\nu_{2i}}$ random variables.

Below, let $T_i$ denote one of the transformations $T_0$ or $T_1$ given in Section 3.1. We have

$$Z_k = T_i(S_i; \nu_{1i}, \nu_{2i}) \sim \mathcal{N}(0, 1) \quad \text{if} \quad \sigma^2_{1i} = \sigma^2_{2i} \quad \text{for each} \quad k$$

but where this is an approximate distribution for the VST $T_1$. Because the $S_i$'s are independent, if the variances within the $K$ studies are in fact equal, then the ordered $Z_k$'s should resemble data randomly generated from a standard normal distribution. Let $Z_{[1]} \leq \ldots \leq Z_{[K]}$ denote the ordered $Z_k$'s. Then one could visually assess the possibility of violations of the equal variance assumption by utilizing quantile–quantile (Q–Q) plots of the

$$Z_{[k]} \quad \text{versus} \quad \Phi^{-1}\left(\frac{k - 0.5}{K}\right)$$

where $\Phi^{-1}(p) = z_p$ is the $p \times 100$th percentile of the standard normal distribution such that $P(Z \leq z_p) = p$ for $Z \sim \mathcal{N}(0, 1)$.

Returning to our motivating example of Section 2, in Figure 1, we provide the Q–Q plots for each of the $T_0$ and $T_1$. The plots are similar with only some minor differences and a more in depth comparison of the VST $T_1$, and its ability to achieve approximate standard normality will be explored in greater depth later in the Appendix. Both Q–Q plots suggest potential problems with the assumption of equal variances with a trend towards larger than expected $Z_k$’s. While these Q–Q plots are informative and easy to obtain, in this example and others, some researchers would like an accompanying test for unequal variances. This is introduced in the next section. For those who prefer to interpret a meta-analysis of effect sizes, in this case the ratio of variances, one can skip to the following section. Both methodologies are based on the same combination of transformed $F$-test statistics.

### 3.3. Omnibus tests for unequal variances

Continuing with the notation introduced in the previous section, it is simple to combine the $Z_k$’s to obtain a simple-to-interpret meta-analytic test statistic to test for unequal variances. In the description that follows, we will assume $Z_k = T_0(S_i) \sim \mathcal{N}(0, 1)$ and note that this standard normality can also be achieved approximately using $T_1$. One obvious test statistic is $Z = \sqrt{K} \times \bar{Z} = K^{-1/2} \sum_{k=1}^{K} Z_k$ that is $\mathcal{N}(0, 1)$ under
the assumption of equal variances. However, this test statistics may lack power in situations for which
the true ratios of variances among studies varies between being less than one and greater than one (as
may be the case when a random effect is present). Alternatively, the test statistic

\[ X^2 = \sum_{k=1}^{K} Z_k^2 \sim \chi^2_K \]  

(3)

under the null hypothesis \( H_0 : \sigma_{k1}^2 = \sigma_{k2}^2 \) for every \( k = 1, \ldots, K \) does not suffer from this problem. That
is, this test statistics will be suitable to provide evidence of either a fixed \( \rho \) across studies that is not equal
to one, a \( \rho \) that varies among studies or a combination of the two.

The aforementioned test statistic applies an equal weighting for each study. However, another pos-
sibility would be to apply a greater weighting to studies with smaller estimator variability on the
non-transformed scale. Hence, a weighted version of the aforementioned is

\[ X_{w}^2 = \sum_{k=1}^{K} w_k Z_k^2 \]  

(4)

where \( w_1, \ldots, w_K \) are nonzero weights where \( \sum_{k=1}^{K} w_k = 1 \). Under the assumption of \( \sigma_{k1}^2 = \sigma_{k2}^2 \), the
distribution of \( X_{w}^2 \) is not straightforward although some approximating distributions do exist (see, for e.g.,
[19]). However, for what we require shortly, it is quite efficient to simply use a Monte Carlo simulation
to accurately obtain what is needed. In terms of the weights, there are many possibilities and we leave
further guidance on this until the next section.

A \( p \)-value associated with the test statistic in (4) is not exactly determinable. However, to obtain an
accurate approximation via simulation, one simply needs to randomly generate \( M X_{w}^2 \)'s under the null
using \( \mathbf{Xw} \) where \( \mathbf{X} \) is an \( M \times K \) matrix of randomly generated \( \chi^2_k \) values and \( \mathbf{w} \) is the column vector of
weights. The simulated \( p \)-value is then simply the proportion of values in the resulting vector that are
greater than the observed test statistic.

Other tests also present themselves as possibilities. For example, if it is reasonable to assume that
when the null is not true that the distribution of the \( Z_k \)’s are normal and identically distributed (i.e., \( N(\mu, \sigma^2) \)
distributed with \( \mu \neq 0 \) and/or \( \sigma^2 \neq 1 \) ) then the joint likelihood exists in a simple form under the alternative
and the maximum likelihood estimators and the sample mean and variances of the \( Z_k \)’s (the observed
\( Z_k \)’s denoted \( \bar{z} \) and \( s \), respectively. Consequently, a generalized likelihood ratio is \( \lambda(\mathbf{z}) = \lambda(z_1, \ldots, z_K) = \prod_{k} \phi(z_k; 0, 1)/\prod_{k} \phi(z_k; \bar{z}, s^2) \) where \( \phi(\cdot; \mu, \sigma^2) \) is the density function for the \( N(\mu, \sigma^2) \) distribution. For
large \( K \), a likelihood ratio test (LRT) can be used where \( -2 \ln[\lambda(\mathbf{z})] \sim \chi^2 \) under the assumption of equal
variances and an approximate \( p \)-value can be computed accordingly. While the assumption of \( N(\mu, \sigma^2) \)
over all studies when there are unequal variances is hard to justify (because the distribution would depend,
for example, on sample sizes within studies), this \( p \)-value is computed under the assumption of equal
variances and is therefore appropriate for large \( K \). A brief simulation is presented later in Section 5.2.

3.4. Estimates of an assumed fixed ratio of variances

In this section, we provide our first insights into estimating an assumed fixed ratio of variances. While
the following estimators are motivated by the VSTs, it should be noted that there is a strong link with
maximum likelihood estimators (MLEs) that we consider in the next section. The MLEs also allow for
the simple introduction of a random effect and this will also be considered in the next section.

Suppose that \( \sigma_{k1}^2/\sigma_{k2}^2 = \rho \) for all \( k = 1, \ldots, K \). That is, the ratio of variances is assumed fixed across all
studies where \( \rho = 1 \) is assumed when utilizing SMDs. In this section, we consider combining evidence
across all studies to obtain point and interval estimates for \( \rho \). Under the assumption that data are sampled
from independent normal distributions, we have that \( S_k/\rho \sim F_{n_{k1}-1, n_{k2}-1} \) so that, from Section 3.3,

\[ T_1(S_k/\rho; v_{k1}, v_{k2}) \sim N(0, 1) \]

and subsequently

\[ Z_{tw} = \frac{1}{\sqrt{\sum_{k=1}^{K} w_k^2}} \times \sum_{k=1}^{K} w_k T_1(S_k/\rho; v_{k1}, v_{k2}) \sim N(0, 1) \]  

(5)
for $T_0$ and approximate standard normality for $T_1$.

Let $c_{k1}$ denote the $k$th study specific $c_1$ used in (2), and also let $E_0(S_k) = E(S_k/\rho) = \nu_{k2}/(\nu_{k2} - 1)$ denote the expected value of $S_k$ under the null hypothesis of equal variances. It is free of $\rho$. Consequently, $Z_{\rho}$ is a pivotal quantity for $\rho$ and an approximate $(1 - \alpha) \times 100\%$ confidence interval for $\rho$ can be obtained by solving

$$|Z_{\rho}| = z_{1-\alpha/2}$$

where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)100\%$ percentile of the $N(0, 1)$ distribution.

For illustrative purposes, we shall focus our attention on $T_1$ from (2). For this simple VST, formulae for the interval estimates of $\rho$ can be derived as closed-form expressions, which will aid in discussion. Estimates for the other transformations can be obtained through simple computational root-solving. Using the definition of $T_1$, simple rearranging of (5) leads to

$$\left(\sum_{k=1}^{K} \frac{w_k}{\sqrt{c_{k1}}}\right)^{-1} \sum_{k=1}^{K} \frac{w_k}{\sqrt{c_{k1}}} \left[\ln(S_k) + \ln\left(\frac{\nu_{k2} - 3}{\nu_{k2} - 1}\right) + \frac{c_{k1}}{2}\right] \approx N[\ln(\rho), V]$$

where

$$V = \left(\sum_{k=1}^{K} \frac{w_k}{\sqrt{c_{k1}}}\right)^{-2} \sum_{k=1}^{K} w_k^2.$$

Therefore, the left hand side of (6) is an estimator of $\ln(\rho)$ and which has variance $V$. Consequently, we denote this estimator as $\hat{\ln}(\rho)$ and a $(1 - \alpha) \times 100\%$ confidence interval for $\ln(\rho)$ is $\hat{\ln}(\rho) \pm z_{1-\alpha/2}\sqrt{V}$. Finally, we exponentiate to obtain our confidence interval for $\rho$ given as

$$\left(\exp[\hat{\ln}(\rho)] - z_{1-\alpha/2}\sqrt{V}, \exp[\hat{\ln}(\rho)] + z_{1-\alpha/2}\sqrt{V}\right)$$

where the $\ln(\rho)$ estimate and $V$ are given in (6). This confidence interval is simple to compute, which is an advantage for the $T_1$ VST. For the other VSTS, the estimates can be obtained computationally which, using a package such as R, is a simple task.

So far in this section, we have not provided any guidance as to suitable choices for the weights $w_1, \ldots, w_K$. In the next section, we show that choice of

$$w_k = \frac{1}{\sqrt{c_{k1}}}, \quad k = 1, \ldots, K$$

in (6) results in an MLE of $\rho$ (with an equivalent variance $V$) based on the log of the ratio of sample variances. This will therefore provide motivation for the choice of weights. However, future work may include different choices of weights that may, for example, be used to provide some degree of robustness in the estimation to protect against outliers.

### 4. Meta-analysis based on maximum likelihood estimation

In the previous section, we considered meta-analysis for the ratio of variances following a variance stabilization of the $F$-test statistics. Another possibility is to consider MLEs for the ratio of variances. We will firstly consider MLE estimation based on the true distribution of the ratio of sample variances, which allows for estimation of an assumed fixed ratio across studies. We will then consider an extension to allow for a simple random effect on the ratio with estimation based on the approximate normal distribution for the log of the ratio of sample variances. Throughout, we will continue to use the notation of the previous sections.

#### 4.1. Maximum likelihood estimator estimation based on the rescaled $F$-distribution

Let $g(f; \nu_1, \nu_2)$ denote the probability density function for the $F$-distribution with degrees of freedom $\nu_1$ and $\nu_2$. Then, under the assumption that the data are sampled from independent normal distributions, it is simple to verify that the probability density function for $S_k$ is
\[
f(s) = \frac{1}{\rho_k} g(s/\rho_k; n_{k1} - 1, n_{k2} - 1) \tag{9}\]

where \(\rho_k = \sigma_{k1}^2/\sigma_{k2}^2\).

It is not difficult to obtain the point and interval MLE estimates for a fixed \(\rho_k = \rho\) \((k = 1, \ldots, K)\) based on the likelihood function previously when using routine optimization functions within a package such as R that includes a computationally computed Hessian matrix.

4.2. Maximum likelihood estimator estimation based on the approximate normal distribution

The simple form of \(T_1\) from Section 3 provides an opportunity to obtain estimates for the ratio of variances based on approximate normality of the log of the ratio of sample variances. For simplicity throughout, let \(\omega = \ln(\rho)\).

4.2.1. Fixed effect model. Using the fact that \(T_1(S_k)\) is approximately \(N(0, 1)\) distributed, we have

\[
\ln(S_k) \overset{approx.}{\sim} N(\mu_k, c_{1k}) \tag{10}\]

where \(\mu_k = \omega + \ln \left\{ \frac{\nu_{2k}}{(\nu_{2k} - 2)} \right\} - c_{1k}/2\), and where \(c_{1k}\) is defined as in \(c_1\) in (2), but specific to the \(k\)th study.

Under the assumption of a common fixed \(\rho = \rho_k\) across all studies, the log-likelihood function obtained from (10) and ignoring constant terms is

\[
l(s_1, \ldots, s_K; \omega) = -\frac{1}{2} \sum_{k=1}^{K} \frac{1}{c_{1k}} \left[ \ln(s_k) - \omega - \ln \left\{ \frac{\nu_{2k}}{(\nu_{2k} - 2)} \right\} + c_{1k}/2 \right]^2. \tag{11}\]

Solving for \(l'(s_1, \ldots, s_K; \ln(\rho)) = 0\) for \(\ln(\rho)\) gives the MLE estimate

\[
\hat{\omega}_{MLE} = \left( \sum_{k=1}^{K} \frac{1}{c_{1k}} \right)^{-1} \sum_{k=1}^{K} \frac{1}{c_{1k}} \left[ \ln(s_k) + \ln \left( \frac{n_{k2} - 3}{n_{k2} - 1} \right) + c_{1k}/2 \right] \tag{12}\]

which is identical to the estimate obtained from (6) when the weights, \(w_k\)s, are chosen to be \(1/\sqrt{c_{1k}}\). A closed-form solution for the variance of the MLE also exists and is given as

\[
V_{MLE}^* = \frac{1}{-l''(s_1, \ldots, s_K; \omega)} = \left( \sum_{k=1}^{K} \frac{1}{c_{1k}} \right)^{-1} \tag{13}\]

which is equivalent to \(V\) from (6) when the weights are chosen to be \(1/\sqrt{c_{1k}}\). This leads to an approximate \((1 - a) \times 100\%\) confidence interval for \(\rho\) as

\[
\left( \exp \left[ \hat{\omega}_{MLE} - z_{1-a/2} \sqrt{V_{MLE}^*} \right], \exp \left[ \hat{\omega}_{MLE} + z_{1-a/2} \sqrt{V_{MLE}^*} \right] \right). \tag{14}\]

4.2.2. Random effects model. As is done for classic random effects models for effect sizes, instead of assuming a fixed \(\rho\) across all studies, we can instead introduce a random effect to explain unmeasurable heterogeneity among the \(\rho_k\)s. From (10), a simple random effect model assumes

\[
\ln(S_k) \overset{approx.}{\sim} N(\mu_k, c_{1k} + \tau^2) \tag{15}\]

where \(\tau^2\) is the variance of an assumed central normal additive random effect for \(\ln(S_k)\). While a closed-form solution for the MLEs of \(\rho\) and \(\tau\) have not been obtained, computationally these are not too difficult to achieve because of the niceties of dealing with the normal distribution.

For \(Y_k = \ln(S_k) \sim N(\mu_k, c_{1k} + \tau^2)\), ignoring constants, a log-likelihood function is

\[
l(y_1, \ldots, y_K; \omega, \tau) = -\frac{1}{2} \sum_{k=1}^{K} \left[ \ln \left( c_{1k} + \tau^2 \right) + \left( \frac{y_k - \mu_k}{c_{1k} + \tau^2} \right)^2 \right]. \tag{16}\]
Taking the first derivative of (15) with respect to both \( \omega \) and \( \tau^2 \), the MLEs for each (denoted \( \hat{\omega}_{MLE} \) and \( \hat{\tau}^2_{MLE} \)) are the solutions to

\[
\frac{\partial l}{\partial \omega} = \sum_{k=1}^{K} \frac{y_k - \mu_k}{c_{k1} + \tau^2} = 0 \quad \text{and} \quad \frac{\partial l}{\partial \tau^2} = \frac{1}{2} \sum_{k=1}^{K} \left[ \frac{(y_k - \mu_k)^2}{(c_{k1} + \tau^2)^2} - \frac{1}{c_{k1} + \tau^2} \right] = 0.
\]

To ensure that the estimate to \( \tau \) is positive, we found that the best approach was to rewrite the likelihood and estimate the parameter \( \ln(\tau) \). Exponentiating was then used to achieve the estimate for \( \tau \).

We can also derive approximate variances for the MLEs (details are in Appendix B),

\[
V_{MLE,\omega} = \left( \sum_{k=1}^{K} \frac{1}{c_{k1} + \hat{\tau}^2_{MLE}} \right)^{-1}, \quad V_{MLE,\tau^2} = 2 \left( \sum_{k=1}^{K} \frac{1}{(c_{k1} + \hat{\tau}^2_{MLE})^2} \right)^{-1}.
\]

When constructing intervals using the appropriate percentile from the normal distribution, our simulation studies showed that the intervals were too narrow. Therefore, our suggested approximate \( (1 - \alpha) \times 100\% \) confidence interval for \( \rho \) is

\[
\left( \exp \left[ \hat{\omega}_{MLE} - t_{K-1,1-\alpha/2} \sqrt{V_{MLE,\omega}} \right], \exp \left[ \hat{\omega}_{MLE} + t_{K-1,1-\alpha/2} \sqrt{V_{MLE,\omega}} \right] \right)
\]

where \( t_{K-1,1-\alpha/2} \) is the \( (1 - \alpha/2) \times 100\% \) percentile from the \( t_{K-1} \) distribution, which results in intervals with improved coverage.

It is not too difficult to implement the estimators in the aforementioned in standard software. For example, using the \texttt{metafor} package [8], estimates for \( \ln(\rho) \) and \( \tau \) can be obtained from the estimated effects and within-study variances using linear mixed effects models. Several estimators of \( \tau \) are available, including the MLE and restricted maximum likelihood estimator (REML, see e.g., [20]), as standard errors and interval estimators for \( \ln(\rho) \), which can be exponentiated as in the aforementioned to achieve intervals for \( \rho \). However, note that \( E[\ln(S_k)] \approx \mu_k \neq \ln(\rho) \) so the adjusted effects

\[
\ln(S_k) - \ln \left( \frac{v_{2k}}{(v_{2k} - 2)} \right) + c_{k1}/2
\]

which has approximate mean \( \ln(\rho) \) should be used to reduce the bias in estimating \( \ln(\rho) \). Similarly to the aforementioned, our simulations revealed lower than nominal coverage for the interval estimators and that using the Knapp and Hartung adjustment [21] resulted in improved coverage.

5. Simulation studies

In this section, we consider several simulation studies, the first of which are focused on the performance of the transformations to achieve approximate standard normality followed by assessment of performance when applied in a meta-analysis setting.

5.1. Simulation 1: meta-analysis for estimation of \( \rho \)

In this section, we consider a meta-analysis for the estimation of the ratio variances (\( \rho \)). We consider both an assumed fixed \( \rho \) across all studies and also a random effect model that allows \( \rho \) to vary across the studies. Simulations are used to obtain approximate values for bias, actual coverage probability and confidence interval width (W) for \( \rho \) as well as the bias of the variance term associated with the random effect for the log of the ratio of variances.

In Table IV, we report the results for data simulated from \( K = 13 \) studies where the sample sizes are chosen to be the same as those for comparison between the BB and Bb/bb groups from Table I. For simplicity, we consider four estimators. Firstly, we consider the interval estimators for \( \rho \) based on the \( T_0 \) and \( T_1 \) transformations (similar results were achieved for the other VSTs considered in the Appendix) with weights according to (8). For \( T_1 \), this is then also identical to the MLE of \( \rho \) based on the log transformation that is reported in Section 4.2.1. The other two methods are the MLE based on the rescaled \( F \)-distribution detailed in Section 4.1 and the estimator based on the random effects model introduced in Section 4.2.2. For the random effects model, we used the \texttt{metafor} R package with the REML of \( \tau \).
Table IV. $K = 13$ studies. Simulated comparisons between $T_0$ and $T_1$, the fixed effects model estimator based on the $F$-distribution (FE), and the random effects model estimator (RE) for bias, coverage probability (CP), and confidence interval width (W) for 1000 simulated runs for various choices of $\rho$. The bias for the estimator of $\tau$ ($\text{Bias}_\tau$) is also included for the RE model. The number of studies and sample sizes chosen are set equal to those for the comparison between the BB and Bb/bb groups from Table I.

| $\tau$ | $\rho$ | Bias | CP  | W   | Bias | CP  | W   | Bias | CP  | W   | Bias | CP  | W   |
|--------|--------|------|-----|-----|------|-----|-----|------|-----|-----|------|-----|-----|
| 0.0    | 0.20   | 0.00 | 0.95| 0.06| 0.00 | 0.95| 0.06| 0.00 | 0.95| 0.06| 0.00 | 0.96| 0.07|
| 0.50   | 0.00   | 0.96 | 0.15| 0.00 | 0.95| 0.15| 0.00 | 0.95| 0.15| 0.00 | 0.96| 0.18|
| 0.80   | 0.00   | 0.95 | 0.24| 0.00 | 0.94| 0.24| 0.00 | 0.94| 0.24| 0.00 | 0.96| 0.29|
| 1.00   | 0.00   | 0.94 | 0.29| 0.00 | 0.94| 0.29| 0.00 | 0.94| 0.29| 0.00 | 0.96| 0.36|
| 1.25   | 0.00   | 0.94 | 0.37| 0.00 | 0.94| 0.37| 0.00 | 0.94| 0.37| 0.00 | 0.95| 0.44|
| 1.50   | 0.01   | 0.94 | 0.45| 0.01 | 0.94| 0.44| 0.01 | 0.94| 0.44| 0.01 | 0.96| 0.54|
| 2.00   | 0.00   | 0.95 | 0.59| -0.00| 0.95| 0.59| -0.00| 0.95| 0.59| -0.00| 0.97| 0.70|
| 5.00   | 0.01   | 0.95 | 1.48| 0.02 | 0.94| 1.48| 0.01 | 0.94| 1.47| -0.00| 0.94| 1.80|
| 0.2    | 0.20   | 0.00 | 0.82 |0.06| 0.00 | 0.82| 0.06| 0.00 | 0.82| 0.06| 0.00 | 0.93| 0.09|
| 0.50   | 0.01   | 0.85 | 0.15| 0.00 | 0.85| 0.15| 0.01 | 0.84| 0.15| 0.00 | 0.94| 0.23|
| 0.80   | 0.01   | 0.83 | 0.24| 0.00 | 0.83| 0.24| 0.01 | 0.82| 0.24| 0.00 | 0.94| 0.37|
| 1.00   | 0.01   | 0.81 | 0.30| 0.01 | 0.82| 0.30| 0.01 | 0.82| 0.30| 0.00 | 0.93| 0.46|
| 1.25   | 0.01   | 0.81 | 0.37| 0.01 | 0.81| 0.37| 0.02 | 0.80| 0.37| 0.00 | 0.92| 0.57|
| 1.50   | 0.02   | 0.84 | 0.45| 0.01 | 0.83| 0.44| 0.03 | 0.84| 0.45| 0.01 | 0.95| 0.68|
| 2.00   | 0.02   | 0.82 | 0.60| 0.01 | 0.82| 0.59| 0.03 | 0.82| 0.60| 0.01 | 0.93| 0.92|
| 5.00   | 0.04   | 0.84 | 1.49| 0.03 | 0.83| 1.48| 0.07 | 0.84| 1.50| 0.01 | 0.93| 2.30|
| 0.4    | 0.20   | 0.01 | 0.62 |0.06| 0.00 | 0.62| 0.06| 0.01 | 0.60| 0.06| 0.00 | 0.94| 0.13|
| 0.50   | 0.02   | 0.62 | 0.15| 0.01 | 0.62| 0.15| 0.03 | 0.60| 0.16| 0.01 | 0.94| 0.34|
| 0.80   | 0.02   | 0.62 | 0.24| 0.01 | 0.61| 0.24| 0.04 | 0.62| 0.25| 0.01 | 0.94| 0.52|
| 1.00   | 0.03   | 0.61 | 0.30| 0.01 | 0.61| 0.30| 0.05 | 0.59| 0.31| 0.00 | 0.93| 0.67|
| 1.25   | 0.04   | 0.59 | 0.38| 0.02 | 0.59| 0.37| 0.08 | 0.60| 0.39| 0.01 | 0.94| 0.84|
| 1.50   | 0.05   | 0.62 | 0.46| 0.03 | 0.62| 0.45| 0.09 | 0.59| 0.47| 0.01 | 0.95| 1.00|
| 2.00   | 0.06   | 0.62 | 0.61| 0.03 | 0.63| 0.60| 0.12 | 0.61| 0.63| 0.01 | 0.95| 1.37|
| 5.00   | 0.19   | 0.61 | 1.53| 0.11 | 0.61| 1.50| 0.33 | 0.61| 1.58| 0.07 | 0.94| 3.36|
Table V. $K = 26$ studies. Simulated comparisons between $T_0$, $T_1$, the fixed effects model estimator based on the $F$-distribution (FE) and the random effects model estimator (RE) for bias, coverage probability (CP) and confidence interval width (W) for 1000 simulated runs for various choices of $\rho$. The bias for the estimator of $\tau$ (Bias$_{\tau}$) is also included for the RE model. The number of studies is twice that used in Table IV where sample sizes have been repeated twice.

| $\tau$ | $\rho$ | $T_0$ Bias | CP | W | $T_1$ Bias | CP | W | FE Bias | CP | W | RE Bias | CP | W | Bias$_{\tau}$ |
|--------|-------|-----------|----|---|------------|----|---|--------|----|---|--------|----|---|-----------|
| 0.0    | 0.20  | 0.00      | 0.94 | 0.04 | 0.00      | 0.95 | 0.04 | 0.00      | 0.94 | 0.04 | 0.00      | 0.95 | 0.05 | 0.05      |
| 0.50   | 0.00  | 0.95      | 0.10 | 0.00 | 0.95      | 0.10 | 0.00 | 0.96      | 0.10 | 0.00 | 0.96      | 0.12 | 0.00 | 0.95      |
| 0.80   | -0.00 | 0.94      | 0.17 | 0.00 | -0.00     | 0.94 | 0.17 | -0.00     | 0.94 | 0.17 | -0.00     | 0.95 | 0.23 | 0.05      |
| 1.00   | 0.00  | 0.95      | 0.21 | 0.00 | 0.95      | 0.21 | 0.00 | 0.95      | 0.21 | 0.00 | 0.95      | 0.23 | 0.00 | 0.96      |
| 1.25   | 0.00  | 0.95      | 0.26 | 0.00 | 0.95      | 0.26 | 0.00 | 0.95      | 0.26 | 0.00 | 0.95      | 0.26 | 0.00 | 0.95      |
| 1.50   | 0.01  | 0.95      | 0.31 | 0.01 | 0.95      | 0.31 | 0.00 | 0.95      | 0.31 | 0.00 | 0.95      | 0.31 | 0.00 | 0.95      |
| 2.00   | -0.00 | 0.96      | 0.42 | -0.00 | 0.96     | 0.41 | -0.00 | 0.95      | 0.41 | -0.00 | 0.95      | 0.47 | 0.00 | 0.95      |
| 5.00   | -0.01 | 0.94      | 1.04 | -0.01 | 0.94     | 1.04 | -0.01 | 0.95      | 1.03 | -0.01 | 0.95      | 1.18 | 0.05 | 0.95      |
| 0.2    | 0.20  | 0.00      | 0.83 | 0.04 | -0.00     | 0.83 | 0.04 | 0.00      | 0.83 | 0.04 | -0.00     | 0.85 | 0.06 | -0.02     |
| 0.50   | 0.00  | 0.81      | 0.10 | 0.00 | 0.81      | 0.10 | 0.01 | 0.81      | 0.11 | 0.00 | 0.81      | 0.15 | 0.00 | -0.01     |
| 0.80   | 0.01  | 0.82      | 0.17 | 0.00 | 0.82      | 0.17 | 0.01 | 0.81      | 0.17 | 0.00 | 0.81      | 0.25 | 0.00 | -0.01     |
| 1.00   | 0.01  | 0.83      | 0.21 | 0.00 | 0.84      | 0.21 | 0.01 | 0.83      | 0.21 | 0.00 | 0.83      | 0.31 | 0.00 | -0.01     |
| 1.25   | 0.01  | 0.83      | 0.26 | 0.00 | 0.82      | 0.26 | 0.02 | 0.83      | 0.26 | 0.00 | 0.83      | 0.38 | 0.00 | -0.02     |
| 1.50   | 0.01  | 0.84      | 0.32 | 0.02 | 0.83      | 0.31 | 0.02 | 0.82      | 0.32 | 0.01 | 0.84      | 0.46 | 0.00 | -0.01     |
| 2.00   | 0.01  | 0.85      | 0.42 | 0.02 | 0.84      | 0.42 | 0.03 | 0.84      | 0.42 | 0.00 | 0.85      | 0.62 | 0.00 | -0.01     |
| 5.00   | 0.05  | 0.82      | 1.05 | 0.09 | 0.82      | 1.05 | 0.09 | 0.93      | 1.53 | 0.05 | 0.93      | 1.53 | 0.02 | -0.02     |
| 0.4    | 0.20  | 0.00      | 0.63 | 0.04 | 0.00      | 0.62 | 0.04 | 0.00      | 0.59 | 0.04 | 0.00      | 0.62 | 0.09 | -0.01     |
| 0.50   | 0.01  | 0.62      | 0.11 | 0.00 | 0.63      | 0.10 | 0.02 | 0.58      | 0.11 | 0.00 | 0.62      | 0.22 | 0.00 | -0.01     |
| 0.80   | 0.01  | 0.62      | 0.17 | -0.00 | 0.62     | 0.16 | 0.03 | 0.58      | 0.21 | -0.00 | 0.62      | 0.22 | 0.00 | -0.01     |
| 1.00   | 0.03  | 0.61      | 0.21 | 0.01 | 0.61      | 0.21 | 0.06 | 0.58      | 0.22 | 0.00 | 0.58      | 0.45 | 0.00 | -0.00     |
| 1.25   | 0.02  | 0.60      | 0.26 | 0.00 | 0.61      | 0.26 | 0.06 | 0.58      | 0.28 | 0.00 | 0.58      | 0.57 | 0.00 | -0.00     |
| 1.50   | 0.03  | 0.59      | 0.32 | 0.01 | 0.60      | 0.31 | 0.08 | 0.58      | 0.33 | 0.00 | 0.58      | 0.67 | 0.00 | -0.01     |
| 2.00   | 0.04  | 0.59      | 0.42 | 0.01 | 0.60      | 0.42 | 0.10 | 0.57      | 0.44 | 0.01 | 0.58      | 0.89 | 0.00 | -0.01     |
| 5.00   | 0.11  | 0.63      | 1.06 | 0.27 | 0.59      | 1.10 | 0.02 | 0.96      | 2.26 | -0.00 | -0.01     |
(because the coverage was slightly better than when using the MLE) and the Knapp and Hartung adjustment. Good results were also achieved for the MLE in Section 4.2.2. Throughout, we will refer to these methods as $T_0$, $T_1$, $FE$, and $RE$ respectively.

For each of the methods described previously, we consider three choices for the variance parameter ($\tau^2$) for the random effects model. The first choice, $\tau^2 = 0$, equates to the fixed effects model for which $T_0$, $T_1$, and $FE$ are well suited where $\tau$ is not estimated. The other choices are for $\tau^2 = 0.2^2$ and $\tau^2 = 0.4^2$, which is a scenario favoring the RE approach. The data are simulated according to the assumed data model (not the approximate normal model) in that the ratio of sample variances are sampled from the rescaled $F$-distribution.

When $\tau = 0$, all approaches achieve very close to nominal coverage of 0.95, small bias in estimating $\rho$ and similar interval widths. The RE estimator of $\tau^2$ has an approximate bias of around 0.07 for $\tau$, which is somewhat expected given that the minimum for the estimate of $\tau$ is 0. The result is a very slightly conservative interval for $\rho$. In the presence of a random effect (i.e., when $\tau = 0.2$ or 0.4 in this simulation), we observe much lower than nominal coverages for $T_0$, $T_1$, and $FE$. While the bias in estimating $\rho$ is often small, these estimators assume that $\tau = 0$, which results in a smaller variance for the estimators and subsequently intervals that are too narrow. The FE method performs typically worse with respect to the bias in estimator $\rho$, indicating that the transformations can adequately isolate the component $\rho$ from the missing random effect. Close to nominal coverage is achieved when using the RE estimator although $\tau$ tends to be slightly underestimated.

We now repeat the simulation from Table IV, but this time we assume twice as many studies (i.e., $K = 26$) and report the results in Table V. To achieve this, the $K = 13$ sample sizes were repeated. For $\tau = 0$, we observe excellent coverage for the interval estimators of $\tau$ and where the intervals are, on average, narrower. When $\tau > 0$, poor coverage is again observed for all methods except for the RE estimator where excellent coverage is obtained. The bias in estimating $\tau$ has also decreased.

Finally, in Table VI, we repeat the simulation again for 26 studies, but this time we double all of the sample sizes. We see the same patterns of coverage as before but with typically narrower intervals. $T_1$ does a good job at estimating $\rho$ in the presence of a random effect with small bias reported although the interval coverage is poor. The RE method has again achieved very good results, and the bias in estimating $\tau$ has decreased.

5.2. Simulation 2: testing for unequal variances.

We now briefly consider the test of $H_0 : \rho_k = 1$ for all $k = 1, \ldots, K$ using the $X^2_w$ and likelihood ratio test from Section 3.3.

In Table VII are shown simulated values for the power of rejecting equal variances for several choices of $\rho$ and $\tau$. The sample sizes chosen were those from Table IV for the comparison between the BB and Bb/bb groups. For $K = 26$, the 13 sample sizes were repeated twice and then also doubled for $K = 26$ with larger sample sizes. When the null is true (i.e., when $\rho = 1$ and $\tau = 0$), the empirical size of the test based on the $X^2_w$ is, as expected, approximately equal to the nominal size of 5%. While for the LRT, the empirical size is a little higher than nominal when $K = 13$, it is closer to nominal when $K = 26$, and in particular when the sample sizes within the studies is larger. The LRT has greater power of rejecting $H_0$ when $\tau = 0$. However, when $\tau = 0.2$ and $\tau = 0.4$, it is the test based on $\bar{X}^2_w$ that is more likely to find evidence against $H_0$. We also note that both tests have a good chance of rejecting $H_0$ when $\tau$ is not too close to zero and even when $\rho = 1$. That is, even when the variances are equal on average, the tests are still capable of detecting the presence of a random effect.

6. Bone mass density example continued

We now illustrate how the theory and simulations previously inform and affect an analysis of the bone mineral density data from Section 2 when comparing the data from the BB and Bb/bb groups from Table I. For simplicity, we focus only on the transformation $T_0$.

In Plot A of Figure 2, we provide the forest plot for the meta-analysis of the ratio of variances. The study specific intervals are those arising from the standard $F$-test, and the estimates are simply the ratio of estimated variances. The estimate and interval labeled ‘Overall (MLE)’ and ‘Overall (REML)’ have arisen from the meta-analysis based on estimation of the random effects model (Section 4.2.2) using MLE and REML for $\tau$, respectively. As can be seen from the forest plot, there appears to be a persistent suggestion that the true ratio is greater than one with two of the studies with the smaller estimates having...
Table VI. Large sample, $K = 26$ studies. Simulated comparisons between $T_0$, $T_1$, the fixed effects model estimator based on the $F$-distribution (FE) and the random effects model estimator (RE) for bias, coverage probability (CP) and confidence interval width (W) for 1000 simulated runs for various choices of $\rho$. The number of studies is twice that used in Table IV where sample sizes have been repeated twice and have been doubled.

| $\tau$ | $\rho$ | $T_0$ Bias | $T_0$ CP | $T_0$ W | $T_1$ Bias | $T_1$ CP | $T_1$ W | FE Bias | FE CP | FE W | RE Bias | RE CP | RE W | Bias$\tau$ |
|--------|--------|------------|----------|--------|------------|----------|--------|--------|------|------|--------|------|------|-----------|
| 0.0    | -0.00  | 0.95      | 0.03     | 0.02   | 0.00       | 0.95     | 0.03   | -0.00  | 0.95 | 0.03 | 0.00   | 0.95 | 0.03 | 0.03      |
|        | -0.00  | 0.94      | 0.07     | 0.05   | -0.00      | 0.95     | 0.07   | -0.00  | 0.95 | 0.07 | -0.00  | 0.96 | 0.08 | 0.03      |
|        | 0.00   | 0.95      | 0.12     | 0.05   | 0.00       | 0.96     | 0.12   | 0.00   | 0.95 | 0.12 | 0.00   | 0.96 | 0.13 | 0.04      |
|        | -0.00  | 0.95      | 0.14     | 0.09   | -0.00      | 0.95     | 0.14   | -0.00  | 0.95 | 0.14 | -0.00  | 0.96 | 0.15 | 0.03      |
|        | 0.00   | 0.96      | 0.18     | 0.11   | 0.00       | 0.96     | 0.18   | 0.00   | 0.96 | 0.18 | 0.00   | 0.96 | 0.20 | 0.04      |
|        | -0.00  | 0.95      | 0.22     | 0.14   | -0.00      | 0.95     | 0.22   | -0.00  | 0.95 | 0.22 | -0.00  | 0.96 | 0.24 | 0.04      |
|        | 0.00   | 0.95      | 0.29     | 0.18   | 0.00       | 0.95     | 0.29   | 0.00   | 0.95 | 0.29 | 0.00   | 0.96 | 0.32 | 0.04      |
|        | 0.00   | 0.96      | 0.73     | 0.46   | 0.00       | 0.96     | 0.73   | 0.00   | 0.95 | 0.72 | 0.00   | 0.96 | 0.79 | 0.03      |
| 0.2    | 0.00   | 0.75      | 0.03     | 0.02   | 0.00       | 0.75     | 0.03   | 0.00   | 0.75 | 0.03 | 0.00   | 0.94 | 0.05 | -0.01     |
|        | 0.00   | 0.73      | 0.07     | 0.05   | -0.00      | 0.73     | 0.07   | 0.00   | 0.73 | 0.07 | -0.00  | 0.94 | 0.12 | -0.01     |
|        | 0.01   | 0.73      | 0.12     | 0.07   | 0.00       | 0.73     | 0.12   | 0.01   | 0.72 | 0.12 | 0.00   | 0.93 | 0.19 | -0.01     |
|        | 0.01   | 0.71      | 0.15     | 0.09   | 0.00       | 0.71     | 0.15   | 0.01   | 0.71 | 0.15 | 0.00   | 0.91 | 0.24 | -0.01     |
|        | 0.01   | 0.76      | 0.18     | 0.11   | 0.00       | 0.76     | 0.18   | 0.02   | 0.76 | 0.18 | 0.00   | 0.94 | 0.31 | -0.01     |
|        | 0.01   | 0.74      | 0.22     | 0.14   | 0.00       | 0.74     | 0.22   | 0.02   | 0.74 | 0.22 | -0.00  | 0.94 | 0.37 | -0.01     |
|        | 0.01   | 0.74      | 0.29     | 0.18   | 0.00       | 0.74     | 0.29   | 0.03   | 0.73 | 0.29 | 0.00   | 0.96 | 0.49 | -0.01     |
|        | 0.02   | 0.74      | 0.73     | 0.46   | 0.01       | 0.74     | 0.73   | 0.06   | 0.73 | 0.74 | -0.00  | 0.94 | 1.22 | -0.01     |
| 0.4    | 0.01   | 0.45      | 0.03     | 0.02   | 0.00       | 0.48     | 0.03   | 0.01   | 0.46 | 0.03 | 0.00   | 0.95 | 0.08 | -0.00     |
|        | 0.02   | 0.43      | 0.07     | 0.05   | 0.00       | 0.47     | 0.07   | 0.03   | 0.45 | 0.08 | 0.00   | 0.94 | 0.20 | -0.00     |
|        | 0.03   | 0.43      | 0.11     | 0.07   | 0.01       | 0.45     | 0.12   | 0.05   | 0.42 | 0.12 | 0.00   | 0.94 | 0.31 | -0.01     |
|        | 0.03   | 0.47      | 0.14     | 0.09   | 0.01       | 0.51     | 0.15   | 0.05   | 0.47 | 0.15 | 0.00   | 0.95 | 0.39 | -0.01     |
|        | 0.05   | 0.42      | 0.18     | 0.12   | 0.01       | 0.45     | 0.18   | 0.07   | 0.41 | 0.19 | 0.01   | 0.92 | 0.48 | -0.01     |
|        | 0.04   | 0.47      | 0.21     | 0.14   | 0.01       | 0.50     | 0.22   | 0.08   | 0.46 | 0.23 | 0.01   | 0.95 | 0.59 | -0.00     |
|        | 0.07   | 0.47      | 0.27     | 0.18   | 0.01       | 0.50     | 0.29   | 0.11   | 0.47 | 0.31 | 0.01   | 0.95 | 0.79 | -0.00     |
|        | 0.14   | 0.47      | 0.70     | 0.46   | 0.03       | 0.49     | 0.73   | 0.27   | 0.47 | 0.77 | 0.02   | 0.94 | 1.96 | -0.01     |
Table VII. Proportion of times that equal variances was rejected based on 10,000 simulation runs and for varying \( \rho \) and \( \tau \). For \( K = 13 \), the samples from Table IV were used for the comparison between the BB and Bb/bb groups. For \( K = 26 \), these sample sizes were repeated twice and \( K = 26 \) (large) the sample sizes were also doubled.

| \( \tau \) | \( K = 13 \) | \( K = 26 \) (larger samples) |
|---|---|---|
| | \( \rho \) | \( X^2_w \) | LRT | \( X^2_w \) | LRT | \( X^2_w \) | LRT |
| 1 | 0.505 | 0.065 | 0.048 | 0.056 | 0.052 | 0.054 |
| 1.2 | 0.417 | 0.507 | 0.619 | 0.792 | 0.911 | 0.991 |
| 1.5 | 0.994 | 0.995 | 0.999 | 1.000 | 1.000 | 1.000 |
| 0.2 | 1.2 | 0.686 | 0.609 | 0.870 | 0.838 | 0.990 | 0.987 |
| 1.5 | 0.979 | 0.990 | 1.000 | 1.000 | 1.000 | 1.000 |
| 0.4 | 1.2 | 0.882 | 0.749 | 0.984 | 0.947 | 0.999 | 0.997 |
| 1.5 | 0.989 | 0.987 | 1.000 | 1.000 | 1.000 | 1.000 |

LRT, likelihood ratio test.

Figure 2. Results of meta-analysis for the BB versus Bb/bb data in Table I. Plot A depicts the forest plot for the analysis where the measure of interest is the ratio of variances. The meta-analytic interval and estimate for a \( \rho \) is given in the last row where the Maximum likelihood estimator assuming the random effects model was used. Plot B displays meta-analysis \( p \)-values for the test for equal variances when studies are added incrementally according to the magnitude of the \(|Z_k|\)s.

As seen in the forest plot, it is possible that this highly significant result is due to the very large estimated ratio of variances from Study 9. In Plot B, we explore this by iteratively adding studies based on the magnitude of their transformed statistics. The \( p \)-values are those based on (4) with weights chosen as high estimator variability and many of the studies with larger estimates having more precision. This is verified by the overall large estimates of 1.36 and 1.34 with corresponding 95% confidence intervals of (1.03, 1.80) and (1.00, 1.79) obtained using the Knapp and Hartung adjustments. Without the adjustments, the intervals were (1.12, 1.65) and (1.08, 1.65), respectively. The MLE and REML estimates for \( \tau^2 \) were \( \hat{\tau}^2 = 0.033 \) and 0.051. When the MLE fixed effect model was assumed, the estimated \( \rho \) was 1.45 with approximate 95% confidence interval (1.26, 1.68). These results should lead to a rejection of using SMDs for the analysis given that the required assumption of equal variances in general is violated. The forest plot is simple to create using the R package `metafor` [8] because it includes functionality for the creation of forest plots that only requires the point and interval estimates.
1/\sqrt{c_{k1}}. The p-value in Plot B for \( k^* = 1 \) is associated with the test based only on the value from study 11 (the study with the smallest \( Z_k \)) and, not surprisingly, it is not significant. For \( k^* = 2 \), the test is run again but this time with both studies 11 and 1 (the two studies with the smallest in magnitude \( Z_k \)’s) and so on. Here, the weights used are calculated based only on the included studies. We can see that as studies with larger \( |Z_k| \) are included, the p-value for the test of equal variances decreases rapidly because the larger studies also have larger \( |Z_k| \)’s. We can see that there is very strong evidence for unequal variances even if Study 9 was ignored. This means it would not simply suffice to remove Study 9 if SMDs were to be used because there is further evidence of a violation of equal variances.

7. Discussion

For a meta-analysis comparing means of two independent groups, a common approach is to consider SMDs, which requires that the population variances within each of the groups within each study are equal. However, when the population variances are not equal, the validity of the meta-analysis should be queried, and the interpretation of the meta-estimated effect may be unreliable. In many cases it appears that assuming equal variances has been for convenience only, and justifications for making this assumption are rarely found. In this paper, we have shown that the meta-analysis of ratio of variances can be used to assess the validity of the equal variances assumption and consequently provide researchers with a justification to move towards other, perhaps less common, estimated effects. Simple Q–Q plots of transformed ratios of sample variances allow visual inspection to determine departures from this assumption. Alternatively, we also introduced tests overall equality of variances and meta-estimates of the ratio of variances in both the fixed effect and the random effect settings. Our recommendation is that one or more of these approaches to querying the equal variances assumption should always be taken and when the assumption is not justified then a different approach, such as using log ratios of means or Glass’ effect, should be employed. In disciplines within which the use of the Cohen’s \( d \) and its variants are commonly used, our methods provide researchers with a very clear approach to justifying other methods. The methods in this paper are simple to implement and R code, including an example using the package metafor, can be provided by the authors upon request.

Appendix A: technical details and further results for the variance stabilizing transformations

In this section of the Appendix, we provided some technical details for \( T_1 \) and introduce three further VSTs.

A.1. Technical details for \( T_1 \)

The technical details for VSTs \( T_1 \) and \( T_2 \) are provided here. The mean and variance of \( S \sim F_{v_1,v_2} \) are (see, for e.g., p.326 of [1])

\[
E[S] = \frac{v_2}{v_2 - 2} \text{ for } v_2 > 2
\]

\[
\text{Var}[S] = \frac{2v_1^2(v_1 + v_2 - 2)}{v_1(v_2 - 2)^2(v_2 - 4)} \text{ for } v_2 > 4.
\]

The variance of \( S \) can be written as function of its mean \( \text{Var}[S] = g(E[S]) \) in a number of ways, including:

\[
g_1(t) = c_1 t^2 \quad \text{where} \quad c_1 = \frac{2(v_1 + v_2 - 2)}{v_1(v_2 - 4)}
\]

\[
g_2(t) = c_2 t \left( t + \frac{v_2}{v_1} \right) \quad \text{where} \quad c_2 = \frac{2}{v_2 - 4}.
\]

The first expression leads (see page 32 of [22]) to the respective VST \( h_1(x) = \ln(x)/\sqrt{c_1} \). Using the approximation given as \( E[h(S)] = h(E[S]) + h''(E[S]) \text{Var}[S]/2 \), we obtain
The statistic given as 

\[ T \]

is closerto zero. However, for this transformation, the adjustment is very small and our simulations reveal it is not necessary. In fact, this transformation, as it is, is quite remarkable in that when compared with 

\[ \frac{2\ln{\left( \frac{\sqrt{S} + \sqrt{S + \nu_2/\nu_1}}{\sqrt{E[S]} + \sqrt{E[S] + \nu_2/\nu_1}} \right)}}{4\sqrt{\nu_2} (2\nu_1 + \nu_2 - 2)} \]

which leads to (2) after recentering to achieve approximate mean zero.

A.2. Three further transformations

We now consider three further competing VSTs of 

\[ S \sim F_{\nu_1,\nu_2} \]

Firstly, consider 

\[ T_2(S) = \frac{2}{\sqrt{\nu_2}} \ln \left( \frac{\sqrt{S} + \sqrt{S + \nu_2/\nu_1}}{\sqrt{E[S]} + \sqrt{E[S] + \nu_2/\nu_1}} \right) + \frac{\sqrt{\nu_2} (2\nu_1 + \nu_2 - 2)}{4\sqrt{\nu_1^2 + 2\nu_1\nu_2 - 2\nu_1}} \]

where \( c_2 = 2/(\nu_2 - 4) \) and \( E[S] = \nu_2/(\nu_2 - 2) \) are the mean for the \( F_{\nu_1,\nu_2} \) distribution for \( \nu_2 > 2 \). This transformation follows from \( g_2(t) \) in Appendix A.1, which leads to \( h_2(x) = 2 \ln \left\{ 2 \left( \sqrt{\frac{x}{\nu_2}} + \sqrt{\frac{1}{\nu_2} + \nu_2/\nu_1} \right) \right\} / \sqrt{c_2} \) and then using a similar approach to establishing \( T_1 \). This VST also has approximate mean zero and variance one.

The next transformation that we consider is Paulson’s normalizing transformation \([23]\) of the \( F_{\nu_1,\nu_2} \) statistic given as 

\[ T_3(S) = \left\{ \left( 1 - \frac{2}{9\nu_2} \right)^{S^{1/3}} - \left( 1 - \frac{2}{9\nu_1} \right)^{S^{1/3}} \right\} \left\{ \left( \frac{2S_{\nu_2}^3}{9\nu_1} + \frac{2}{9\nu_2} \right) \right\}^{-1/2}. \]

As was the case with \( T_1 \) and \( T_2 \), a recentering of \( T_3 \) could also be carried out so that its expected value is closer to zero. However, for this transformation, the adjustment is very small and our simulations reveal it is not necessary. In fact, this transformation, as it is, is quite remarkable in that when compared with \( T_0(S) = \Phi^{-1} \left[ F(S) \right] \), the results are almost identical so that \( T_3 \) provides a simple alternative. For example, when simulating 1000 observations from the \( F_{\nu_1,\nu_2} \) distribution for all choices of \( \nu_1 \) and \( \nu_2 \) in 5, …, 100, the minimum correlation between the transformed data using each of the transformations was 0.99969.

The final transformation we now describe a slight variation of the VST introduced in Chapter 23 of Kulinskaya et al. \([10]\). While it is more complicated than the other transformations, an advantage it has over \( T_1 \) and \( T_2 \) is that it exists for all degrees of freedom. It is also more applicable in general than the other transformations because it can be applied to non-central \( F \)-distributed test statistics. Consider a test statistic \( S_\lambda \) for which \( S_\lambda \sim F_{\nu_1,\nu_2}(\lambda) \) where \( \lambda \) is the non-centrality parameter. Let \( \Delta = F_{\nu_1,\nu_2}^{-1}(0.5) \) where \( F_{\nu_1,\nu_2}^{-1} \) denotes the inverse cumulative distribution function so that \( \Delta \) is the median of the \( F_{\nu_1,\nu_2} \) (i.e., with \( \lambda = 0 \); the central \( F \)) distribution. Let 

\[ S'_{\lambda} = \left\{ \begin{array}{ll} S_\lambda & \text{if } S_\lambda > \Delta \\ F_{\nu_1,\nu_2}^{-1} \left[ 1 - F_{\nu_1,\nu_2}(S_\lambda) \right], & \text{if } S_\lambda \leq \Delta \end{array} \right. \]

where \( F_{\nu_1,\nu_2} \) is the cumulative distribution function of \( S_\lambda \).

For \( \text{sign}(S_\lambda; \Delta) = 1 \) if \( S_\lambda > \Delta \) and \( -1 \) otherwise, the transformation function is 

\[ T_4(S_\lambda; \nu_1, \nu_2) = \left( \frac{\nu_2}{\nu_2 + 1} \right) \text{sign}(S_\lambda; \Delta) \times \sqrt{\frac{\nu_2}{2}} \cosh^{-1} \left( \frac{\nu_1 S'_\lambda + \nu_2}{\sqrt{\nu_1^2 \Delta + \nu_2^2}} \right) \]

\[ -\cosh^{-1} \left( \sqrt{\nu_1 \Delta/\nu_2 + 1} \right) \]

This VST is that of Kulinskaya et al. \([10, \text{p.199}]\) but multiplied by \( \nu_2/(\nu_2 + 1) \). We have used this adjustment because our extensive simulations reveal that it improves performance when \( \nu_2 \) is small. Kulinskaya et al. \([10]\) show that \( h(S; \nu_1, \nu_2) \) is approximately normally distributed with variance one even when the degrees of freedom are small and that it has mean 0 under the null and mean increasing in \( \lambda \) under alternatives. They also provide applications for the ANOVA \( F \)-tests.
Figure A.1. Contour plots for the empirical size of the test following variance stabilization when the nominal size is $\alpha = 0.05$. Plots A, B, C, and D are for the transformations in (2), (A.1), (A.2), and (A.4), respectively. Ten thousand iterations were used in the simulation for each combination of degrees of freedom.

While we have referred to these transformations as VSTs, this is strictly only true in general for $T_1$ for any $\rho$. The other transformations are VSTs for $\rho$ near one although the transformations are still useful in other contexts.

A.3. Some simulated comparisons for the variance stabilizing transformations

In Section 3 we noted that a correction to the $T_4$ transformation function can be useful to improve achieving approximate standard normality. This was discovered by considering contour plots of the size of the tests (tail probabilities) from Section 3.3 under the assumption of equal variances. We provide some examples of these plots here for $T_1$, $T_2$, $T_3$, and $T_4$ in Figure A.1.

The simulated size of the test is depicted, which is the proportion of times that $X^2$ exceeded the cutoff 1.96$^2$ for a nominal size of $\alpha = 5\%$ where $X^2$ is from (3) but for a single study. That is, it is simply the squared VST transformed test statistics using the VSTs $T_1$ (Plot A) through to $T_4$ (Plot D). While all the transformations perform well across most of the degrees of freedom combinations (because the empirical size is close to nominal), the Paulson transformation $T_3$ is quite remarkable in the sense that it consistently results in an excellent empirical size across all of the choices. As noted previously, $T_3(S) = \Phi^{-1}[F(S)]$ for $S \sim F_{v_1,v_2}$. The white space in Plots A and B are for when the transformations are not defined for $T_1$ and $T_2$, which occurs when $v_2 \leq 4$. VST $T_4$ is defined across all degrees of freedom, and therefore does not suffer from this same problem, but for consistency $T_3$ is the overall best performer when it comes to size. It should be pointed out that when the degrees of
freedom are small, transformations such as \( T_1 \) are sensitive to which group is used in the denominator and which is used in the numerator for the ratio. So from a purely testing point of view, \( T_3 \) is the safest option. However, this is for a single study only and in a meta-analysis, the problem is likely to only be problematic if there is a persistent small sample size in one group across several studies.

Similar results were also found when we considered sizes of \( \alpha = 0.01 \) and \( \alpha = 0.1 \) (not shown). While all of the VSTs performed well, it was the Paulson transformation that consistently provided excellent results across all combinations for the degrees of freedom.

We now consider the performance of the transformation function in achieving standard normality over varying \( \nu_1 \) and \( \nu_2 \).

In Figure A.2, we provide histograms of 10,000 observations generated from the \( F_{\nu_1,\nu_2} \) distribution after application of each of VSTs \( T_1, T_2, T_3, \) and \( T_4 \). Some combinations of \( \nu_1 \) and \( \nu_2 \) are from \( \{15, 30\} \), and the black line depicts the standard normal density function. All VSTs achieve at least approximate normality for these combinations of degrees of freedom. However, it is the \( T_3 \) and \( T_4 \) VSTs that are the best performers followed by \( T_1 \) provided both degrees of freedom are not too small. We also tried smaller degrees of freedom values (e.g., combinations from 10 to 20; not shown) and the performance of \( T_2 \) diminished the most. \( T_1 \) still did a reasonable job of normalizing the data, and \( T_3 \) and \( T_4 \) performed exceptionally well. When increasing the degrees of freedom, all methods did very well in normalizing \( F \)-statistics.
Appendix B: variances for the random effects model maximum likelihood estimators

For simplicity, when needed, let $v = \tau^2$. Using the notation from Section 4.2.2,

\[
\frac{\partial^2 I}{\partial \omega^2} = -\sum_{k=1}^{K} \frac{1}{c_{1k} + \tau^2},
\]

\[
\frac{\partial^2 I}{\partial v^2} = \sum_{k=1}^{K} \left[ \frac{1}{2} \cdot \frac{1}{(c_{1k} + \tau^2)^2} - \frac{(y_k - \mu_k)^2}{(c_{1k} + \tau^2)^3} \right],
\]

\[
\frac{\partial^2 I}{\partial \omega \partial v} = -\frac{1}{2} \sum_{k=1}^{K} \frac{y_k - \mu_k}{(c_{1k} + \tau^2)^2}.
\]

Taking the expectation with respect to random $\ln(S_1), \ldots, \ln(S_K)$ gives $E[\partial^2 I/(\partial v^2)] = - (1/2) \sum_{k=1}^{K} (c_{1k} + \tau^2)^{-1}$ and $E[\partial^2 I/(\partial \omega \partial v)] = 0$. By inverting the negative of the information matrix, we then have approximate variances for the MLEs.

Acknowledgements

The authors are very thankful to an Associated Editor and two referees for providing many excellent suggestions that lead to a much improved manuscript. The authors give thanks to the Editor Professor Louise Ryan, for allowing a resubmission after an extensive revision of the paper and to Professor Elena Kulinskaya for her feedback on an earlier draft.

References

1. Johnson N, Kotz S, Balakrishnan N. Continuous Univariate Distributions, Vol. 2. John Wiley & Sons: New York, 1995.
2. Cohen J. Statistical Power Analysis for the Behavioral Sciences (2nd edn). Lawrence Erlbaum Associates: New Jersey, 1988.
3. Thakkinian A, D’Este C, Eisman J, Nguyen T, Attia J. Meta-analysis of molecular association studies: vitamin D receptor gene polymorphisms and BMD as a case study. *Journal of Bone and Mineral Research* 2004; 19:419–428.
4. Borenstein M, Hedges LV, Gurevitch J, Ruston I. The arcsine difference as a measure of treatment effect in meta-analysis with zero cells. *Statistics in Medicine* 2009; 28:721–738.
5. Kulinskaya E, Morgenthaler S, Staudte RG. Meta Analysis: A Guide to Calibrating and Combining Statistical Evidence. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd.: Chichester, 2008.
6. Hedges LV. Conducting meta-analyses in R with the *metafor* package. *Journal of Statistical Software* 2010; 36(3):1–48.
7. Hedges L, Olkin I. *Statistical Methods for Meta-Analysis*. Academic Press: Orlando, 1981.
8. Glass G, McGaw B, Smith ML. *Meta-Analysis in Social Research*. Sage Publications: Beverly Hills, California, 1981.
9. Viechtbauer W. Using the notation from Section 4.2.2, gives $E[\partial^2 I/(\partial \omega \partial v)] = 0$.
10. Kulinskaya E, Morgenthaler S, Staudte RG. Variance stabilizing the difference of two binomial proportions. *American Statistician* 2010; 64(4):350–356.
11. Prendergast LA, Staudte RG. Better than you think: interval estimators of the difference of binomial proportions. *Journal of Statistical Planning and Inference* 2014; 148:38–48.
12. Prendergast LA, Staudte RG. Transforming the model t: random effects meta-analysis with stable weights. *Statistics in Medicine* 2013; 32(11):1842–1864.
13. Jackson D, Bowden J, Baker R. Approximate confidence intervals for moment-based estimators of the between-study variance in random effects meta-analysis. *Research Synthesis Methods* 2015. DOI: 10.1002/jrsm.1162.
14. Yuan KH, Bentler PM. Two simple approximations to the distributions of quadratic forms. *British Journal of Mathematical and Statistical Psychology* 2010; 63(2):273–291.
20. Viechtbauer W. Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics* 2005; 30(3):261–293.

21. Knapp G, Hartung J. Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine* 2003; 22:2693–2710.

22. Bickel PJ, Doksum KA. *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden–Day: San Francisco, 1977.

23. Paulson F. An approximate normalization of the analysis of variance distribution. *Annals of Mathematical Statistics* 1942; 13:233–235.