Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression

Corresponding author:
Elizabeth Tipton
Assistant Professor of Applied Statistics
Department of Human Development
Teachers College, Columbia University
525 W. 120th St., Box 118
New York, NY 10027
(212) 678-3844
tipton@tc.columbia.edu

James E. Pustejovsky
Assistant Professor
Educational Psychology Department
University of Texas at Austin
1 University Station D5800
Austin, TX 78712
pusto@austin.utexas.edu

ELIZABETH TIPTON is an Assistant Professor of Applied Statistics in the Department of Human Development at Teachers College, Columbia University, 425 W 120th Street, New York, NY 10027; email: tipton@tc.columbia.edu. Her research interests are in the design and analysis of large-scale randomized experiments and meta-analysis.

JAMES E. PUSTEJOVSKY is an Assistant Professor in the Educational Psychology Department at the University of Texas at Austin, Department of Educational Psychology, 1 University Station D5800, Austin, TX 78712; e-mail: pusto@austin.utexas.edu. His interests include statistical methods for meta-analysis and statistical models and analytic methods for single-case research.

Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression

Many research syntheses include studies that contribute multiple effect size estimates based on a common pool of subjects. While it is not generally reasonable to treat such effect sizes as independent, standard methods for quantitative synthesis provide no easy way to handle them. Rather, univariate meta-analysis methods are premised on the assumption that all of the effect size estimates are independent, while multivariate meta-analysis methods are premised on the assumption that the dependence structure of the effect size estimates is fully known. In real applications, neither approach is without problems.

In univariate meta-analysis, dependent effect sizes are conventionally handled by averaging them into a single synthetic, study-level effect, which creates problems when the goal is to meta-analyze moderator variables that vary within study. As a partial remedy, some analysts apply a "shifting unit of analysis" approach (H. M. Cooper, 2010), wherein effect sizes within a study are aggregated only if they have the same value of a categorical moderator. For example, if several studies report both writing and reading measures, three different average effect sizes are reported. First, the writing and reading measures within each study are averaged and then pooled together across studies in order to calculate an overall average effect size. Second and third, average-writing and average-reading effect sizes are calculated by pooling separately the writing and reading outcomes across studies. While this approach does meet the assumptions of univariate meta-analysis, it does not permit statistical comparisons between the writing and reading effects, because it does not account for dependence between effect sizes that are drawn from the same study but have different values of the moderator variable.

A more principled approach to handling dependent effect sizes is to apply a fully multivariate meta-analytic model. However, the multivariate model requires good estimates of the correlation between dependent effect sizes, which typically are not available to the meta-analyst. For example, a synthesis of experimental studies may include some that report treatment effects on both oral and written ability measures; estimating the dependence between these effect sizes requires knowledge of the correlation between the outcome measures (Gleser & Olkin, 2009), but this information is often not available from published reports. Lacking estimates of the dependence among the effect sizes nested within each study, the standard errors, confidence intervals, and hypothesis tests pertaining to grand-average effect sizes and meta-regression coefficients will be inaccurate.

A recent innovation in meta-analysis is the introduction of a robust variance estimator [RVE] that allows for the inclusion of multiple, correlated effect sizes in a meta-analysis (Hedges, Tipton, & Johnson, 2010). RVE is appealing because it allows for the inclusion of dependent effect size estimates without requiring full knowledge about their correlation structure. This allows the analyst to estimate average effect sizes and meta-regression coefficients (such as contrasts between categorical moderator variables), without having to aggregate effect size estimates in ad-hoc fashion or to collect information about the correlation among dependent effect size estimates. Due to these advantages, RVE has been widely employed, including meta-analyses in education (e.g., Wilson, Lipsey, Tanner-Smith, Huang, & Steinka-Fry, 2011), psychology (e.g., Samson, Ojanen, & Hollo, 2012; Uttal et al., 2013), and intervention science (e.g., De Vibe, Bjørndal, Tipton, Hammerstrøm, & Kowalski, 2012; Tanner-Smith, Wilson, & Lipsey, 2013). Software implementations of RVE are available in both R

(robumeta package; Fisher & Tipton, 2014) and Stata (robumeta macro; Hedberg, 2011; Tanner-Smith & Tipton, 2014), and to a limited degree in SPSS (see Tanner-Smith & Tipton, 2014).

The statistical theory behind RVE is asymptotic, in that it provides an approximately unbiased estimator of the true sampling variance if the number of independent studies is large. However, when the number of studies is not sufficiently large, the estimator is biased downward and the Type I error rate of hypothesis tests based on RVE can be much too liberal (Hedges et al., 2010; Tipton, 2014). This is a serious limitation, given that at least half of meta-analyses in education and the social sciences contain fewer than 40 studies (Ahn, Ames, & Myers, 2012; Polanin, 2013). To address this shortcoming, Tipton (Tipton, 2014) proposed small-sample corrections for hypothesis tests of single meta-regression coefficients (i.e., t-tests), which have close to nominal Type-I error even when the number of studies is small. However, as of yet no methods exist for multiple contrast hypothesis tests.

The goal of the present investigation is to develop small-sample corrections for multiple contrast hypothesis tests (i.e., F-tests), such as the omnibus test of meta-regression fit or a test for equality of three or more levels of a categorical moderator. For example, a collection of results to be synthesized might include studies conducted on students at three different grade-levels, categorized as "elementary," "middle," or "high" school. In order to answer the questions "Does the effectiveness of this intervention vary in relation to grade-level?" the analyst would normally use an F-test for the null hypothesis of no differences between the average effect sizes in each category.

In this paper, we consider five different small-sample corrected tests, all of which are based on approximations to the distribution of a Wald test statistic. The approximations are developed by generalizing previous work on simpler, specific cases of tests based on cluster-

robust variance estimation, such as the multiple-group Behrens-Fisher problem and tests of regression coefficients based on heteroskedasticity-robust variance estimation. In the following sections, we present new analytic work describing several potential approximations and a large simulation study that compares these potential solutions. After a discussion regarding these findings for practice, we illustrate the practical implications of the proposed small sample corrections in an example based on a meta-analysis conducted by Wilson, Tanner-Smith, Lipsey, Steinka-Fry, and Morrison (2011).

Robust variance estimation in multi-variate meta-regression

We will develop methods under the general meta-regression model

$$\mathbf{T}_j = \mathbf{X}_j\boldsymbol{\beta} + \boldsymbol{\varepsilon}_j \tag{1}$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of coefficients, $\mathbf{T}_j$ is a $k_j \times 1$ vector of effect size estimates from study j; $\mathbf{X}_j$ is a $k_j \times p$ matrix of covariates; and $\boldsymbol{\varepsilon}_j$ is a $k_j \times 1$ vector of errors with mean zero and covariance matrix $\boldsymbol{\Sigma}_j$, all for $j = 1,\ldots,m$. This general meta-regression model encompasses estimation of an overall average effect size (i.e., using an intercept-only model), models with categorical moderators, and meta-regressions that include quantitative predictors. The model also encompasses univariate meta-analysis, where each study contributes $k_j = 1$ effect size. For ease of notation, denote $\mathbf{y} = \left(\mathbf{y}_1^T,\ldots,\mathbf{y}_m^T\right)^T$, $\mathbf{X} = \left(\mathbf{X}_1^T,\ldots,\mathbf{X}_m^T\right)^T$, $\boldsymbol{\varepsilon} = \left(\boldsymbol{\varepsilon}_1^T,\ldots,\boldsymbol{\varepsilon}_m^T\right)^T$, and $\boldsymbol{\Sigma} = \text{diag}\left(\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_m\right)$. Let $K = \sum_{j=1}^{m} k_j$ denote the total number of effect sizes.

We shall consider tests for null hypotheses of the form H_0: $\mathbf{Cb} = \mathbf{c}$ for fixed $q \times p$ contrast matrix $\mathbf{C}$ and q x 1 vector $\mathbf{c}$. For example, an omnibus test of regression specification might be written $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$, with $q = p - 1$. A test for a single meta-regression coefficient

β_s is a special case where $q = 1$, with $\mathbf{c} = 0$ and $\mathbf{C}$ set to a $1 \times p$ vector with entry s equal to one and all other entries equal to zero.

Dependence between the effect sizes within a given study can arise from any of several sources. Two common forms of dependence found in meta-analysis, as introduced by Hedges et al (2010), are "correlated effects" and "hierarchical effects." Correlated effects are common when primary studies report multiple outcomes measured on the same individuals. For example, measures of writing and reading may be collected, or measures may be collected at multiple time points. Hierarchical effects arise when outcomes are collected on different groups of individuals, but those groups may share some common influences. For example, a primary study may report the results of two separate experiments conducted in the same lab, with the same subject pool, and same laboratory protocols. In practice, both dependence types can occur within the same study; however, the RVE approach asks practitioners to choose one or the other as a "working model" for the covariances $\boldsymbol{\Sigma}_1,\ldots,\boldsymbol{\Sigma}_m$, which are used to create approximately inverse-variance weighting matrices $\mathbf{W}_1,\ldots,\mathbf{W}_m$. Exact forms for suggested working models and weighting matrices can be found in Appendix A. While RVE does not require the use of inverse-variance weights, using them tends to increase the precision of the estimator to the extent that the working model is a reasonable approximation for the true covariance structure.

Writing $\mathbf{W} = \text{diag}\left(\mathbf{W}_1,...,\mathbf{W}_m\right)$, the weighted least-squares estimate of $\boldsymbol{\beta}$ is

$$\mathbf{b} = \mathbf{M}\left(\sum_{j=1}^{m} \mathbf{X}_j^T \mathbf{W}_j \mathbf{T}_j\right)$$

where $\mathbf{M} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}$. The exact variance of $\mathbf{b}$ is

$$\text{Var}\left(\mathbf{b}\right) = \mathbf{M}\left[\sum_{j=1}^{m} \mathbf{X}_j^T \mathbf{W}_j \boldsymbol{\Sigma}_j \mathbf{W}_j \mathbf{X}_j\right]\mathbf{M},$$

which is a function of the weights $\mathbf{W}_j$, design-matrices $\mathbf{X}_j$, and study-specific covariances $\mathbf{\Sigma}_j$. If the structure of $\mathbf{\Sigma}_j$ is fully known and the weights are defined so that $\mathbf{W}_j = \mathbf{\Sigma}_j^{-1}$ for $j = 1,...,m$, then the formula reduces to Var(**b**) = **M**. However, the assumption that the covariance structure is correctly modeled will often be difficult to meet in practice. The crucial advantage of RVE is that it provides a means to estimate the variance of **b** without relying on this stringent assumption, and makes use of the working covariance model only to improve the precision of the meta-regression estimates and the small-sample performance of hypothesis tests based on RVE.

The robust variance estimator

A general expression for the RVE estimator of Var(**b**) is given by

$$\mathbf{V}^R = \mathbf{M}\left[\sum_{j=1}^{m} \mathbf{X}_j^T \mathbf{W}_j \mathbf{A}_j \mathbf{e}_j \mathbf{e}_j^T \mathbf{A}_j^T \mathbf{W}_j \mathbf{X}_j\right]\mathbf{M} \tag{2}$$

where $\mathbf{e}_j = \mathbf{T}_j - \mathbf{X}_j\mathbf{b}$ is the vector of residuals for study j and $\mathbf{A}_j$ is a $k_j \times k_j$ adjustment matrix. In the original formulation of Hedges et al. (2010), the adjustment matrices were set to identity matrices of dimension k_j; Tipton (Tipton, 2014) considered other forms of adjustment matrices, which will be described below. Note that in RVE, the true covariance $\mathbf{\Sigma}_j$ is estimated by $\mathbf{e}_j\mathbf{e}_j^T$. Although the estimate for any given study may be quite inaccurate when considered in isolation, under some general assumptions, $\mathbf{V}^R$ nonetheless converges in probability to Var(**b**) as the number of studies increases to infinity (see Hedges et al., 2010 Appendix A).

The RVE estimator can be used to construct Wald statistics for single- and multi-parameter hypothesis tests. A Wald-type test statistic for H_0: **Cb** = **c** is given by

$$Q = \left(\mathbf{Cb} - \mathbf{c}\right)^T \left(\mathbf{C}\mathbf{V}^R\mathbf{C}^T\right)^{-1} \left(\mathbf{Cb} - \mathbf{c}\right) \tag{3}$$

It can be shown that, under the null hypothesis, Q follows a chi-square distribution with q degrees of freedom when the number of independent studies is sufficiently large. However, this

asymptotic approximation can be quite poor when the number of studies is small or moderate, as we demonstrate in a later section. Furthermore, it is not always clear when the sample is sufficiently large to trust the asymptotic approximation.

Small-sample corrections for t-tests

When testing a single meta-regression coefficient H_0: $\beta_s = 0$, the Wald statistic becomes simply

$$t_s = b_s / \sqrt{V_{ss}^R}$$

where b_s denotes entry s of **b** and V_{ss}^R is the s^{th} diagonal entry of $\mathbf{V}^R$. In large samples, t_s follows a standard normal distribution under the null hypothesis. For moderate sample sizes, Hedges et al. (2010) suggested using the adjustment matrices $\mathbf{A}_j = \sqrt{m/(m-p)}\mathbf{I}_{k_j}$ to calculate $\mathbf{V}^R$ (rather than identity matrices) and comparing the statistic to a t-distribution with $m - p$ degrees of freedom. However, evidence from several simulation studies indicates that, even with these corrections, the test has inflated Type-I error when there are fewer than 40 independent studies (Hedges et al., 2010; Tipton, 2013, 2014; Williams, 2012).

Tipton (2014) proposed alternative methods for testing single meta-regression coefficients, which involve different adjustment matrices and a Satterthwaite approximation for the degrees of freedom. The best-performing test used adjustment matrices proposed by McCaffrey, Bell, and Botts (2001), which make use of a working model for the study-specific covariances $\mathbf{\Sigma}_1,...,\mathbf{\Sigma}_m$. Following Tipton (2014), we use adjustment matrices based on the same model that motivates the choice of approximately efficient weights $\mathbf{W}_j$. In this case, the adjustment matrices are given by

$$\mathbf{A}_j = \mathbf{W}_j^{-1/2}\left[\mathbf{W}_j^{-1/2}\left(\mathbf{W}_j^{-1} - \mathbf{X}_j\mathbf{M}\mathbf{X}_j^T\right)\mathbf{W}_j^{-1/2}\right]^{-1/2}\mathbf{W}_j^{-1/2} \quad (4)$$

where $\mathbf{U}^{-1/2}$ denotes the inverse of the symmetric square root of the matrix $\mathbf{U}$, which satisfies $\mathbf{U}^{-1/2}\mathbf{U}\mathbf{U}^{-1/2} = \mathbf{I}$ (see Appendix A for examples). This adjustment matrix is such that when the working covariance model is correct, the robust variance estimator $\mathbf{V}^R$ based on these adjustment matrices is an exactly unbiased estimate of the variance of $\mathbf{b}$ (McCaffrey et al., 2001). In a simulation study, Tipton (2014) showed that using the adjustment matrices given in Equation (4) typically results in a small increase in $\mathbf{V}^R$, thus improving Type I error. However, in many situations use of the adjustment matrices on their own is not enough to bring the error within nominal levels; to do so, an additional correction to the degrees of freedom is required.

Tipton (2014) considered a Satterthwaite approximation for the degrees of freedom of t_s, again based on a working covariance model. Under the working model with $\mathbf{W} = \mathbf{\Sigma}^{-1}$, the mean and variance of V_{ss}^R can be calculated from the design matrices $\mathbf{X}_j$, weighting matrices $\mathbf{W}_j$, and adjustment matrices $\mathbf{A}_j$. (Exact expressions for these quantities will be given in the next section.) The degrees of freedom corresponding to V_{ss}^R can thus be approximated by

$$\nu_s = 2\,\mathrm{E}\left(V_{ss}^R\right)^2 / \mathrm{Var}\left(V_{ss}^R\right). \tag{5}$$

A test of H_0: $\beta_s = 0$ is then obtained by comparing t_s to a t-distribution with ν_s degrees of freedom.

In a large simulation study, Tipton (2014) showed that using the Satterthwaite degrees of freedom led to larger improvements in Type I error than using only the corrected adjustment matrices given in Equation (4). The accuracy of Type I error rates depends more closely on the degrees of freedom (ν_s) than on the sample size (m), and tests with $\nu_s < 4$ or 5 typically have Type I error above nominal. Furthermore, different coefficients within a single meta-regression can have widely discrepant degrees of freedom. Beyond just the number of studies, the degrees of freedom are influenced by a combination of covariate features and tend to grow smaller when

the covariate in question is highly unbalanced or has high leverage. Because the degrees of freedom are diagnostic, in that they indicate the circumstances when small-sample corrections are needed, Tipton recommended that these small sample corrections should be used by default in RVE.

Small sample corrections to F-tests

This section describes several small sample corrections for multiple-contrast hypotheses that are akin to the corrections proposed by Tipton (2014) for single-parameter hypothesis tests. In all of the methods to be considered, we will use the adjustment matrices proposed by McCaffrey and colleagues (2001), as given in Equation (4), which are derived under a working covariance model with inverse-variance weights.

A simple, ad-hoc correction for a multiple-contrast hypothesis test would be to use the test statistic Q / q, compared to an F-distribution with q and $m - p$ degrees of freedom. This correction is analogous to the degrees of freedom correction suggested by Hedges et al. (2010); in the simulation study described below, we will refer to it as the "naïve F-test" and use it as a baseline against which to compare other corrected tests. While offering some improvement over the asymptotic chi-squared test, a potential shortcoming of the naïve F-test is that it uses the same degrees of freedom regardless of the contrasts being tested. More principled corrections would instead take into account the features of the design matrix. The challenge in finding such corrections is that they involve the distribution of the random matrix $\left(\mathbf{C}\mathbf{V}^{R}\mathbf{C}^{T}\right)^{-1}$, which is more difficult to approximate than that of the single variate V_{ss}^{R}.

Various solutions to the problem of approximating test statistics for multiple-contrast hypothesis tests have been proposed for special cases of robust variance estimators. For example, a wide variety of procedures have been proposed for testing equality among several group means

(including both univariate and multivariate means) when the groups have unequal variances or covariance matrices (for a review, see Coombs, Algina, & Oltman, 1996). For the case of univariate means (i.e., heteroskedastic one-way ANOVA), several such tests are based on comparing the analogue of the Q statistic to a multiple of an F distribution with denominator degrees of freedom that are determined from the data (e.g., Welch, 1951; Zhang, 2013). In contrast, Alexander and Govern (1994) proposed a novel test statistic that normalizes the components of Q in order to more closely follow a chi-square distribution when the per-group sample size is small. For the case of testing equality of multivariate means in two groups, Nel and van der Merwe (1986) proposed approximating $\mathbf{C}\mathbf{V}^{R}\mathbf{C}^{T}$ by a Wishart distribution, so that Q approximately follows Hotelling's T^2 distribution; Krishnamoorthy and Yu (2004) suggested a modification so that the test is invariant to affine transformation. Zhang (2012) proposed a similar approach, also using Hotelling's T^2 as an approximating distribution, for heteroskedastic multivariate analysis of variance.

Another special case of the problem arises when heteroskedasticity-robust variance estimation is used with multiple regression. In this setting, Lipsitz, Ibrahim, and Parzen (Lipsitz, Ibrahim, & Parzen, 1999) and Kauermann and Carroll (2001) described tests of single regression coefficients that maintain close-to-nominal Type-I error even in small samples, though these tests are not readily extended to multiple-contrast hypotheses. Drawing on earlier work by Alexander and Govern (1994) and Fai and Cornelius (1996), Cai and Hayes (2008) proposed a method for testing multiple-contrast hypotheses for multiple regression; their method uses the eigen-decomposition of $\mathbf{C}\mathbf{V}^{R}\mathbf{C}^{T}$ to derive a test statistic with more accurate Type-I error than the asymptotic chi-squared test based on a Q statistic.

Finally, robust methods are also used in connection with generalized estimating equations (GEE), where they are often called "sandwich" estimators. Several small-sample corrections for GEE sandwich estimators have been proposed, along with corresponding small-sample adjustments for single-contrast hypothesis tests (e.g., Fay & Graubard, 2001; Mancl & DeRouen, 2001; McCaffrey & Bell, 2006). Notably, Pan and Wall (2002) proposed a small-sample correction for Wald-type tests of multiple-contrast hypotheses in GEE models. Their approach uses a Hotelling's T^2 approximation for Q, with degrees of freedom derived by approximately matching estimates of the mean and variance of $\mathbf{C}\mathbf{V}^R\mathbf{C}^T$ to the moments of a Wishart distribution.

In this paper, we consider two broad strategies for small-sample correction that draw on methods that perform well in these special cases. Following Pan and Wall (2002), Krishnamoorthy and Yu (2004), and Zhang (2012, 2013), the first strategy is to approximate the sampling distribution of $\mathbf{C}\mathbf{V}^R\mathbf{C}^T$ using a Wishart distribution, which leads to test statistics that approximately follow Hotelling's T^2 distribution (a multiple of an F distribution). The second strategy is based on the eigen-decomposition of $\mathbf{C}\mathbf{V}^R\mathbf{C}^T$, and extends Cai and Hayes' (2008) test for models with heteroskedasticity to the more general case of models with dependent errors.

Moments of $\mathbf{V}^R$

Both of the broad strategies for approximating the sampling distribution of $\mathbf{C}\mathbf{V}^R\mathbf{C}^T$ involve estimating the mean and variance of linear combinations of the entries in $\mathbf{V}^R$. Before describing the approximations in detail, we first derive expressions for these moments, as doing so simplifies notation later. Assume that $\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_m$ are normally distributed with covariances $\text{Var}(\boldsymbol{\varepsilon}_j) = \boldsymbol{\Sigma}_i$. Let $\boldsymbol{\Omega}$ denote the true variance of $(\mathbf{C}\mathbf{b} - \mathbf{c})$, i.e., $\boldsymbol{\Omega} = \mathbf{C}\text{Var}(\mathbf{b})\mathbf{C}^T$, and note that the Q statistic can be written as

$$Q = \mathbf{z}'\mathbf{D}^{-1}\mathbf{z} \tag{6}$$

for $\mathbf{z} = \boldsymbol{\Omega}^{-1/2}(\mathbf{C}\mathbf{b} - \mathbf{c})$ and $\mathbf{D} = \boldsymbol{\Omega}^{-1/2}\mathbf{C}\mathbf{V}^{R}\mathbf{C}^{T}\boldsymbol{\Omega}^{-1/2}$. Under the null hypothesis H_0: $\mathbf{C}\mathbf{b} = \mathbf{c}$, $\mathbf{z}$ is normally distributed with mean zero and covariance $\mathbf{I}_q$. Furthermore, if the weighting matrices are exactly inverse variance, then $\mathbf{z}$ is independent of $\mathbf{e}_1,\ldots,\mathbf{e}_m$ and therefore independent of $\mathbf{D}$. The moments of $\mathbf{D}$ are given in the following theorem.

Theorem 1. Let $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{u}_3$, $\mathbf{u}_4$ denote arbitrary, fixed q × 1 vectors. If $\boldsymbol{\varepsilon}_1,\ldots,\boldsymbol{\varepsilon}_m$ are normally distributed with covariances $\text{Var}(\boldsymbol{\varepsilon}_j) = \boldsymbol{\Sigma}_i$, then

$$E\left(\mathbf{u}_1^T\mathbf{D}\mathbf{u}_2\right) = \mathbf{u}_1^T\left(\sum_{j=1}^{m}\mathbf{B}_j\boldsymbol{\Sigma}\mathbf{B}_j^T\right)\mathbf{u}_2 \tag{7}$$

and

$$\text{Cov}\left(\mathbf{u}_1^T\mathbf{D}\mathbf{u}_2, \mathbf{u}_3^T\mathbf{D}\mathbf{u}_4\right) = \sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{u}_1^T\mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_j^T\mathbf{u}_4\mathbf{u}_2^T\mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_j^T\mathbf{u}_3 + \mathbf{u}_1^T\mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_j^T\mathbf{u}_3\mathbf{u}_2^T\mathbf{B}_i\boldsymbol{\Sigma}\mathbf{B}_j^T\mathbf{u}_4\,, \tag{8}$$

where the $q \times K$ matrices $\mathbf{B}_1,\ldots,\mathbf{B}_m$ are defined as

$$\mathbf{B}_j = \boldsymbol{\Omega}^{-1/2}\mathbf{C}\mathbf{M}\mathbf{X}_j\mathbf{W}_j\mathbf{A}_j\left(\mathbf{I}_K - \mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{W}\right)_j \tag{9}$$

and the subscript on the final term denotes the rows of the matrix corresponding to study j. Proof is given in Appendix B.

In practice, the moments of $\mathbf{D}$ will need to be estimated. One approach would be to estimate the $\boldsymbol{\Sigma}_j$ empirically using the residuals $\mathbf{e}_j\mathbf{e}_j^T$. Another approach, introduced by Bell and McCaffrey (2002) and followed by Tipton (2014), is to instead use the same working model used to develop the weighting matrices $\mathbf{W}_j$ and adjustment matrices $\mathbf{A}_j$. Results from simulation studies conducted by Bell and McCaffrey (2002) found that using an empirical estimate tended to result in tests that were extremely conservative. For this reason, we focus on the second, model-based approach. When the working model is correct, so that $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$, $\mathbf{V}^R$ is an exactly

unbiased estimator of Var(**b**) = **M**, by which it follows that $E\left(\mathbf{u}_1^T \mathbf{D} \mathbf{u}_2\right) = \mathbf{u}_1^T \mathbf{I}_q \mathbf{u}_2 = \mathbf{u}_1^T \mathbf{u}_2$ and **Ω** = $\mathbf{CMC}^T$. Consequently, the $\mathbf{B}_j$ matrices can be calculated from the contrast matrix **C**, design matrix **X**, and weighting matrix **W**.

Hotelling's T^2 approximations

We first consider approximating the sampling distribution of **D** by a Wishart distribution with scale matrix $\mathbf{I}_q$, for some degrees of freedom η. Under this approximation, Q approximately follows Hotelling's T^2 distribution with dimensionality q and degrees of freedom η, so that

$$\frac{\eta - q + 1}{\eta q} Q \sim \mathrm{F}\left(q, \eta - q + 1\right). \tag{10}$$

The question is then how to choose the degrees of freedom η that well-approximate the distribution of **D**.

Let d_{st} denote the entry in row s and column t of **D** and let $I(j = k)$ be the indicator function that equals one if $j = k$ and zero otherwise. For a random $q \times q$ matrix **D** such that η**D** follows a Wishart distribution with η degrees of freedom and scale matrix $\mathbf{I}_q$, the covariances among the entries of **D** satisfy

$$\eta \operatorname{Cov}\left(d_{st}, d_{uv}\right) = I(s = u) I(t = v) + I(s = v) I(t = u) \tag{11}$$

for $s,t,u,v = 1,\ldots,q$ (Muirhead, 1982, Section 3.2.2),. For $q > 1$, the constraints on the covariances will not be satisfied exactly, and so an approximating value of η must be found. One choice, following Pan and Wall (2002), is to minimize the sum of squared differences between the left-hand and right-hand sides, which leads to

$$\eta_A = 2\left[\sum_{s=1}^{q}\sum_{t=1}^{q} \operatorname{Var}\left(d_{st}\right)\right] \Big/ \left[\sum_{s=1}^{q}\sum_{t=1}^{q}\sum_{u=1}^{q}\sum_{v=1}^{q} \operatorname{Cov}\left(d_{st}, d_{uv}\right)\right]. \tag{12}$$

An alternative, which Pan and Wall (2002) noted but did not examine, is to minimize the sum of squared differences only over the unique entries in the covariance matrix of the lower-triangle entries in **D**, which leads to

$$\eta_B = 2\left[\sum_{s=1}^{q}\sum_{t=1}^{s}\mathrm{Var}\left(d_{st}\right)\right] \Big/ \left[\sum_{s=1}^{q}\sum_{t=1}^{s}\left(\sum_{u=1}^{s-1}\sum_{v=1}^{u}\mathrm{Cov}\left(d_{st},d_{uv}\right)+\sum_{v=1}^{t}\mathrm{Cov}\left(d_{st},d_{sv}\right)\right)\right]. \tag{13}$$

A further alternative, proposed by Zhang (2012, 2013) for the special cases of heteroskedastic one-way analysis of variance and multivariate analysis of variance, is to match the total variation in **D** (i.e., the sum of the variances of its entries) to the total variation of a Wishart distribution. This approach leads to

$$\eta_Z = \frac{q(q+1)}{\sum_{s=1}^{q}\sum_{t=1}^{q}\mathrm{Var}\left(d_{st}\right)}. \tag{14}$$

Note that the variance of d_{st} can be calculated from Equation **Error! Reference source not found.** using the vectors $\mathbf{u}_1 = \mathbf{u}_3 = \mathbf{j}_s$ and $\mathbf{u}_2 = \mathbf{u}_4 = \mathbf{j}_t$, where the vector $\mathbf{j}_s$ has entry s equal to 1 and all remaining entries equal to zero and $\mathbf{j}_t$ is similarly defined. Likewise, the covariance between d_{st} and d_{uv} can be calculated $\mathbf{u}_1 = \mathbf{j}_s$, $\mathbf{u}_2 = \mathbf{j}_t$, $\mathbf{u}_3 = \mathbf{j}_u$, $\mathbf{u}_4 = \mathbf{j}_v$.

In the simulation studies described in a later section, we will refer to the tests based on the approximation in Equation (10) as T^2-A, T^2-B, and T^2-Z, depending on whether the degrees of freedom are based on (12), (13), or (14), respectively. For tests of single parameter hypotheses, all three of the T^2 degrees of freedom evaluate to $\eta = 2/\mathrm{Var}\left(d_{11}\right)$ and the resulting tests are equivalent to a t-test with Satterthwaite degrees of freedom, as described by Tipton (2014). However, the T^2-A, T^2-B, and T^2-Z tests are not exactly equivalent to the tests proposed by Pan and Wall (2002) or by Zhang (2012, 2013) because we estimate the variability of **D** under

the working covariance model, whereas these other tests are based on empirical estimates of the variability of **D**.

Eigen-decomposition approximations

Fai and Cornelius (1996) developed small-sample corrections for multi-parameter hypothesis tests in linear mixed models that better account for the uncertainty in variance component estimates. In that context, they employ the eigen-decomposition of the estimated covariance matrix and ignored the sampling variation in the eigenvectors, instead focusing solely on the sampling distributions of the eigenvalues. We consider two similar small-sample corrections for test statistics based on RVE.

Denote the eigen-decomposition of **D** as $\mathbf{P}\mathbf{\Lambda}\mathbf{P}^{\mathrm{T}}$, where **P** is an orthonormal matrix with eigenvectors $\mathbf{p}_1,\ldots,\mathbf{p}_q$ and $\mathbf{\Lambda}$ is a diagonal matrix of eigenvalues $\lambda_1,\ldots,\lambda_q$. The Wald test statistic can then be written as

$$Q = \mathbf{z}^T\mathbf{P}\mathbf{\Lambda}\mathbf{P}^T\mathbf{z} = \sum_{s=1}^{q}\frac{\left(\mathbf{p}_s^T\mathbf{z}\right)^2}{\lambda_s} = \sum_{s=1}^{q} t_s^2, \tag{15}$$

for $t_s = \mathbf{p}_s^T\mathbf{z}/\sqrt{\lambda_s}$. If we treat **P** as constant, then $\mathbf{p}_1^T\mathbf{z},\ldots,\mathbf{p}_q^T\mathbf{z}$ are distributed as independent, standard normal variates when the null hypothesis holds. Furthermore, $\mathrm{E}(\lambda_s) = 1$ and $\mathrm{Var}(\lambda_s)$ can be calculated from **Error! Reference source not found.** with $\mathbf{u}_1 = \mathbf{u}_2 = \mathbf{u}_3 = \mathbf{u}_4 = \mathbf{p}_s$. We can thus approximate the distribution of $t_1,\ldots,t_m$, by t-distributions with degrees of freedom

$$f_s = \left[\sum_{i=1}^{m}\sum_{j=1}^{m}\left(\mathbf{p}_s^T\mathbf{B}_i\mathbf{\Sigma}\mathbf{B}_j^T\mathbf{p}_s\right)^2\right]^{-1} \tag{16}$$

for $s = 1,\ldots,q$. Note that in multi-parameter tests, the degrees of freedom $f_1,\ldots,f_q$ are typically not the same as the degrees of freedom for t-tests of single coefficients. It can be shown that the two

sets of degrees of freedom are identical only if the covariates being tested are orthogonal, which rarely occurs in practice.

Eigen-decomposition F (EDF) test. Fai and Cornelius (1996) considered an adjusted F test based on the above approximation, derived by matching the first two moments of $\delta Q / q$ to an $F(q, \nu_F)$ distribution for some constant δ and some degrees of freedom ν_F. Assuming that $t_1,\ldots,t_m$ are uncorrelated, the correction terms are given by

$$\delta = \frac{E_Q^2(q-2) + 2qV_Q}{qE_Q\left(V_Q + E_Q^2\right)}, \quad \nu_F = 4 + \frac{2E_Q^2(q+2)}{qV_Q - 2E_Q^2}, \tag{17}$$

where

$$E_Q = \sum_{s=1}^{q} \frac{f_s}{f_s - 2} \quad \text{and} \quad V_Q = 2\sum_{s=1}^{q} \frac{f_s^2\left(f_s - 1\right)}{\left(f_s - 2\right)\left(f_s - 4\right)}.$$

It is possible that some f_s from Equation (16) may be less than 4, which may lead to infeasible values of V_Q. We therefore truncate the f_s at a value slightly higher than 4 when evaluating E_Q and V_Q. Note that when $q = 1$, and assuming that $f_s > 4$, the constants will evaluate to $\delta = 1$ and $\nu_F = f_1$; Q will therefore be compared to an $F(1, f_1)$ reference distribution, which is equivalent to using a t-test with Satterthwaite degrees of freedom.

Eigen-decomposition and transformation (EDT) test. A final small-sample correction also employs the eigen-decomposition of the Wald test statistic. We use a technique similar to an approach proposed by Alexander and Govern (1994) for heteroskedastic, one-way ANOVA, and further developed by Cai and Hayes (2008) in the context of multiple-contrast hypothesis tests based on heteroskedasticity-robust variance estimation. Assume that the variates $t_1,\ldots,t_q$ from Equation (15) follow independent t-distributions with degrees of freedom $f_1,\ldots,f_q$, respectively. Let $g(t; f)$ be a transformation function that normalizes a t-distribution with f degrees of freedom,

so that $g(t_s; f_s)$ will be approximately unit-normal. It follows that the squared sum of the transformed variates

$$C = \sum_{s=1}^{q} g^2\left(t_s; f_s\right) \tag{18}$$

will be approximately distributed as a chi-square with q degrees of freedom. Following Cai and Hayes (2008), we use a normalizing transformation proposed by Hill (1970). Let

$$a_s = f_s - 1/2, \quad b_s = 48a_s^2, \quad c_s = \sqrt{a_s \ln\left(1 + \frac{t_s^2}{f_s}\right)}.$$

The transformation function is then given by

$$g\left(t_s; f_s\right) = c_s + \frac{c_s^3 + 3c_s}{b_s} - \frac{4c_s^7 + 33c_s^5 + 240c_s^3 + 855c_s}{10b_s^2 + 8b_s c_s^4 + 1000b_s}.$$

In contrast to the other tests under consideration, all of which are based on comparing the Q statistic to a scaled F distribution, the EDT test is derived by altering the internal structure of Q. For $q = 1$ the EDT approach amounts to a method for evaluating the p-value corresponding to a t-statistic after transforming it to a chi-squared statistic; the result will be equal to a Satterthwaite approximation so long as the transformation function is accurate. Finally, it should be noted that the EDT approach is not exactly equivalent to the test proposed by Cai and Hayes (2008) because we estimate the degrees of freedom via the working model, rather than from the empirical residuals.

Other approaches

In the course of this research, we also investigated other possible extensions to these approaches. For example, it is possible to combine a T^2 approximation with the eigen-decomposition approach; to combine the use of the EDT adjustment with estimated degrees of freedom; and to approximate the moments of Q directly via Taylor expansions, without assuming

that **D** follows a Wishart distribution. These approaches did not result in estimators that performed anywhere near as well as those included here, and for this reason we do not include results related to them. Interested readers should contact the authors for further information.

Simulation study

In order to assess and compare the performance of the five hypothesis testing procedures described in the previous section, we conducted a large simulation study. The main aim of the study was to determine which of the five tests provides the most accurate Type-I error rates over a range of conditions likely to be encountered in practice. We focused in particular on tests that are level-α, where the empirical Type-I error rate does not exceed the nominal rate under all conditions studied. Throughout we use the naïve F-test as a benchmark because it is analogous to the original t-test presented in Hedges et al (2010) and because its performance is uniformly superior to the asymptotic chi-squared test. The simulations are limited to models where the null hypothesis is correct; considerations of power for specific alternative hypotheses remains a topic for future research.

Simulation design

The design of the simulation study largely followed that of Simulation Study #2 in Tipton (2014), which examined the performance of Satterthwaite-type corrections for single-contrast hypothesis tests. As in the previous work, we focused on correlated standardized mean difference effect sizes, as would be obtained from an intervention study that evaluated treatment effects on several outcome variables.[1] We set the correlation between outcomes within study (denoted ρ) to values of .0, .5, or .8 in order to represent null, moderate, or high levels of within-study correlation. We also varied the proportion of the total variation in effect size estimates due to between-study heterogeneity in true effect sizes (denoted I^2) across a very wide range, including

values of .00, .33, .50, .75, .90. Non-zero values of ρ or I^2 lead to dependence among the effect size estimates within a given study. We varied the number of independent studies per meta-analysis (m) over a very wide range, using values of 10, 15, 20, and 30 as well as values from 40 to 100 in increments of 20. Finally, each simulated meta-analysis included studies contributing as few as one and as many as ten effect sizes. The per-group sample size per study (n_j) ranged from 32 to 130 and included the same values as in the second simulation from Tipton (2014). We chose to use a variable number of effect sizes per study and variable sample sizes in order to emulate conditions likely to be observed in practice, as well as to induce a greater degree of imbalance in the designs, thus creating a more challenging scenario for the tests being evaluated.

Tipton (2014) showed that besides sample size, the performance of the asymptotic approximations is strongly influenced by features of the covariates and contrasts involved in the hypothesis test. In particular, the Satterthwaite degrees of freedom for a t-test depend on the degree of balance and leverage in the covariate (see also MacKinnon, 2013; McCaffrey et al., 2001). Covariates with large imbalances (e.g., a binary variable with very few 1's) or with cases with extreme leverage (e.g., most values between 0 and 30, with a couple of values as high as 100) had much smaller degrees of freedom, even when the number of studies was moderate to large.

In order to represent a range of covariate types encountered in practice, the simulation used five covariate vectors, including two binary and three continuous covariates. Exact values of the covariate vectors are available in an online supplement. X_1 is a binary, study-level covariate that exhibits large imbalances; it is equal to 1 in 15% of the studies. X_2 is a binary covariate that varies within study and is also highly unbalanced, equal to 1 for approximately 10% of the effect sizes overall and between 0% and 20% of effects within a given study. X_3 is a

continuous, study-level covariate that is roughly normally distributed. X_4 is a continuous covariate, also roughly normal, that varies within study. X_5 is a continuous, effect-size level covariate that has a highly skewed distribution. This variable is generated from a log-normal distribution, leading to cases with leverage values 4 to 70 times the average. Tipton (2014; Table 2) showed that the Satterthwaite degrees of freedom associated with individual t-tests of these covariates range from very small (X_5: df = 4.35; X_1: df = 5.21) to small (X_2: df = 9.36; X_3: df = 12.41; X_4: df = 15.84) even with as many as 40 studies. This combination of covariates allows us to explore how well these corrections perform in the "worst" cases; an estimator that performs well in these conditions will likely perform well in practice.

The RVE estimator and the degrees of freedom in the proposed hypothesis tests all involve specification of a working model for the covariance structure. In order to maintain consistency with previous work and with the way RVE is applied in practice, the simulations used a working model based on the "correlated effects" weights proposed by Hedges et al. (2010), which are described in greater detail in Appendix A. Throughout, we assume a working model in which there is no between-study heterogeneity in true effects (i.e., $I^2 = 0$), which corresponds to the use of fixed-effects weighting. It is important to note that the working model (used for weights and adjustments) and the actual data-generating model are not equivalent; instead, the two models diverge as ρ and I^2 increase, which allows us to examine the extent to which the methods are robust to model misspecification. This feature of the simulation also emulates what the analyst will encounter in practice, where the true data-generating model is unknown and the working model will seldom be correctly specified.

For each combination of simulation conditions, we fit several different regression specifications and tested many different contrasts involving different combinations of the

covariates. The combinations of regression specifications fall into two categories. First, we ran "omnibus" tests based on models with $p - 1$ covariates (plus an intercept term with coefficient β_0), where the null hypothesis was H_0: $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$. For each simulated meta-analysis, we fit all possible model specifications with 2, 3, 4, or 5 covariates (for a total of 26 specifications) and calculated the omnibus test with dimension $q = p - 1$. Second, based on a model specification including all $p - 1 = 5$ covariates, we ran "subset" tests of all possible combinations of $q = 2$, 3, or 4 of the covariates (for a total of 25 tests). For each simulated meta-analysis, we recorded the p-value from each of the 51 tests. For each of the tests, we determined the Type-I error rates corresponding to $\alpha = .01$, .05, and .10 by calculating the proportion of p-values less than α, across 5000 replications. Monte Carlo margins of error are therefore approximately 0.28% for $\alpha = .01$, 0.62% for $\alpha = .05$, and 0.85% for $\alpha = .10$. In the figures that follow, upper bounds on the empirical error rate are depicted in figures using dashed lines.

Results

This section describes five main findings from the simulation study. The first three findings address the question of which of the proposed tests performs best in terms of both Type I error and power. The final two findings focus on the roles of degrees of freedom and model misspecification. For each finding, we include a figure illustrating selected results of the simulations; in many cases, we also relate the findings to corresponding findings from Tipton (2014) regarding single-parameter t-tests. Full simulation results are available in an online supplement.

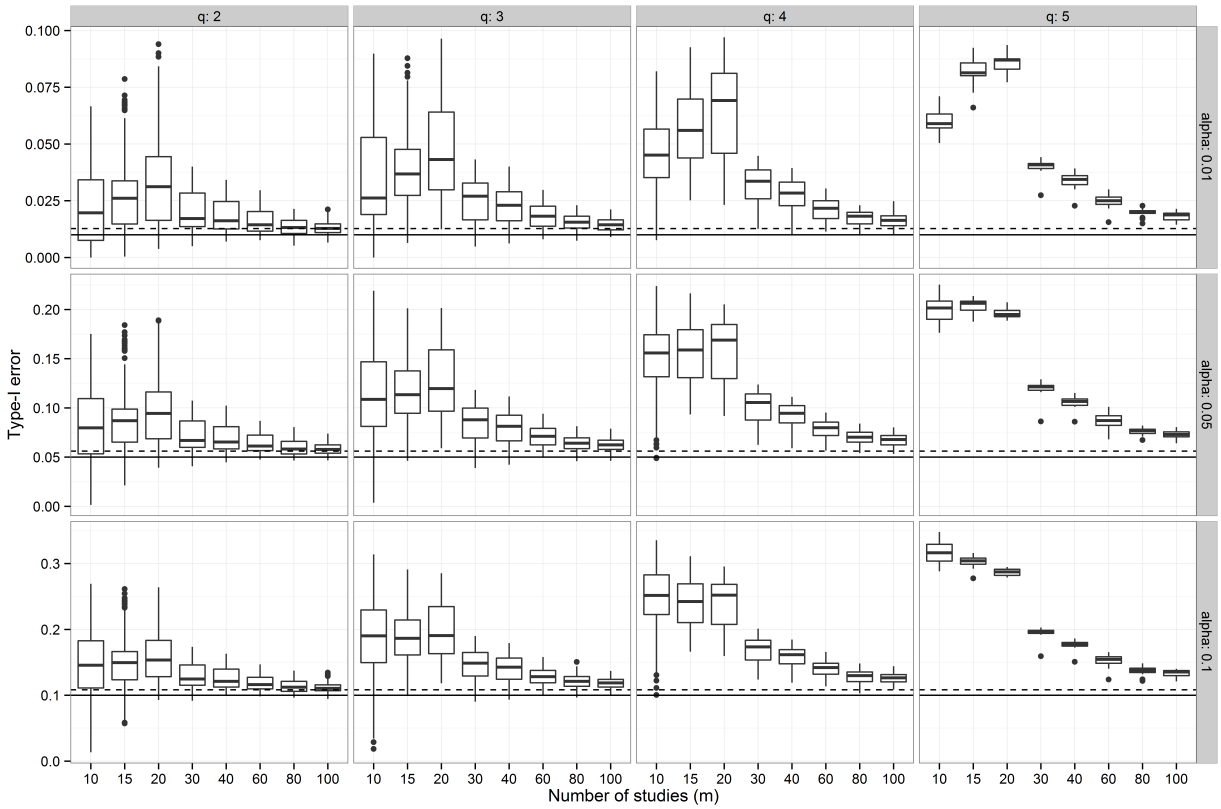
Finding 1: Naïve F-test performance

The naïve F-test compares Q / q to an F-distribution with q and $m - p$ degrees of freedom. This test is a multivariate analogue to the small sample corrections originally provided by

Hedges et al (2010) for the t-test, which Tipton (2014) found to be accurate only in rather large samples. For multiple contrast hypothesis tests, we found that the Naïve F adjustment does not control Type I error except in very large samples; additionally, the performance of the test varies greatly in relation to the number of parameters tested (q).

Figure 1 presents results comparing the sample size (m) to the Type I error rate, for each of the three α-levels under study and for four values of q (=2,3,4,5). Note that typically the true Type I error is above nominal, and that even in the best case (when q = 2 and m = 100), the results are not within the simulation error bounds for a level-α test. When there are fewer than 40 studies, the Type I error is often over twice as large as nominal (e.g., 0.10 for α=0.05). Even with 100 studies, the true Type I error is often 50% larger than nominal (e.g., 0.075 for α = 0.05; 0.15

FIGURE 1: Type I error of naïve F-test by number of studies and values of *q*.



for α = 0.10) or even larger (e.g., 0.02 for α =0.01). As the number of parameters being tested increases, these trends become more pronounced, with the largest Type I error rates occurring when q = 5 and the number of studies is small.

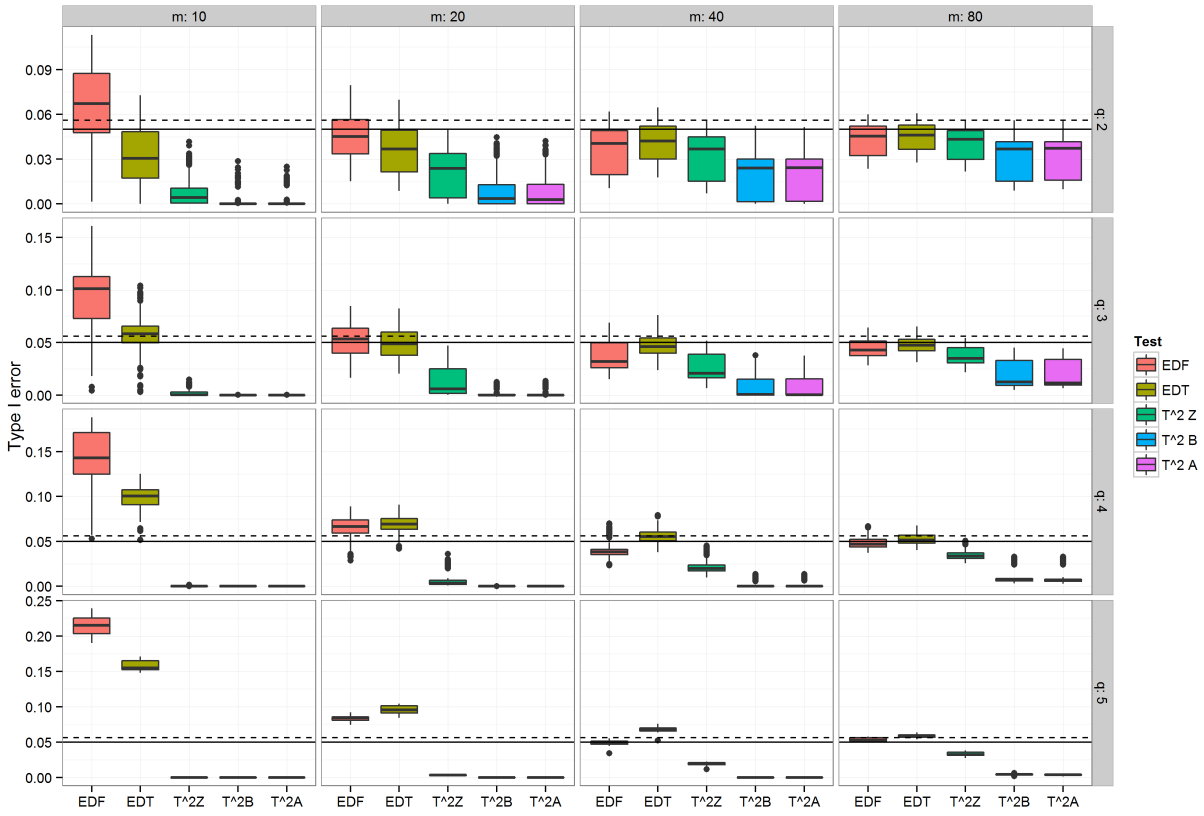
Finding 2: T^2 approximations are nearly level-α

Figure 2 presents the Type I error rates of the five proposed tests in relation to the number of studies (m) and the number of parameters in the hypothesis test (q). The figure is limited to error rates for α = 0.05, although we note that the trends are very similar for α = 0.01 and 0.10 as well. We do not include the naïve F-test in this figure because all 5 of these tests outperform the naïve F-test under every condition studied.

Examining this figure reveals several trends. First, Type I error for the EDF and EDT tests typically approach the nominal values from above, whereas the T^2-A, T^2-B, and T^2-Z tests approach the nominal values from below. This trend holds in relation to both m and q. For example, when there are 20 studies, as q increases, the Type I error rates of the EDF and EDT tests increase to values far above nominal (close to 0.10), while the error rates decrease towards zero for the T^2-A, T^2-B, and T^2-Z tests. For each value of q, the error rates of all five tests converge toward the nominal values as the number of studies increases.

Second, the EDF and EDT tests have Type I error rates that are closest to nominal, on average. However, averaging over a variety of design conditions, including covariate types and degrees of model misspecification conceals variation in the full range of errors. For example, when m = 20 and q = 3, the average Type I error of the EDT test is nominal (0.05), yet under some design conditions the error can be as large as 0.08. Because it is not possible to know a priori the true data-generating model corresponding to any real data analysis, it makes more

FIGURE 2: Type I error for α = 0.05 of 5 alternative tests

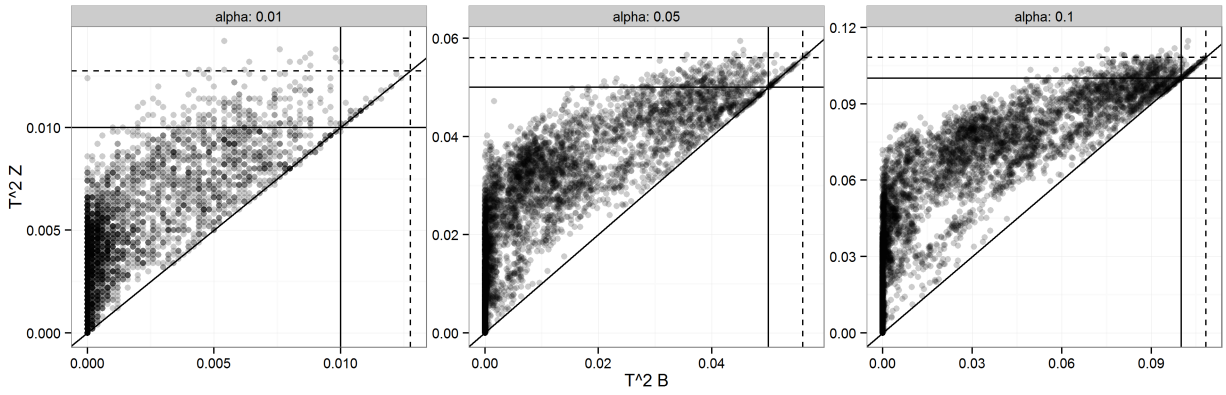


sense to compare the full range of Type I errors for each of these estimators. Doing so indicates that while the T^2-A, T^2-B, and T^2-Z tests are typically more conservative, they are also nearly always level-α. Across simulation conditions, the maximum Type I error is only 0.059, which is less than the largest Type I error found for single-parameter t-tests in Tipton, 2014. Further findings therefore focus on the three T^2 approximations.

Finding 3: T^2-Z is uniformly most powerful

Ideally, a hypothesis testing procedure should not only control the Type I error rate to be at most α, but should have Type I error as close to the nominal level as possible, so as to maximize power. Having established that T^2-A, T^2-B, and T^2-Z are all level-α under nearly all of the conditions studied, we examined which of the tests is most powerful. To address this

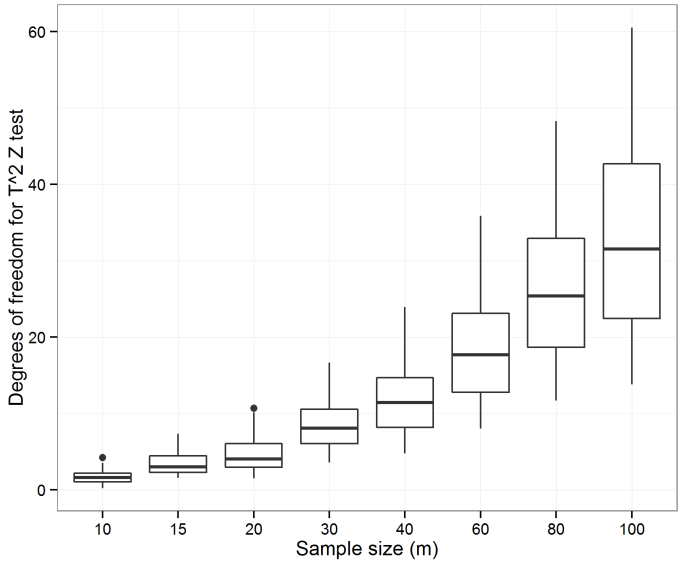
FIGURE 3: Type I error for α = 0.05 for T^2-B versus T^2-Z



question, we compared the Type I error rates of each pair of the three tests for each of the 6,120 contrasts and parameter combinations in the simulation. We found that T^2-A is less powerful than T^2-B or T^2-Z. Figure 3 plots the Type I error rates of the latter two tests, with separate panels for each α level under study. In this figure, the solid vertical and horizontal lines indicate the exact α-level, while the dashed lines indicate values within two-standard deviations of the simulation error. Points above the diagonal line represent tests and parameter combinations for which the Type I error of the T^2-Z test is larger than that of the T^2-B test.

It can be seen that the maximum Type I error of the T^2-B test is slightly smaller than that for the T^2-Z test, but these differences are minimal. It is also apparent that the Type I error of the T^2-Z test is closer to the stated α-level for every parameter combination, although both tests are conservative. For example, when the Type I error of T^2-B is nearly 0.00, the error of T^2-Z ranges from 0 – 0.008, 0 – 0.04, and 0 – 0.06 for α = 0.01, 0.05, 0.10 respectively. Furthermore, the median Type I error for the α = 0.05 test is 0.0254 for T^2-Z but only 0.0044 for T^2-B; in over

FIGURE 4: Degrees of freedom of T2-Z test by number of studies (m)



79% of cases, the Type I error of T^2-B is less than the median error of T^2-Z. In summary, T^2-Z is the most powerful level-α test across all contrasts and parameter combinations considered.

Finding 4: Relationship between m and degrees of freedom

Tipton (2014) found that for single-parameter t-tests, the performance of the test was more closely related to the degrees of freedom, which depend in part on the degree of balance or leverage in the covariate, than to the number of independent studies. We observed similar results in the simulation results for multiple contrast tests. Figure 4 illustrates the relationship between the number of studies and the degrees of freedom for the T^2-Z test. For the covariate combinations we studied here, the degrees of freedom are always smaller than the number of studies, and the range of degrees of freedom grows as the number of studies increases. For example, the degrees of freedom range from 1.5 to 10.7 with 20 studies, while they range from 13.8 to 60.5 with 100 studies. The fact that the degrees of freedom are strongly dependent on features other than the number of studies suggests that it is not possible to specify a clear rule of thumb for when the sample size is large enough to rely on asymptotic approximations alone.

Rather, the best indicator of the need for a small-sample correction would appear to be the degrees of freedom of the small-sample correction itself.

Finding 5: Robustness to model misspecification

The results reported so far have looked at overall performance, without attending to how performance relates to the parameters of the data-generating model, such as the correlation (ρ) and the between-study effect size variability (I^2). In the simulation study, the tests were always constructed based on a on the "correlated effects" working model with $\rho = 1$ and $I^2 = 0$. By simulating the data using other values of ρ and I^2, we are able to investigate the extent to which the T^2-Z test is to misspecification of the working model, as is likely to occur in practice. In further analysis, we found no consistent pattern in the relationship between Type-I error and ρ, but did find that the maximum Type I error tends to increase slightly as I^2 increases. We also observed that the T^2-Z test is *always* level-α when $I^2 < 1/2$. Given that the working matrices were developed based on $I^2 = 0$, this suggests that the model misspecification must be large before the performance of the small sample corrections is compromised. Additionally, even when this misspecification is large, the test still contains Type I error to a much greater extent than the naïve-F test or the asymptotic chi-squared test.

Discussion

We have compared five possible alternatives to the naïve F-test for multiple contrast hypothesis tests based on RVE. The simulation results indicate that the naïve F-test performs very poorly in nearly all of the conditions under consideration, even when the meta-analysis includes as many as 100 studies. The poor performance of the naïve F-test is due to the fact that the ad-hoc degrees of freedom of the naïve F-test do not take into account covariate features, such as the degree of balance or leverage, that impact the Type I error rates. Given the typical

sizes of meta-analyses in the educational and social sciences, the chi-squared test and naïve F-test should rarely be used. Instead, a small-sample corrected test should always be employed for multiple contrast hypothesis tests based on RVE.

While the T^2-A, T^2-B, and T^2-Z tests presented here all control the Type I error rate—even in very small samples—the T^2-Z test clearly outperforms the others. Except under conditions of extreme model misspecification, the T^2-Z test maintains the nominal level-α and has Type I error rates closer to the stated α-level than any other test studied. Furthermore, even with large model misspecification, the maximum Type I error observed was only slightly beyond simulation error (e.g., 0.0594 when $\alpha = .05$). It is important to highlight, too, that in practice we would not expect to encounter such extreme model misspecification, because the default practice in RVE is to estimate the degree of between study heterogeneity instead of assuming $I^2 = 0$.

A further advantage of the T^2-Z test is that it performs well even with degrees of freedom close to zero. This is counter to the results found in Tipton (2014) for t-tests, where the Type I error was sometimes too liberal when the Satterthwaite degrees of freedom were smaller than 4 or 5. To see why these results differ, note that the dimension of the test (q) impacts both the test-statistic itself and the reference distribution: as q increases, the multiplier $[(\eta - q + 1)/\eta]$ decreases, as do the denominator degrees of freedom (i.e., $\eta - q + 1$). This difference is starkest when comparing $q = 1$ and $q = 2$ for a model with a fixed number of covariates. When $q = 1$, the multiplier reduces to 1, whereas for $q = 2$, the multiplier is strictly less than 1; similarly, the denominator degrees of freedom shift from η to $\eta - 1$. These two factors penalize the power of the test for any value of $q > 1$, thus leading to a test that is level-α, even when the degrees of freedom are near zero.

Application

In order to illustrate the performance of the T^2-Z test in practice, as well as to better understand differences between it and the naïve F-test, we apply both tests to a meta-analysis of dropout prevention and intervention programs conducted by Wilson et al (2011). This systematic review focused on school- or community-based programs aimed at primary and secondary students; included studies measured school completion or school dropout outcomes, and were reported between 1985 and 2010. The review included 385 log-odds-ratio effect sizes from 317 independent samples; although only a single outcome was reported for most samples, multiple outcomes were reported for 48 (15%) of the samples. The samples were nested within 152 reports of experimental or quasi-experimental studies; 102 studies reported results for just a single sample, with the remainder reporting results for between 2 and 30 samples.

In the original meta-analysis, a meta-regression model was provided predicting the logged odds ratio effect size for general programs (Model 3, Table 3.4, Wilson et al); this model included methodological characteristics (e.g., study design indicators, attrition, an independent evaluator scale), participant characteristics (% male, %white, average age), and program characteristics (e.g., implementation quality, indicators of program format). The model was estimated using RVE, with weights corresponding to the "correlated effects" working model.

In order to understand the performance of the naïve F-test and T^2-Z, we calculated five separate F-tests for this model, including tests of the role of study design ($q = 2$); the outcome measure type ($q = 3$); evaluator independence ($q = 3$); implementation quality ($q = 2$); and program format ($q = 3$). In the original study, evaluator independence and implementation quality were included as scales, varying from 1 to 4 or 1 to 3, respectively. For the sake of example we treat these covariates as categorical and model them with indicator variables.

Notably, the original study did not report F-tests for any of the covariates because multiple-contrast hypothesis test functions were not available in RVE software at the time.

Table 1 reports the results of the five F-tests based on two different modeling approaches. In the first model, reported in the top panel, we treat the 317 unique samples as independent and use an RVE model with correlated effects weights (with $\rho = 0.80$). This approach is consistent with the original analysis reported by Wilson and colleagues. An alternative approach is to instead treat the 152 studies as independent, but to allow for dependence between samples from common studies. The lower panel of Table 1 reports the results based on this "hierarchical effects" working model. The latter model specification leads to a substantially smaller number of independent samples, across which the 385 effect sizes are quite unequally distributed.

Under the correlated effects model, the naïve F-test and the T^2-Z test lead to consistent conclusions for four out of the five tests. Both tests indicate that the effects of study design and

Table 1: Comparison of Naïve-F tests and T^2-Z tests in Dropout Example

			Naïve F test*		T^2-Z test		
Cluster level	**Factor**	**q**	**F**	**p-val**	**F**	**df**	**p-val**
Sample							
	Study design	2	0.11	0.894	0.11	103	0.895
	Outcome measure	3	1.86	0.136	1.81	64	0.155
	Evaluator independence	3	3.36	0.019	3.05	20	0.052
	Implementation quality	2	9.46	<0.001	9.38	110	<0.001
	Program format	3	4.31	0.005	4.19	74	0.008
Study							
	Study design	2	0.23	0.796	0.22	43	0.800
	Outcome measure	3	0.92	0.436	0.84	22	0.488
	Evaluator independence	3	3.11	0.029	2.78	17	0.073
	Implementation quality	2	14.15	<0.001	13.78	37	<0.001
	Program format	3	3.85	0.011	3.65	38	0.021

Notes:
df = degrees of freedom. p-val = p-value.
** The Naïve F test uses 295 degrees of freedom for the model with sample-level clustering and 130 degrees of freedom for the model with study-level clustering.*

outcome measure are *not* statistically significant, while the effects of implementation quality and program format *are* significant (all $p < 0.05$). In contrast, the effect of evaluator independence is statistically significant when using the naïve F-test ($p = 0.019$), but not when using the T^2-Z test ($p = 0.052$). For this covariate, the small-sample correction substantially affects both the test statistic (the T^2-Z F statistic is 11% smaller than the naïve F) and the degrees of freedom used in the reference distribution (20 versus 290). The low degrees of freedom for evaluator independence are due to the unequal allocation of effect sizes to the implementation quality categories (i.e., 6, 33, 43, and 303 effect sizes in the four categories). It is interesting to note that the degrees of freedom used in the T^2-Z test vary quite markedly across the five covariates that were tested. Even though the small-sample correction did not lead to substantial changes in the significance level for most of the covariates, the differences in degrees of freedom still illustrates that the T^2-Z test provides useful diagnostic information.

In the bottom panel of Table 1, the test results based on the hierarchical effects model are quite similar to the results of the correlated effects model, despite using a substantially smaller number of independent clusters. Again, the naïve F-test and the T^2-Z test lead to a consistent pattern of statistical significance for four of the five covariates, and the test of evaluator independence changes significance level as a result of the small-sample correction. The main difference between the two model specifications is that the degrees of freedom are mostly much smaller in the hierarchical effects model than in the correlated effects model; for example, they decrease from $df = 109$ to $df = 37$ for implementation quality.

Overall, this example illustrates two points, each of which echoes results from Tipton (in press) regarding small sample t-tests. First, the degrees of freedom in the T^2-Z test can vary widely across covariates and tests, making it difficult to know a priori when small sample

corrections are required. Even in studies with a large number of clusters, the degrees of freedom for some tests can be very small, suggesting that the T^2-Z test is best used as a default in practice. Second, the degrees of freedom in all of these tests are typically much smaller than the number of clusters, with the maximum observed here being less than one third of the number of clusters. This suggests that in small and even more moderate sized meta-analyses, F-tests may often have relatively low power, particularly for tests driven by highly unbalanced covariates.

Conclusion

Robust variance estimation has rapidly become a widely used tool for combining effect sizes in meta-analysis. The fact that RVE does not require any assumptions regarding either the error distribution or covariance structure has allowed analysts to summarize relationships across all collected effect sizes, instead of reducing the data to fit the statistical method. However, practical applications of RVE are limited by the fact that, while the method performs well in large-samples, hypothesis tests based on RVE have inflated Type I error rates in small and moderate samples. Building on corrections to t-tests found in Tipton (2014), this paper has developed similar corrections to Wald-type tests of multiple-contrast hypotheses. Based on initial simulation evidence, we recommend that the T^2-Z test be applied in meta-analyses that use RVE methods, and that the asymptotic chi-square test and naïve F-test be avoided. In combination with earlier work, the methods that we have described now provide means for hypothesis testing of single covariates, multiple covariates (e.g., categorical moderators), and omnibus model-fit tests within the RVE framework, all with Type I error at or below the stated α-level.

There are a few limitations to the work presented in this paper, which point towards outstanding questions where continued research is needed. While we have examined the performance of a variety of small-sample corrections, our scope was limited to closed-form

corrections for Wald-type test statistics. Thus, we have not considered iterative procedures such as bootstrapping, although some recent work in econometrics has proposed promising techniques for bootstrapping in settings with cluster-dependent observations (Webb & MacKinnon, 2013). Neither have we considered corrections such as saddlepoint approximations. McCaffrey and Bell (2006) present evidence that, for tests of single regression coefficients, a saddlepoint approximation may provide even more accurate error control than the Satterthwaite approximation studied in Tipton (2014). However, extending this approach to Wald-type test statistics for multiple contrast hypotheses is not straight-forward.

Like Tipton (2014) we have also employed a working model approach when estimating the degrees of freedom. Some results from previous research on small-sample corrections for cluster-robust variance estimation suggest that this approach may lead to less conservative tests than using an empirical estimate of the covariance matrix (McCaffrey et al., 2001), although we have not investigated this in our simulations. Finally, our examination of different small-sample corrections has focused exclusively on Type I error, rather than also considering power. Future work will also need to examine the power of the recommended small-sample corrected tests under non-null alternative hypotheses. It would be particularly useful to compare the RVE approach (with the T^2-Z test) with other approaches such as fully model-based multivariate meta-analysis.

As with any simulation study, our conclusions are limited by the data-generating models and parameters considered. While we have included a wide range of values, we have not studied all possible conditions encountered in meta-analysis, and have focused only on standardized mean difference effect sizes. While work by Tipton (2013) and Williams (2012) suggests that in moderate samples, RVE performs equally well for log-odds ratios and regression coefficients, no

work to date has investigated the performance of small sample corrections with families of effect sizes other than standardized mean differences.

Additionally, our simulation study focused on a single type of covariance structure–the correlated effects model–in conjunction with a working model and weight matrices developed in Hedges et al (2010). This weighting scheme is not necessarily fully efficient, even when the between-study variance is estimated (see Appendix A). Future work should investigate the extent to which the choice of working model and weights influences the size and power of the tests in small samples.

Finally, we note that the corrections developed here are not limited in application to meta-analysis. In future work we plan to investigate the performance of the corrections we have described under more general models with cluster dependence, such as hierarchical linear models and generalized estimated equations.

Note

1. The exact data-generating model was as follows. Let $\mathbf{1}_j$ denote a $k_j \times 1$ vector of 1's; let $\mathbf{I}_j$ denote a $k_j \times k_j$ identity matrix; and let $\mathbf{\Psi}_j = \rho 1_j 1_j^T + (1-\rho)\mathbf{I}_j$ be a compound-symmetric correlation matrix with intra-class correlation ρ. Each meta-analysis contained a total of m studies. For study j, we generated k_j standardized mean differences by simulating the numerator and denominator. The numerators were generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\tau^2 \mathbf{1}_j \mathbf{1}_j^T + (2/n_j)\mathbf{\Psi}_j$, where n_j represents the per-group sample size and τ^2 represents the between-study variance in true effects. The denominators were generated by simulating diagonal elements from a Wishart distribution with $2n_j$ - 2 degrees of freedom and scale matrix $\mathbf{\Psi}_j$, dividing by $2n_j - 2$, and taking the square roots. For ease of

interpretation, we re-parameterized the between-study variance using the I^2 measure of heterogeneity.

References

Ahn, S., Ames, A. J., & Myers, N. D. (2012). A Review of Meta-Analyses in Education: Methodological Strengths and Weaknesses. *Review of Educational Research*, *82*(4), 436–476. doi:10.3102/0034654312458162

Alexander, R. A., & Govern, D. M. (1994). A new and simpler approximation for ANOVA under variance heterogeneity. *Journal of Educational and Behavioral Statistics*, *19*(2), 91–101. doi:10.3102/10769986019002091

Bell, R. M., & McCaffrey, D. F. (2002). Bias reduction in standard errors for linear regression with multi-stage samples. *Survey Methodology*, *28*(2), 169–181.

Cai, L., & Hayes, A. F. (2008). A new test of linear hypotheses in OLS regression under heteroscedasticity of unknown form. *Journal of Educational and Behavioral Statistics*, *33*(1), 21–40. doi:10.3102/1076998607302628

Coombs, W., Algina, J., & Oltman, D. (1996). Univariate and multivariate omnibus hypothesis tests selected to control type I error rates when population variances are not necessarily equal. *Review of Educational Research*, *66*(2), 137–179.

Cooper, H. M. (2010). *Research Synthesis and Meta-Analysis* (4th ed.). Thousand Oaks, CA: SAGE Publications.

De Vibe, M., Bjørndal, A., Tipton, E., Hammerstrøm, K., & Kowalski, K. (2012). Mindfulness based stress reduction (MBSR) for improving health, quality of life, and social functioning in adults. *Campbell Systematic Reviews*, *2012*(3). doi:10.4073/csr.2012.3

Fai, A. H.-T., & Cornelius, P. (1996). Approximate F-tests of multiple degree of freedom hypotheses in generalized least squares analyses of unbalanced split-plot experiments. *Journal of Statistical Computation and Simulation*, *54*(4), 363–378.

Fay, M. P., & Graubard, B. I. (2001). Small-sample adjustments for Wald-type tests using sandwich estimators. *Biometrics*, *57*(4), 1198–206.

Fisher, Z., & Tipton, E. (2014). robumeta: An R-package for robust variance estimation in meta-analysis. *Journal of Statistical Software*.

Gleser, L. J., & Olkin, I. (2009). Stochastically dependent effect sizes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (2nd ed., pp. 357–376). New York, NY: Russell Sage Foundation.

Hedberg, E. C. (2011). ROBUMETA: Stata module to perform robust variance estimation in meta-regression with dependent effect size estimates. Statistical Software Components S457219, Boston College Department of Economics.

Hedges, L. V, Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. doi:10.1002/jrsm.5

Hill, G. W. (1970). Algorithm 395: Student's t quantiles. *Communications of the ACM*, *13*, 617–619.

Kauermann, G., & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, *96*(456), 1387–1396.

Krishnamoorthy, K., & Yu, J. (2004). Modified Nel and Van der Merwe test for the multivariate Behrens–Fisher problem. *Statistics & Probability Letters*, *66*(2), 161–169. doi:10.1016/j.spl.2003.10.012

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Lipsitz, S. R., Ibrahim, J. G., & Parzen, M. (1999). A degrees-of-freedom approximation for a t-statistic with heterogeneous variance. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *48*(4), 495–506. doi:10.1111/1467-9884.00207

MacKinnon, J. G. (2013). Thirty years of heteroskedasticity-robust inference. In X. Chen & N. R. Swanson (Eds.), *Recent Advances and Future Directions in Causality, Prediction, and Specification Analysis*. New York, NY: Springer New York. doi:10.1007/978-1-4614-1653-1

Mancl, L. A., & DeRouen, T. A. (2001). A covariance estimator for GEE with improved small-sample properties. *Biometrics*, *57*(1), 126–134.

McCaffrey, D. F., & Bell, R. M. (2006). Improved hypothesis testing for coefficients in generalized estimating equations with small samples of clusters. *Statistics in Medicine*, *25*(23), 4081–98. doi:10.1002/sim.2502

McCaffrey, D. F., Bell, R. M., & Botts, C. H. (2001). Generalizations of biased reduced linearization. In *Proceedings of the Annual Meeting of the American Statistical Association*.

Nel, D., & van der Merwe, C. (1986). A solution to the multivariate Behrens-Fisher problem. *Communications in Statistics - Theory and Methods*, *15*(12), 3719–3735.

Pan, W., & Wall, M. M. (2002). Small-sample adjustments in using the sandwich variance estimator in generalized estimating equations. *Statistics in Medicine*, *21*(10), 1429–41. doi:10.1002/sim.1142

Polanin, J. (2013). *Addressing the Issue of Meta-Analysis Multiplicity in Education and Psychology*. Loyola University. Retrieved from http://ecommons.luc.edu/luc_diss/539/

Samson, J. E., Ojanen, T., & Hollo, A. (2012). Social goals and youth aggression: Meta-analysis of prosocial and antisocial goals. *Social Development*, *21*(4), 645–666. doi:10.1111/j.1467-9507.2012.00658.x

Skinner, C. J. (1989). Domain means, regression and multivariate analyses. In C. J. Skinner, D. Holt, & T. M. F. Smith (Eds.), *Analysis of Complex Surveys* (pp. 59–88). New York, NY: Wiley.

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*(1), 13–30. doi:10.1002/jrsm.1091

Tanner-Smith, E. E., Wilson, S. J., & Lipsey, M. W. (2013). The comparative effectiveness of outpatient treatment for adolescent substance abuse: a meta-analysis. *Journal of Substance Abuse Treatment*, *44*(2), 145–58. doi:10.1016/j.jsat.2012.05.006

Tipton, E. (2013). Robust variance estimation in meta-regression with binary dependent effects. *Research Synthesis Methods*, *4*(2), 169–187. doi:10.1002/jrsm.1070

Tipton, E. (2014). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*. doi:10.1037/met0000011

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352–402. doi:10.1037/a0028446

Webb, M., & MacKinnon, J. G. (2013). *Wild Bootstrap Inference for Wildly Different Cluster Sizes* (No. 1314). Kingston, Ontario, Canada.

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

Williams, R. T. (2012). *Using robust standard errors to combine multiple regression estimates with meta-analysis*. Loyola University. Retrieved from http://ecommons.luc.edu/luc_diss/405/

Wilson, S. J., Lipsey, M. W., Tanner-Smith, E., Huang, C. H., & Steinka-Fry, K. T. (2011). Dropout prevention and intervention programs: Effects on school completion and dropout Aaong school-aged children and youth: A systematic review. *Campbell Systematic Reviews*, *7*(8).

Zhang, J.-T. (2012). An approximate Hotelling T2 -test for heteroscedastic one-way MANOVA. *Open Journal of Statistics*, *2*, 1–11.

Zhang, J.-T. (2013). Tests of linear hypotheses in the ANOVA under heteroscedasticity. *International Journal of Advanced Statistics and Probability*, *1*(2), 9–24.

Appendix A: RVE working models and weights

The methods presented in Hedges et al. (2010), in Tipton (2014), and in this paper require the specification of a working model for the covariance structure. Hedges et al. (2010) provide two possible working models likely to be found in meta-analyses: correlated effects and hierarchical effects. In the correlated effects model, the effect sizes from study j are assumed to have covariance

$$\mathbf{\Sigma}_{cj} = \tau^2\mathbf{J}_j + \rho v_j(\mathbf{J}_j - \mathbf{I}_j) + v_j\mathbf{I}_j,$$

where $\mathbf{I}_j$ is a $k_j \times k_j$ identity matrix; $\mathbf{J}_j$ is a $k_j \times k_j$ matrix of 1's; τ^2 is a measure of the variation in study-average effect sizes across studies; v_j is the estimation error variances for the k_j effect sizes in study *j*, which is assumed constant within studies ($v_{ij} = v_j$); and ρ is an assumed constant correlation between effect sizes. In the hierarchical effects model, the effect sizes are assumed to follow

$$\mathbf{\Sigma}_{hj} = \tau^2\mathbf{J}_j + \omega^2\mathbf{I}_j + \mathbf{V}_j,$$

where ω^2 is a measure of the within-study variation in true effect sizes and $\mathbf{V}_j = \text{diag}(v_{1j},\ldots,v_{kjj})$ is a $k_j \times k_j$ diagonal matrix of the estimation error variances in study *j*. Importantly, these working models are simplified versions of what could happen in practice. For example, in $\mathbf{\Sigma}_{cj}$, the correlation ρ is assumed constant between all effect sizes and across all studies. Hedges et al provide method-of-moments estimators for τ^2 and ω^2.

Based on these working models and estimated variance components, Hedges et al propose the use of approximately inverse variance weights. For the correlated effects model, they propose to use $\mathbf{W}_{cj} = 1 / [k_j(v_{.j}+\tau^2)]\ \mathbf{I}_j$, where $v_{.j} = \Sigma v_{ij}/k_j$ is the average effect size variance in study *j*; for the hierarchical effects model, they propose to use $\mathbf{W}_{hj} = \text{diag}(w_{1j},\ldots,w_{kjj})$, where w_{ij}

$= 1/(v_{ij} + \tau^2 + \omega^2)$. Note that even under the assumed covariance structures $\mathbf{\Sigma}_{cj}$ and $\mathbf{\Sigma}_{hj}$, the proposed weights are not exactly inverse-variance. For example, $\mathbf{\Sigma}_{cj}$ and $\mathbf{\Sigma}_{hj}$ are non-diagonal while $\mathbf{W}_{cj}$ and $\mathbf{W}_{hj}$ are diagonal. The fact that the weights are only approximately inverse variance is not problematic, since in practice we have found that once weights are within the right ballpark, changes to the weights have only small effects on precision.

Appendix B: Mean and variance of V^R

This section provides a proof of Theorem 1. Begin by noting that D, which is a function of the RVE estimator, can be written as

$$\mathbf{u}_1^T \mathbf{D} \mathbf{u}_2 = \sum_{j=1}^{m} \mathbf{u}_1^T \mathbf{\Omega}^{-1/2} \mathbf{C} \mathbf{M} \mathbf{X}_j^T \mathbf{W}_j \mathbf{A}_j \mathbf{e}_j \mathbf{e}_j^T \mathbf{A}_j^T \mathbf{W}_j \mathbf{X}_j \mathbf{M} \mathbf{C}^T \mathbf{\Omega}^{-1/2} \mathbf{u}_2$$

Substituting $\mathbf{e}_j = \left(\mathbf{I}_K - \mathbf{X}\mathbf{M}\mathbf{X}^T\mathbf{W}\right)_j \boldsymbol{\varepsilon}$ into this equation, it can be seen that $\mathbf{u}_1^T \mathbf{D} \mathbf{u}_2$ is a quadratic form in $\boldsymbol{\varepsilon}$:

$$\mathbf{u}_1^T \mathbf{D} \mathbf{u}_2 = \sum_{j=1}^{m} \mathbf{u}_1^T \mathbf{B}_j \boldsymbol{\varepsilon} \boldsymbol{\varepsilon}^T \mathbf{B}_j^T \mathbf{u}_2 = \boldsymbol{\varepsilon}^T \left(\sum_{j=1}^{m} \mathbf{B}_j^T \mathbf{u}_2 \mathbf{u}_1^T \mathbf{B}_j \right) \boldsymbol{\varepsilon},$$

where $\mathbf{B}_j$ is defined in Equation (9). From the properties of quadratic forms for multivariate normal random variables, it follows that

$$E\left(\mathbf{u}_1^T \mathbf{D} \mathbf{u}_2\right) = \mathrm{tr}\left[\left(\sum_{j=1}^{m} \mathbf{B}^T \mathbf{u}_2 \mathbf{u}_1^T \mathbf{B} \right) \mathbf{\Sigma} \right] = \sum_{j=1}^{m} \mathbf{u}_1^T \mathbf{B} \mathbf{\Sigma} \mathbf{B}^T \mathbf{u}_2,$$

as given in Equation (7). Furthermore,

$$\begin{aligned}
\mathrm{Cov}\left(\mathbf{u}_1^T \mathbf{D} \mathbf{u}_2, \mathbf{u}_3^T \mathbf{D} \mathbf{u}_4\right) &= \mathrm{tr}\left[\left(\sum_{j=1}^{m} \mathbf{B} \mathbf{u}_2 \mathbf{u}_1^T \mathbf{B}^T \right) \mathbf{\Sigma} \left(\sum_{j=1}^{m} \mathbf{B} \left[\mathbf{u}_4 \mathbf{u}_3^T + \mathbf{u}_3 \mathbf{u}_4^T \right] \mathbf{B}^T \right) \mathbf{\Sigma} \right] \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} \mathrm{tr}\left(\mathbf{B} \mathbf{u}_2 \mathbf{u}_1^T \mathbf{B}^T \mathbf{\Sigma} \mathbf{B} \mathbf{u}_4 \mathbf{u}_3^T \mathbf{B}^T \mathbf{\Sigma} + \mathbf{B} \mathbf{u}_2 \mathbf{u}_1^T \mathbf{B}^T \mathbf{\Sigma} \mathbf{B} \mathbf{u}_3 \mathbf{u}_4^T \mathbf{B}^T \mathbf{\Sigma} \right) \\
&= \sum_{i=1}^{m} \sum_{j=1}^{m} \left(\mathbf{u}_1^T \mathbf{B}^T \mathbf{\Sigma} \mathbf{B} \mathbf{u}_4 \mathbf{u}_3^T \mathbf{B}^T \mathbf{\Sigma} \mathbf{B} \mathbf{u}_2 + \mathbf{u}_1^T \mathbf{B}^T \mathbf{\Sigma} \mathbf{B} \mathbf{u}_3 \mathbf{u}_4^T \mathbf{B}^T \mathbf{\Sigma} \mathbf{B} \mathbf{u}_2 \right),
\end{aligned}$$

which is equivalent to Equation (8).