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

May 28, 2015

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.

Abstract

Meta-analyses often include studies that report multiple effect sizes based on a common pool of subjects, or that report effect sizes from several samples that were treated with very similar research protocols. The inclusion of such studies introduces dependence among the effect size estimates. When the number of studies is large, robust variance estimation (RVE) provides a method for pooling dependent effects, even when information on the exact dependence structure is not available. When the number of studies is small or moderate, however, test statistics and confidence intervals based on RVE can have inflated Type I error. This paper describes and investigates several small-sample adjustments to F-statistics based on RVE. Simulation results demonstrate that one such test, which approximates the test statistic using Hotelling's T^2 distribution, is level-α and uniformly more powerful than the others. An empirical application demonstrates how results based on this test compare to the large-sample F-test.

Keywords: cluster robust, meta-analysis, sandwich estimator, F-test, bias-reduced-linearization

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.

Dependent effect sizes are most often handled by averaging them into a single, synthetic effect size for each study. This approach allows the estimation of an average effect across studies (using univariate meta-analysis methods), but it leads to difficulties when researchers are interested in the relationship between effect sizes and moderator variables that vary within a single study. When such moderator relationships are of interest, 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, consider a synthesis of experimental studies, some of which report separate effect sizes for writing and reading measures. In order to estimate an overall average effect size, the writing and reading measures within each study would be averaged, then pooled together across studies. However, one might also be interested in determining if the findings differ for writing versus reading. In order to examine this, separate univariate meta-analyses could be conducted by combining only the writing (or reading) measures across studies. While this approach allows for average writing and reading measures to be calculated (thus providing descriptive information regarding their

differences) it does not allow for statistical comparisons to be conducted because the effects are correlated.

A more principled approach to handling dependent effect sizes is to apply a multivariate meta-analytic model. If the effect sizes reported in the same study are conditionally independent – as occurs when they are calculated from different groups of individuals – then a hierarchical linear modeling approach can be used (Raudenbush & Bryk, 2002). However, when at least some effect sizes are collected on the same individuals, the assumptions of this hierarchical approach are not met and full multivariate meta-analysis is required. The multivariate meta-analysis model requires good estimates of the correlation between dependent effect sizes (Gleser & Olkin, 2009), which typically are not available to the meta-analyst. For example, in a meta-analysis that includes both writing and reading measures, one would need to know the correlation between the measures in order to estimate the dependence between the effect sizes, 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 that addresses the challenges of handling dependent effect size estimates is robust variance estimation [RVE] (Hedges, Tipton, & Johnson, 2010). RVE is appealing because it allows for the inclusion of multiple, correlated effect sizes in a single meta-analysis, 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 writing and reading measures, as in the previous example), 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 & Pigott, 2014). To address this shortcoming, 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. This correction has enabled RVE to be implemented in meta-analyses with as few as 5 studies.

In addition to single-coefficient tests, multiple-contrast hypothesis tests are also common in meta-analysis. In univariate meta-analysis, these tests are used to assess model fit, to compare nested models (i.e., incremental tests), and to determine if effect sizes are moderated by categorical variables with multiple levels. Such tests are all based on Q-statistics (e.g., Q-between, Q-model), which are compared to a chi-squared reference distribution. However, no such tests are currently available in RVE. The aim of this paper is therefore to develop

procedures for testing multiple-contrast hypotheses, with a focus on finding a test that performs well in small samples. Although the tests that we develop are all generalizations of the methods developed in Tipton (2014) for small sample t-tests, the extension to multiple-contrast hypotheses involves non-trivial complications due to the multivariate features of the contrasts being tested.

In the remainder of the paper, we consider five possible multiple-contrast hypothesis tests for RVE, all of which are based on approximations to the distribution of a Q-statistic (i.e., a Wald test statistic). We develop each approximation 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. After reviewing RVE and the literature on t-tests, we present new analytic work describing several potential approximations. We then describe a large simulation study that compares these potential solutions. Finally, 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) and discuss implications for meta-analytic practice.

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 models 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 $\mathbf{y} = \left(\mathbf{y}_1^T, ..., \mathbf{y}_m^T\right)^T$, $\mathbf{X} = \left(\mathbf{X}_1^T, ..., \mathbf{X}_m^T\right)^T$, $\boldsymbol{\varepsilon} = \left(\boldsymbol{\varepsilon}_1^T, ..., \boldsymbol{\varepsilon}_m^T\right)^T$, and $\boldsymbol{\Sigma} = \text{diag}\left(\boldsymbol{\Sigma}_1, ..., \boldsymbol{\Sigma}_m\right)$. 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 could be written as $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$, 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.

Let $\mathbf{W} = \text{diag}\left(\mathbf{W}_1, ..., \mathbf{W}_m\right)$ be a block-diagonal matrix of weights, which for the time being we will treat as arbitrary. Given this set of weights, 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 $\boldsymbol{\Sigma}_j$. If the structure of $\boldsymbol{\Sigma}_j$ is fully known and the weights are defined so that $\mathbf{W}_j = \boldsymbol{\Sigma}_j^{-1}$ for $j = 1,...,m$, then the variance of $\mathbf{b}$ reduces to simply $\mathbf{M}$. However, in practice it is often difficult to meet the assumption that the covariance structure is correctly modeled. The crucial advantage of RVE is

that it provides a means to estimate the variance of **b** without relying on this stringent assumption.

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 (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 (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. Moreover, it is not always clear when the sample is sufficiently large to trust the asymptotic approximation.

Choice of weights

Thus far, we have introduced RVE for a general set of weights $\mathbf{W} = \text{diag}(\mathbf{W}_1, \mathbf{W}_2, \ldots, \mathbf{W}_m)$. While the RVE approach makes no requirements on these weights, the use of approximately inverse variance weights can improve the efficiency of the meta-regression estimates. However, when the correlation structure is unknown (thus necessitating RVE), it is not possible to calculate exact inverse-variance weights. Noting that the only role for weights is efficiency gains, Hedges et al (2010) provided two simplified options for weighting based upon "working" covariance models: "correlated effects" and "hierarchical effects." Correlated effects are used when primary studies report multiple outcomes measured on the same individuals. For example, a primary study might report measures of both writing and reading performance, or might report outcomes from multiple follow-up times. Hierarchical effects are used when outcomes are collected on different groups of individuals, but those groups may share some common influences. For example, a primary study might report the results of two separate experiments conducted in the same lab, with the same subject pool and same laboratory protocols. In practice, it is common for both types of dependence to occur within the same study; the analyst then typically chooses the "working" model that matches the most common type of dependence structure in the studies to be meta-analyzed. Exact forms for suggested working models and weighting matrices can be found in Appendix A.

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 $\mathbf{b}$ 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 changes to the adjustment matrices and to 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 $\mathbf{\Sigma}_1,...,\mathbf{\Sigma}_m$. Tipton proposed using 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 the nominal level; 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 are approximated by

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

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 an extensive simulation, Tipton (2014) found that the Satterthwaite degrees of freedom correction led to larger improvements in Type I error than using adjustment matrices alone. Her results also showed that t-tests based on combining both corrections are level-α so long as the degrees of freedom (ν_s) are larger than 4 or 5. When ν_s is smaller than 4 or 5, the empirical size of the test can be higher (or lower) than the stated α-level, leading to the suggestion that p-values not be reported for tests when ν_s is this small.

Tipton also found that what was considered a "small" sample depended more on the degrees of freedom than on the number of studies. Depending upon features of the covariates, the degrees of freedom could vary wildly even if the sample size (m) is held constant. For example, consider a meta-regression analysis based on $m = 40$ independent studies, where interest is in testing a single factor that has two levels (A and B). If the studies are divided evenly across the levels, the degrees of freedom for that factor will be moderate (i.e., $\nu_s = m - p = 38$). In contrast, if most of the studies are in one level (e.g., 36 in level B), then the degrees of freedom will be considerably smaller, indicating a smaller effective sample ($\nu_s = 5.21$; see Table 2 of Tipton, 2014). Because it is not possible to know if the sample is "small" without calculating these degrees of freedom, Tipton argued that the small-sample t-test should be used in all RVE analyses.

Small sample corrections to F-tests

The goal of this paper is to develop a method for conducting multiple-contrast hypothesis tests based on RVE. As with the t-test developed by Tipton (2014) for single-parameter hypothesis tests, we seek a test that performs well in small samples, thus enabling broad application in meta-analysis. Following the degrees of freedom correction suggested by Hedges et al (2010), a simple, ad hoc test would be to use the test statistic $F = Q / q$, compared to an F-distribution with q and $m - p$ degrees of freedom. As Tipton (2014) found for the analogous version of the t-test, our simulation study will show that this "naïve F-test" performs adequately only in very particular circumstances. One reason for its poor performance is that it uses the same degrees of freedom regardless of the contrasts being tested. Like the t-test proposed by Tipton (2014), a more principled test would 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 the distribution of the single variate V_{ss}^{R}.

In order to identify potential approaches to approximating the distribution of Q, we first reviewed the literature on robust variance estimation outside of the meta-analysis context. Simpler forms of robust variance estimation are used in ANOVA, multivariate ANOVA, and multiple regression when there is concern about heteroskedasticity; robust variance estimation is also used in connection with generalized estimating equations. In each of these areas, previous research has investigated the small sample properties of hypothesis tests based on special cases of RVE. In this paper, we consider two broad strategies that draw on methods that perform well in these special cases. Importantly, in all of the methods to be considered, we will use the same adjustment matrices as implemented by Tipton (2014), the form of which is given in Equation

(4). These adjustment matrices are derived under a working covariance model with approximate inverse-variance weights.

The first strategy is to approximate the sampling distribution of $\mathbf{CV}^{R}\mathbf{C}^{T}$ using a Wishart distribution, which leads to test statistics that approximately follow Hotelling's T^2 distribution (a multiple of an F distribution). Versions of this approach are found in the literature on heteroskedastic ANOVA (Zhang, 2013); MANOVA with unequal variance-covariance structures (Krishnamoorthy & Yu, 2004; Nel & van der Merwe, 1986; Zhang, 2012); and GEE models (Pan & Wall, 2002).

The second strategy uses the spectral-decomposition of $\mathbf{CV}^{R}\mathbf{C}^{T}$ to approximate the distribution of Q as a sum of independent univariate random variables. Using this decomposition, two specific approaches are considered. One approach, which has been considered in the literature on small-sample corrected hypothesis tests for linear mixed models (Fai & Cornelius, 1996), uses a Satterthwaite approximation with estimated degrees of freedom. The other approach involves transforming the independent univariate random variables so that their sum more closely follows a chi-squared distribution. This second approach has been studied for the special case of heteroskedasticity robust variance estimation in ANOVA (Alexander & Govern, 1994) and in multiple regression (Cai & Hayes, 2008).

In the remainder of this section, we develop five possible approaches based upon these two broad strategies, including three tests based on Hotelling's T^2 distribution and two tests based on the spectral decomposition approach. While these approaches take inspiration from the broader literature, none of the estimators have been developed or studied for the problem of robust variance estimation with correlated, dependent errors, as occur in multi-variate meta-analysis.

Moments of $\mathbf{V}^{\mathbf{R}}$

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 the later presentation. 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 also 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 $q \times 1$ 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 **D** will need to be estimated. One approach would be to estimate the $\mathbf{\Sigma}_j$ 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 that was 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 empirical estimates 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 $\mathbf{W} = \mathbf{\Sigma}^{-1}$, $\mathbf{V}^R$ 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 $\mathbf{\Omega} = \mathbf{C}\mathbf{M}\mathbf{C}^T$. 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 degrees of freedom η that approximate well the distribution of **D**.

Let d_{st} denote the entry in row s and column t of **D** and let $I(s = t)$ be the indicator function that equals one if $s = t$ 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

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

An alternative, which Pan and Wall (2002) mentioned 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 ANOVA and MANOVA, 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 (8) 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 AHA, AHB, and AHZ (where "AH" refers to an "Approximate Hotelling" test), 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 AH degrees of freedom evaluate to $\eta = 2/\text{Var}\left(d_{11}\right)$ and the resulting tests are equivalent to the t-test with Satterthwaite degrees of freedom, as described by Tipton (2014). However, the AHA, AHB, and AHZ 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. Their method is based on the eigen-decomposition of the estimated covariance matrix and ignores 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

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

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 Equation (8) 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

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

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

$$d = \frac{E_Q^2 (q - 2) + 2 q V_Q}{q E_Q \left(V_Q + E_Q^2 \right)}, \quad n_F = 4 + \frac{2 E_Q^2 (q + 2)}{q V_Q - 2 E_Q^2}, \tag{17}$$

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

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

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 involves 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 a variety of other possible corrections. For example, it is possible to combine the Hotelling's T^2 approximation with the eigen-decomposition approach; to combine the eigen-decomposition approach with a simulation-based reference distribution; 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 describe them further. 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 any condition studied. We used 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 were 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 led 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 (i.e., 10, 15, 20, 30, 40, 60, 80, 100). 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 of extreme leverage (e.g., most values between 0 and 30, with a couple of values as high as 100) have much smaller degrees of freedom, even when the number of studies was moderate to large.

The simulation used five covariates representing a range of variable types often encountered in practice, including two binary and three continuous covariates. Exact values of

the covariate vectors are available in the online supplementary materials. 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 allowed 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 assumed 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 were not equivalent; instead, the two models diverged as ρ and I^2 increased, which allowed us to examine the extent to which the tests are robust to model misspecification. This feature of the simulation also emulates

what the analyst will encounter in practice, when the true data-generating model is unknown and the working model will seldom be correctly specified.

This combination of parameters resulted in a total of 120 conditions for the data-generating model under study (8 sample sizes × 3 values for ρ × 5 values for I^2). For each combination of simulation conditions, we fit several different regression specifications and tested several contrasts involving different combinations of the covariates. The combinations of regression specifications fell 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 5 covariates, we ran "subset" tests of all possible combinations of q = 2, 3, or 4 of the covariates (for a total of 25 unique hypotheses tested). For each simulated meta-analysis and each of the five tests, we recorded the p-value for each combination of model specifications and hypotheses. In total, each of the five tests was therefore evaluated under 6,120 unique conditions (120 unique data-generating models × 51 unique model specifications and hypotheses). For each of the tests under each of the conditions, we determined the Type-I error rates corresponding to α = .01, .05, and .10 by calculating the proportion of p-values less than α, across 10,000 replications. Monte Carlo margins of error are therefore approximately 0.19% for α = .01, 0.42% for α = .05, and 0.59% for α = .10. In the figures that follow, the dashed lines depict upper bounds on the empirical error rate of a level-α test.

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 both terms of both absolute and relative Type I error. The final two findings focus on the roles of degrees of freedom and model misspecification. For all but the last 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 the online supplementary materials.

Finding 1: Naïve F-test performance

The naïve F-test compares $F = Q / q$ to an F-distribution with q and $m - p$ degrees of freedom, where q is the number of contrasts being tested, m is the total number of studies in the meta-analysis, and p is the number of parameters in the meta-regression model. This test is the multivariate analogue to the small sample correction 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 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

for α = 0.10) or even larger (e.g., 0.02 for α =0.01). Interestingly, when *q* is small and the sample is small (*m* < 30), the naïve F-test tends to do better when *m* is smaller rather than larger (and particularly for α = 0.01). This trend occurs because the degrees of freedom correction exerts a larger penalty when *m* is very small. Of course, even this relatively large penalty is inadequate, as the Type I error remains far above nominal in all cases. Furthermore, the performance of the Naïve F degrades as the number of parameters being tested increases, with the largest Type I error rates occurring when *q* = 5 and the number of studies is small.

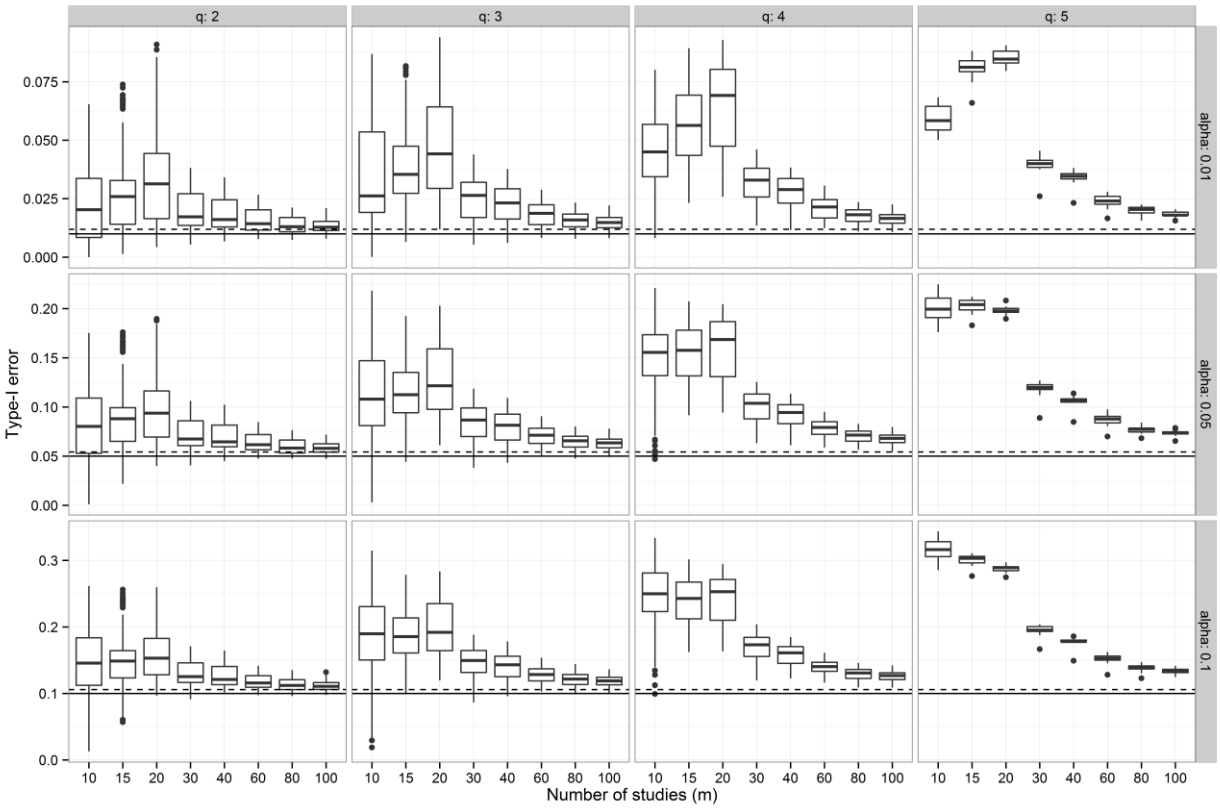


FIGURE 1: Type I error of naïve F-test by number of studies and values of *q*.

Note: Solid lines indicate the stated α level; dashed lines indicate bounds for simulation error.

Finding 2: AH tests are nearly level-α

Figure 2 presents the range of 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 $\alpha = 0.05$, although we note that the trends are very similar for $\alpha = 0.01$ and 0.10. 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.

Figure 2 reveals several trends. First, Type I error for the EDF and EDT tests typically approach the nominal values from above, whereas the AHA, AHB, and AHZ 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 AHA, AHB, and AHZ 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 cover a wide range of values across the parameters and hypothesis specifications under study (as indicated by the long whiskers on each box). Because it is not possible to know a priori in which design condition a particular analysis will fall, it makes more sense to compare the *maximum* Type I error observed across tests. While the EDT and EDF tests have Type I error rates that are closest to nominal *on average*, they also exhibit error rates that are far above nominal under a large number of design conditions that cannot be identified a priori. In comparison, the AHA, AHB, and AHZ tests are typically more conservative and are also nearly always level-α, with a maximum error rate of 0.059 across all conditions studied. In describing further trends, we therefore focus only on the three AH tests.

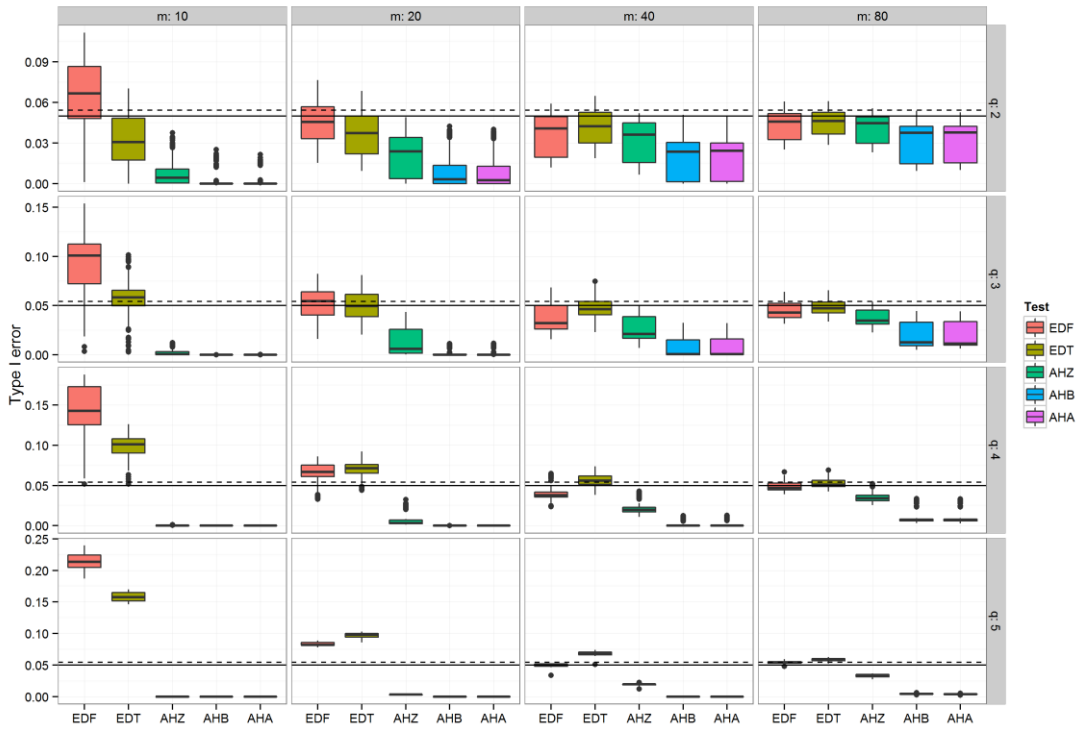


FIGURE 2: Type I error for α = 0.05 of 5 alternative tests

Note: Solid lines indicate the stated α level; dashed lines indicate bounds for simulation error.

Finding 3: Type I error of AHZ is closest to nominal

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. Having established that AHA, AHB, and AHZ are all level-α under nearly all of the conditions studied, we examined which of the tests have empirical Type I error rates closest to the stated values. To address this question, we compared the Type I error rates of each pair of the three tests for each of the parameter combinations in the simulation. We do not depict the comparisons involving AHA because its performance was clearly worse than the other two tests. Figure 3 plots the Type I error rates of AHZ against those of AHB for each of the 6,120 unique conditions in the simulation; separate panels correspond to each α level under study. In this figure, points above

the diagonal line represent tests and parameter combinations for which the Type I error of the AHZ test is closer to α than that of the AHB test.

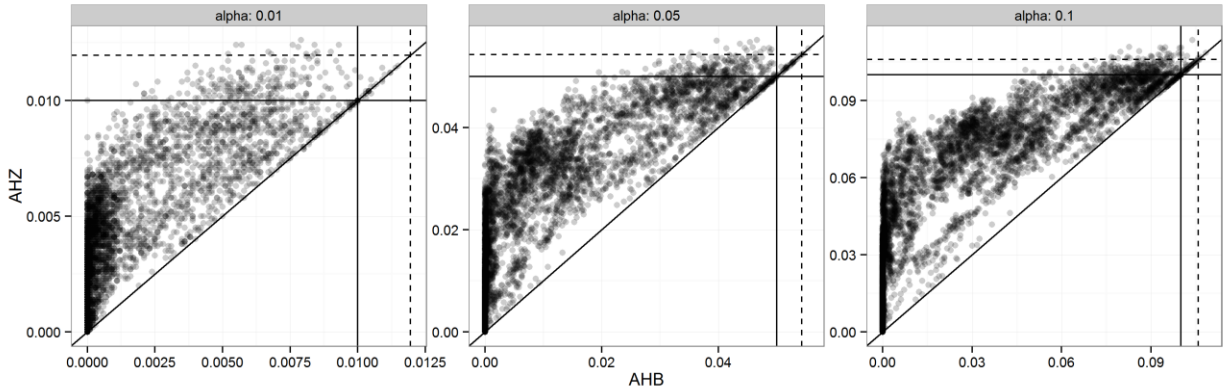


FIGURE 3: Type I error for α = 0.05 for AHB versus AHZ

Note: The solid and dashed horizontal lines indicate the stated α level and simulation error bound, respectively, for the AHZ test; the vertical lines indicate the same quantities for the AHB test. Points above the diagonal line correspond to conditions where the Type I error rate of AHZ is closer to nominal than that of AHB.

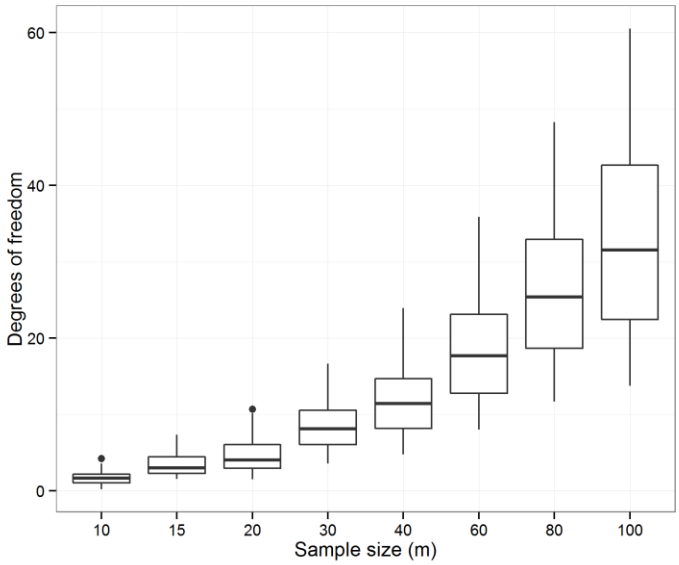
Figure 3 illustrates that the maximum Type I error of the AHB test is slightly smaller than that for the AHZ test, but these differences are minimal. It is also apparent that the Type I error of the AHZ test is closer to the stated α-level for every parameter combination, although both tests tend to be conservative. For example, when α = 0.05 and the Type I error of AHB is nearly 0.00, the error of AHZ ranges from 0.00 to 0.04. Furthermore, the median Type I error for the α = 0.05 test is 0.0254 for AHZ but only 0.0044 for AHB; in over 79% of cases, the Type I error of AHB is less than the median error of AHZ. In summary, AHZ is the most accurate level-α test across nearly all combinations of data-generating conditions and hypothesis specifications 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 on the degree of balance or

leverage in the covariate, than to the sample size alone. We observed a similar pattern for multiple-contrast tests. Figure 4 illustrates the relationship between the number of studies and the degrees of freedom for the AHZ 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. Put another way, degrees of freedom smaller than 20 (indicating "small" effective samples) can be observed across a wide range of actual sample sizes (i.e., from m = 40 to 100). Just as Tipton (2014) concluded, this means that 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.

FIGURE 4: Degrees of freedom of AHZ test by number of studies (m)



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). While the other parameters under study (e.g., the number of independent studies, the number of effect sizes per study, features of the covariates) are known to researchers, these two parameters are unknown. Recall that in RVE, and throughout this paper, a "working" model of the covariance structure is required for calculation of weights, adjustment matrices, and development of the degrees of freedom. In the simulation study, the tests were always constructed based on the "correlated effects" working model wherein it was assumed that $\rho = 1$ and $I^2 = 0$ (i.e., fixed effects weights). An important question, then, is how well the AHZ test performs when this assumed structure is incorrect, as is likely to occur in practice.

In order to investigate the robustness of the AHZ test to misspecification of the working model, we simulated data using a wide range of values of ρ and I^2 and then compared the Type I error. In conducting this 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 when fixed effects weights (i.e., assuming $I^2 = 0$) were used, the AHZ test was always level-α so long as the actual I^2 was less than 1/2. Values of I^2 above ½ correspond to a high degree of working model misspecification, particularly given that the default in most RVE analyses is to use random effects weights. Additionally, even when the working model is badly mis-specified (true $I^2 > ½$), the test still contains Type I error to a much greater extent than the naïve-F, EDF, or EDT tests.

Application

To illustrate the performance of the AHZ test in practice, as well as to better understand differences between it and the benchmark of the naïve F-test, we applied 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 that measured school completion or school dropout outcomes and that were reported between 1985 and 2010. The review included 152 independent reports of experimental or quasi-experimental studies, which together included 317 independent samples. Multiple outcome measures were reported for some of the samples, resulting in a total of 385 log-odds-ratio effect sizes.

In the original meta-analysis, a meta-regression model was provided predicting the log-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, level of attrition, an independent evaluator scale), participant characteristics (% male, %white, average age), and program characteristics (e.g., implementation quality, indicators of program format). We illustrate the use of multi-parameter hypothesis tests using the same model specification, though our analysis deviates from Wilson et al. in one important respect. The original analysis treated the 317 samples as independent (ignoring the nesting of samples within studies) and used RVE with a "correlated effects" working model. In contrast, our re-analysis treated the 152 studies as independent and employed RVE with a "hierarchical effects" working model. We took this approach because it led to a smaller sample size (though still very large compared to most meta-analyses). Additionally, we repeated the analysis using a subset of 32 studies, which is a typical sample size for meta-analyses in education and the social sciences (Ahn et al., 2012; Polanin &

Pigott, 2014). The subset included the 32 studies that reported 3 or more effect sizes (whether from independent samples or multiple measures).

For each of these re-analyses, we tested five separate multiple-contrast hypotheses about the meta-regression 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 illustration we treated these covariates as categorical and modeled them using indicator variables. Notably, the original study did not report hypothesis tests for any of the covariates because multiple-contrast hypothesis testing procedures were not available in RVE software at the time.

Table 1 reports the results of the naïve F and AHZ tests for each of the five hypotheses. The upper panel includes results based on the small subset of 32 of the studies; the lower panel includes results based on all 152 studies. While not illustrated here, the results found in the second panel are qualitatively similar to those found when using the "correlated effects" approach applied by Wilson and colleagues.

Several notable patterns are apparent from Table 1. First, the F-statistic for the AHZ test is always smaller than that for the Naïve F test, reflecting the multiplicative correction involved. Second, the degrees of freedom for the AHZ test are usually smaller – and sometimes much smaller – than those for the naïve F-test, and vary depending on the hypothesis being tested. Such variation reflects imbalances and skewness in covariate distributions. For example, across both panels, evaluator independence has the smallest degrees of freedom of any of the hypotheses (i.e., $df = 6$; $df = 17$); these low degrees of freedom are due to unequal allocation of effect sizes to the four categories in both the subset (i.e., 0, 10, 25, and 195 effect sizes) and full samples (i.e., 6, 33, 43, and 303 effect sizes). Third, the combination of smaller F-statistics and degrees of freedom results in p-values that are uniformly smaller for the AHZ test compared to the naïve F-test. For some tests and conditions these p-value differences do not impact conclusions (as with implementation quality in the full sample) while in other cases the

Table 1: Comparison of Naïve-F tests and AHZ tests in Dropout Example

			Naïve F test*		AHZ test		
Sample	**Factor**	**q**	**F**	**p-val**	**F**	**df**	**p-val**
32 studies							
	Study design	2	3.19	0.081	2.93	11.0	0.096
	Outcome measure	3	1.05	0.407	0.84	7.7	0.512
	Evaluator independence	2	0.32	0.735	0.26	4.6	0.781
	Implementation quality	2	4.02	0.049	3.69	11.0	0.059
	Program format	3	1.19	0.357	0.98	9.1	0.444
152 studies							
	Study design	2	0.23	0.796	0.22	42.9	0.800
	Outcome measure	3	0.91	0.436	0.84	21.5	0.488
	Evaluator independence	3	3.11	0.029	2.78	16.8	0.073
	Implementation quality	2	14.15	<0.001	13.78	36.9	<0.001
	Program format	3	3.85	0.011	3.65	37.5	0.021

Notes:
df = degrees of freedom. p-val = p-value.
** The Naïve F test uses 11 degrees of freedom in the small sample (m = 32) and 130 degrees of freedom in the full sample (m = 152).*

inferences change. For example, in the small sample, the p-value for *implementation quality* is less than .05 for the naïve F-test but greater than .05 for the AHZ test. Similarly, the inference for evaluator independence based on the full sample of 152 studies is sensitive to whether the small-sample corrections are employed.

Discussion and 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 strong assumptions about either the error distribution or covariance structure has allowed analysts to summarize relationships across all collected effect sizes, instead of artificially reducing the data to fit the statistical method. However, while the method performs well in large-samples, practical applications of RVE are limited by the fact that hypothesis tests based on RVE have inflated Type I error rates in small and moderate samples.

In this paper, we developed and compared five possible multi-contrast hypothesis tests for use in RVE in meta-analysis. We also examined the performance of the asymptotic chi-squared test and a simple, ad hoc test: the naïve F test. The simulation results indicate that the asymptotic test and the naïve F-test perform 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 its degrees of freedom do not account for covariate features, such as the degree of balance or leverage, that impact the Type I error of the test. Given the typical sizes of meta-analyses in the educational and social sciences, the chi-squared test and naïve F-test should not be used in practice. Instead, a small-sample corrected test should always be employed for multiple contrast hypothesis tests based on RVE.

While the AHA, AHB, and AHZ tests presented here all control the Type I error rate—even in very small samples—the AHZ test clearly outperforms the others. Except under conditions of extreme model misspecification, the AHZ 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 highly mis-specified models, its maximum Type I error was only slightly beyond simulation error (e.g., 0.0594 when α = .05). It is important to highlight, too, that we would seldom expect to encounter such extreme model misspecification in practice because the default approach in RVE is to estimate the degree of between study heterogeneity, rather than assuming $I^2 = 0$.

A further advantage of the AHZ 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 the tests behave differently, note that the dimension of the test (q) impacts both the test-statistic itself and the reference distribution: as q increases, the multiplier $[(\eta - q + 1)/\eta]$ and the denominator degrees of freedom $(\eta - q + 1)$ both decrease. The 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.

The example presented in the previous section demonstrates that using a test that appropriately controls Type I error – here AHZ – can have two implications for the findings of a meta-regression analysis. First, the estimated degrees of freedom in a meta-regression model using AHZ can vary considerably from covariate to covariate, and are typically far below those

used by the naïve F-test (i.e., $m - p$). When a covariate is extremely imbalanced (e.g., evaluator independence in the example), the degrees of freedom can be particularly small. This behavior highlights that a small-sample test should *always* be preferred, even in what may seem to be a very large meta-analysis, Second, the impact of the degrees of freedom can lead to different conclusions.

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 a variety of tests that have good small-sample performance, 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, developing saddlepoint approximations to Wald-type test statistics for multiple contrast hypotheses is not straight-forward.

Like Tipton (2014) we have also employed a working model approach to 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. 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.

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. Work by Tipton (2013) and Williams (2012) suggests that RVE performs similarly for log-odds-ratios, log-risk-ratios, risk differences and regression coefficients, but no work to date has investigated the performance of small sample tests with families of effect sizes other than standardized mean differences. Additionally, our examination has focused exclusively on Type I error, and has not considered power. Future work will need to examine the power of the recommended tests under non-null alternative hypotheses. It would be particularly useful to compare the RVE approach (with the AHZ test) with other approaches such as fully model-based multivariate meta-analysis. Finally, we note that the tests developed here are not limited in application to meta-analysis. In future work we plan to investigate the performance of these tests under other models that involve 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 $\mathbf{\Psi}_j = \rho \mathbf{1}_j \mathbf{1}_j^T + (1-\rho)\mathbf{I}_j$ 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 $\tau^2 \mathbf{1}_j \mathbf{1}_j^T + \left(2 / n_j\right) \mathbf{\Psi}_j$, 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

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.

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.

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

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

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*.

Muirhead, R. J. (1982). *Aspects of Multivariate Statistical Theory*. New York, NY: John Wiley & Sons.

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. R., & Pigott, T. D. (2014). The use of meta-analytic statistical significance testing. *Research Synthesis Methods*, (August 2014), n/a–n/a. doi:10.1002/jrsm.1124

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

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

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.

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 $\mathbf{V}^{\mathbf{R}}$

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

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

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

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

as given in Equation (7). Furthermore,

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

which is equivalent to Equation (8).