Abstract Title Page

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

Authors and Affiliations:
Elizabeth Tipton, *Teachers College, Columbia University*
James E. Pustejovsky, *University of Texas at Austin*

Abstract Body

Background / Context:

Randomized experiments are commonly used to evaluate the effectiveness of educational interventions. The main focus in randomized experiments is often on the average treatment effect across all participants in the study, yet when the effectiveness of an intervention varies, a single summary effect may be of limited utility. Instead, understanding what works for whom, when, and where matters. The questions of this conference – regarding the optimal age for an intervention and the effect of the intervention on outcomes at different time points – are inherently questions of this type. Questions regarding moderation can be addressed in several different ways: 1) through the inclusion of multiple cohorts of students (e.g., K and 3rd graders) or through longitudinal designs (e.g., outcomes at 1, 2, and 3 years) within individual experiments; 2) through the accumulation of evidence across studies, synthesized using meta-analysis. This paper focuses on this second approach, which we argue is particularly important because individual studies are rarely powered adequately to detect treatment effect interactions.

Over the past 30 years, meta-analysis has been widely used in education research. 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, and Johnson, 2010); to date, this method has been used in over 50 meta-analyses in as diverse fields as ecology, education, psychology, and intervention studies. An advantageous feature of RVE is that it does not require information on the true correlation structure of the estimates within a given study, which are rarely reported in practice.

The statistical theory behind the robust variance estimation method is asymptotic; in large-enough samples, it has been shown to be an unbiased estimator of the true sampling variance. In small samples, however, the estimator can be biased and the Type I error rate of tests based upon the RVE method can be much too liberal (Hedges et al, 2010; Tipton, 2013). This represents a serious limitation, given that as many as half of recent meta-analyses in education contained fewer than 40 studies (Ahn, Ames, & Myers, 2012). 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.

Purpose / Objective / Research Question / Focus of Study:

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, studies might be conducted on students of different ages, resulting in a covariate *grade-level* with three levels: "elementary", "middle", or "high" school. In order to answer the questions "Does the effectiveness of this intervention vary in relation to age?" an F-test would need to be conducted. Currently, it is not possible to conduct F-tests of this type in RVE. Lacking valid testing methods, researchers are left to either rely on asymptotic approximations, which can be seriously in error, or to cobble together ad-hoc methods, such as using RVE with all effect sizes to conduct t-tests, but ANOVA with study-aggregated effect sizes to conduct F-tests.

Significance / Novelty of study:

Drawing on work that addresses related, simpler problems and special cases of cluster-robust variance estimation, we develop three small-sample tests based on different approximations to the distribution of a robust Wald test statistic. In the remainder, we describe our modeling assumptions, proposed tests, and some initial simulation result. These approximations are drawn from a wide array of areas within statistics, ranging from econometrics to survey sampling. The paper presents both new analytic work describing these small-sample corrected test statistics and the results of a large simulation study that compares these potential solutions, as well as a discussion of the implications of our findings for practice.

Statistical, Measurement, or Econometric Model:

We develop the methods under the general meta-regression model

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i$$

where $\mathbf{y}_i$ is a vector of k_i effect size estimates from study i, $\mathbf{X}_i$ is an $k_i \times p$ matrix of covariates, and $\boldsymbol{\varepsilon}_i$ is a vector of (potentially correlated) errors with $\text{Var}(\boldsymbol{\varepsilon}_i) = \boldsymbol{\Sigma}_i$. Importantly, the structure of $\boldsymbol{\Sigma}_i$ is typically unknown, and may involve a combination of several correlation structures.

Let $\mathbf{W}_i$ be a $k_i \times k_i$ weighting matrix based on a "working" covariance model (see Tipton, 2014 for a discussion of how to choose the working model). The WLS estimator of $\boldsymbol{\beta}$ is

$$\mathbf{b} = \mathbf{M}\left(\sum_{i=1}^{m}\mathbf{X}_i^T\mathbf{W}_i\mathbf{y}_i\right), \text{ where } \mathbf{M} = \left(\sum_{i=1}^{m}\mathbf{X}_i^T\mathbf{W}_i\mathbf{X}_i\right)^{-1}.$$

Note that if the working model is correct, as is typically assumed in univariate meta-analysis, then $\mathbf{W}_i = \boldsymbol{\Sigma}_i^{-1}$ for $i = 1,\ldots, m$ and $\text{Var}(\mathbf{b}) = \mathbf{M}$. Following Tipton (2014), we employ an unbiased form of the robust variance estimator developed by McCaffrey, Bell, and Botts (2001), given by

$$\mathbf{V}^R = \mathbf{M}\left[\sum_{i=1}^{m}\mathbf{X}_i^T\mathbf{W}_i\mathbf{A}_i\mathbf{e}_i\mathbf{e}_i^T\mathbf{A}_i^T\mathbf{W}_i\mathbf{X}_i\right]\mathbf{M},$$

where $\mathbf{e}_i = \mathbf{y}_i - \mathbf{X}_i\mathbf{b}$. The $k_i \times k_i$ matrices $\mathbf{A}_i$ are chosen such that if the weights are truly inverse-variance (i.e. $\mathbf{W}_i = \boldsymbol{\Sigma}_i^{-1}$), then the variance estimator is exactly unbiased: $\text{E}(\mathbf{V}^R) = \mathbf{M}$. Note that Tipton (2014) derives the appropriate $\mathbf{A}_i$ for the two commonly used models in RVE – hierarchical and correlated effects.

In this paper, we are interested in testing the null hypothesis H_0: $\mathbf{C}\boldsymbol{\beta} = \mathbf{0}$ for the $q \times p$ contrast matrix $\mathbf{C}$. For example, an omnibus test might be written H_0: $\beta_1 = \beta_2 = \ldots = \beta_{p\text{-}1} = 0$, or a test of a categorical variable with three levels might be written H_0: $\beta_1 = \beta_2 = 0$. A Wald-type test statistic for the multi-parameter hypothesis H_0 is given by

$$Q = \mathbf{b}^T\mathbf{C}^T\left(\mathbf{C}\mathbf{V}^R\mathbf{C}^T\right)^{-1}\mathbf{C}\mathbf{b} = \mathbf{z}^T\mathbf{D}^{-1}\mathbf{z},$$

where $\mathbf{z} = \left(\mathbf{C}\mathbf{M}\mathbf{C}^T\right)^{-1/2}\mathbf{C}\mathbf{b}$ and $\mathbf{D} = \left(\mathbf{C}\mathbf{M}\mathbf{C}^T\right)^{-1/2}\mathbf{C}\mathbf{V}^R\mathbf{C}^T\left(\mathbf{C}\mathbf{M}\mathbf{C}^T\right)^{-1/2}$. As m increases, the distribution of Q approaches that of a chi-squared random variate with q degrees of freedom (Wooldridge, 2002). However, the asymptotic distribution may provide a very poor approximation when m is small, leading to actual type I error rates far in excess of the nominal level. Furthermore, it is often unclear when one has a sufficient sample of studies to trust the asymptotic test; as Tipton (2014) shows with t-tests, the degrees of freedom for the associated tests depend not only on the number of studies (m), but also on features of the covariates.

Methods

In this paper, we consider three adjusted tests for H_0 that have improved type I error rates. All three tests are based on quantities derived from the variances and covariances of the entries in the matrix **D** under a working covariance structure. Let d_{st} denote the $(s,t)^{th}$ entry in **D**. Under the working covariance model and assuming that the errors are normally distributed, we can obtain expressions for $\mathrm{Var}\left(d_{st}\right)$ and $\mathrm{Cov}\left(d_{st}, d_{uv}\right)$ based on the fact that d_{st} is a quadratic form. (We omit the expressions here due to space constraints.)

Approximate Satterthwaite correction. The first test employs a Satterthwaite-type correction, wherein we find a multiplier δ and degrees of freedom η such that the first two moments of δQ approximately match those of an $F(q, \eta)$ distribution. We can show that

$$\mathrm{E}(Q) \approx q + T = E_Q$$

$$\mathrm{Var}(Q) \approx 2q + 6T + 2qT + S - T^2 = V_Q$$

where $T = \sum_{s=1}^{q}\sum_{t=1}^{q} \mathrm{Var}\left(d_{st}\right)$ and $S = \sum_{s=1}^{q}\sum_{u=1}^{q} \mathrm{Cov}\left(d_{ss}, d_{uu}\right)$. It then follows that

$$\delta = \frac{E_Q^2(q-2) + 2qV_Q}{qE_Q\left(V_Q + E_Q^2\right)} \quad \text{and} \quad \eta = 4 + \frac{2E_Q^2(q+2)}{qV_Q - 2E_Q^2}.$$

***T*^{2} approximation.** An alternative test can be derived by approximating the distribution of **D** with a Wishart distribution. This approach has been considered previously for special cases of cluster-robust variance estimation, including one-way heteroskedastic ANOVA (Zhang, 2013) and the multivariate Behrens-Fisher problem (Krishnamoorthy & Yu, 2004), but never in the general case. Following Zhang (2013), we derive the approximation by matching the expectation and total variance of **D** to those of a Wishart distribution with v degrees of freedom. Approximating the distribution of **D** by a Wishart implies that Q approximately follows Hotelling's T^2 distribution when H_0 is true. From the properties of the T^2 distribution, it then follows that

$$\frac{v - q + 1}{vq} Q \sim \mathrm{F}\left(q, v - q + 1\right),$$

which can be used to test H_0. For a one-dimensional contrast ($q = 1$), the approximation is exactly equivalent to the Satterthwaite approximation studied by Tipton (2014).

Spectral decomposition and transformation (SDT). Whereas the previous two tests sought approximations to the distribution of Q, the final test involves altering its internal structure, using an approach very similar to one developed by Alexander and Govern (1994) for heteroskedastic one-way ANOVA and by Cai and Hayes (2008) for heteroskedasticity-robust variance estimation (both of which are simpler cases than the cluster-robust variance estimation methods considered here). The SDT test entails first expressing Q as a sum of q squared t-variates, then applying a normalizing transformation to each variate, yielding a test statistic that is closer to chi-square distributed.

Research Design: Simulation Study

In order to evaluate these potential small-sample corrections, as well as to determine

when the methods are needed (i.e., the line between "small" and "large"), we conducted a simulation study. The simulation follows a structure similar to the second study reported in Tipton (2014). This design included a single meta-regression model with 5 covariates – of these, two are constant at the study level (a common feature in meta-analyses) and 3 vary at the effect size level; additionally, one of the covariates has high leverage and another has large imbalances (two conditions that Tipton found have large effects on performance). We simulated correlated standardized mean difference effect sizes, as might be found in randomized experiments that report treatment effects on multiple outcome measures. We used a diagonal weight matrix for the working covariance model; conditions with non-zero correlation between outcomes or non-zero between-study variability therefore represent varying degrees of model misspecification. We considered hypothesis tests for each of the 26 possible subsets of two or more covariates, including the omnibus test of model fit H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$. **Table 1** summarizes the design of the simulation. For each combination of factor levels, we simulated 5,000 meta-analyses.

Findings / Results: Results of Simulation Study

Due to space constraints, we describe the results only for the nominal $\alpha = .05$ level. Furthermore, here we focus on the results in relation to the number of studies in the meta-analysis (m); in the paper, we also investigate the role of the degrees of freedom. **Figure 1** summarizes the range of Type-I error rates for the conventional Wald test across the various combinations of simulation factors; each panel displays the results for the hypothesis tests of the same dimension (q). The error rates of the Wald test far exceed the nominal level, particularly for higher q. Even at the largest sample size considered, the Wald test has unacceptably inaccurate error rates. **Figure 2** plots the range of Type-I error rates for each of the tests described above. All three corrected tests are more accurate than the Wald test. The T^2 test is generally conservative, with error rates that seldom exceed the nominal level. The error rates of the Satterthwaite test sometimes exceed .05, but are mostly quite accurate when m is 30 or larger. The error rates of the SDT test tend to exceed .05 and are more variable than those of the Satterthwaite test. In the paper, we further elucidate the conditions under which the Satterthwaite and T^2 tests are most appropriate, including a discussion of power.

Usefulness / Applicability of Method:

In order to illustrate the usefulness of the method, we include an example based on a meta-analysis by Tanner-Smith and Lipsey (2013). This meta-analysis combined results of randomized-experiments evaluating the effectiveness of brief alcohol interventions on subjects of different ages (i.e., adolescents and young adults) and over multiple time points and waves.

Conclusions:

The results of the simulation study indicate that the asymptotic chi-squared test does not perform well unless the number of studies (m) is very large relative to the dimension of the test (q). In this paper, we investigated several small sample corrections and found two that performed best, in terms of both Type I and II error. Finally, while this paper focuses on the RVE context, we expect that these same techniques will have use in other contexts, including analysis of cluster-randomized trials (using hierarchical linear models) and econometric analysis of panel data.

Appendices

Appendix A. 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.

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

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

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

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*.

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

Wooldridge, J. M. (2002). *Econometric Analysis of Cross Section and Panel Data*. Cambridge, MA: MIT Press.

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 B. Tables and Figures

Table 1. Simulation study design

Factor	Levels
Independent studies (m)	10, 15, 20, 30, 40 – 200 (in units of 20)
Effect sizes per study ($k_1,\ldots,k_m$)	constant at 10 or varied, ranging from 1 to 10
Sample size per study	constant at 30 or varied, ranging from 32 to 130
Correlation between the outcome measures	.0, .5, .8
Between-study variability in true effect sizes as a proportion of total variation (I^2)	.00, .33, .50

Figure 1. Type-I error rates of conventional Wald test by sample size (*m*) and test dimension (*q*)
for nominal α = .05

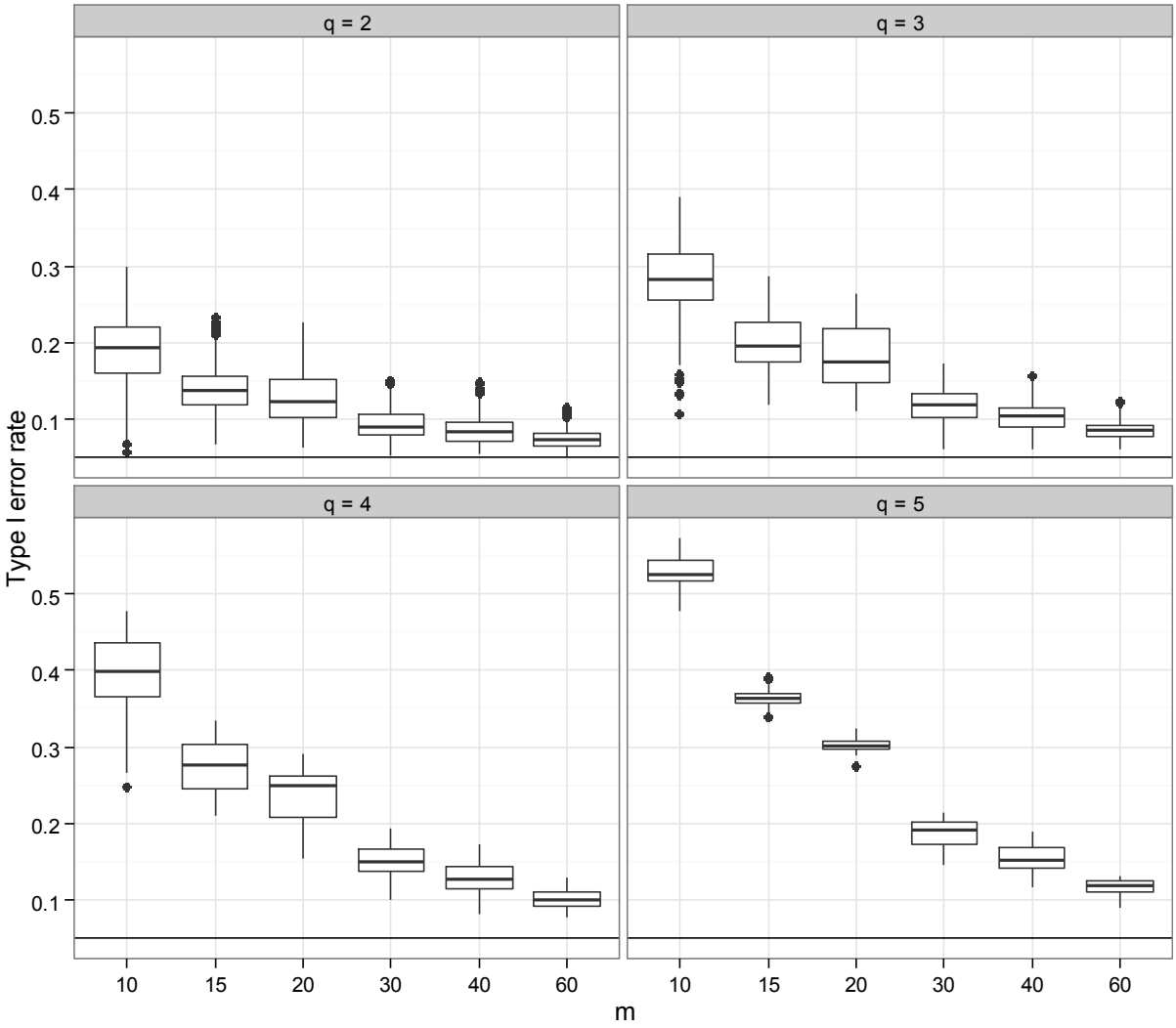

Figure 2. Type-I error rates of Wald, T^2, Satterthwaite, and SDT tests by sample size (m) for nominal α = .05