3-1-2015

The Use of Meta-Analytic Statistical Significance Testing

Terri D. Pigott
Loyola University Chicago, tpigott@luc.edu

Joshua R. Polanin
Loyola University Chicago, jpolanin@luc.edu

Author Manuscript
This is a pre-publication author manuscript of the final, published article.

Recommended Citation

Pigott, Terri D. and Polanin, Joshua R.. The Use of Meta-Analytic Statistical Significance Testing. Research Synthesis Methods, 6, 1: 63-73, 2015. Retrieved from Loyola eCommons, School of Education: Faculty Publications and Other Works, http://dx.doi.org/10.1002/jrsm.1124

The Use of Meta-Analytic Statistical Significance Testing

Joshua R. Polanin, Ph.D.

Vanderbilt University

Terri D. Pigott, Ph.D.

Loyola University Chicago

Author Notes

Joshua R. Polanin, Ph.D., Peabody Research Institute, Vanderbilt University & Terri D. Pigott, Ph.D., School of Education, Loyola University Chicago. Research for the current study was partially supported by Loyola University Chicago, Advanced Dissertation Fellowship. Opinions expressed herein do not necessarily reflect those of Loyola University Chicago. Correspondence regarding this article should be addressed to Joshua R. Polanin, Peabody Research Institute, 230 Appleton Place, Nashville, TN 37203.

Email: joshua.r.polanin@vanderbilt.edu; Phone: 217.369.2046

Abstract

Meta-analysis multiplicity, the concept of conducting multiple tests of statistical significance within one study, is an underdeveloped literature (Tendal, Nüesch, Higgins, Jüni, & Gøtzsche, 2011). We address this issue by considering how Type I~~I~~ errors can impact meta-analytic results, suggest how statistical power may be affected through the use of multiplicity corrections, and propose how meta-analysts should analyze multiple tests of statistical significance. The context for this study is a meta-review of meta-analyses published in two leading review journals in education and psychology. Our review of 130 meta-analyses revealed a strong reliance on statistical significance testing without considering of Type ~~I~~I errors or the use of multiplicity corrections. In order to provide valid conclusions, meta-analysts must consider these issues prior to conducting the study.

Keywords: Meta-analysis; ~~S~~statistical significance testing, multiplicity corrections, Type I errors, moderator analyses, power analysis, review of reviews.

The Use of Meta-Analytic Statistical Significance Testing

The concept of multiplicity, conducting multiple tests of statistical significance within one study, has received much attention in primary research over the last seven decades (Hochberg & Tamhane, 1987; Keselman, Cribbie, & Holland, 1999; Neyman & Pearson, 1928; Tukey, 1949). The major focus of this attention is due to the fact that conclusions derived from multiple tests have an increased probability of falsely rejecting a true null hypothesis (i.e., Type ~~I~~I error). Researchers and statisticians, as such, have developed methods to control the probability of making such false conclusions. Far less research, however, has been conducted on multiplicity in meta-analysis (Tendal, Nüesch, Higgins, Jüni, & Gøtzsche, 2011). In fact, multiple calls have been made for meta-analysis methodologists to address this important issue (Bender et al., 2008; Sutton & Higgins, 2008), yet few scientific advances have been put forth since Hedges and Olkin's (1985) discussion of this topic.

Given the prolific dissemination and increased usage of meta-analysis (Williams, 2012), it is paramount to investigate and ensure the validity of reviews' results. Matt and Cook (2009) developed a validity framework for meta-analytic results using the paradigm delineated in Cook and Campbell (1979) for primary studies. One threat discussed by Matt and Cook is the threat of "capitalizing on chance in meta-analysis" (pg. 544). They summarized the inherent problem: "Although research syntheses may combine findings from hundreds of studies and thousands of respondents, they are not immune to inflated ~~t~~Type ~~I~~I error when many statistical tests are conducted without adequate control for error rate" (pg. 545). A second issue raised by Matt and Cook (2009) is the lack of statistical power for detecting an association. The statistical power for any meta-analytic test depends on the number of studies, the sample sizes within studies, the size of the effect of interest, and for random effects models, the between-studies variance component (Hedges and Pigott, 2001; 2004). The power or Type II error rate within a meta-analysis also depends on the type of test conducted, with the test of the mean effect size generally having more power than tests of moderators in effect size models.

While researchers have been concerned with both Type I and Type II errors in meta-analysis, few researchers have examined both issues simultaneously. For example, Bender et al. (2008) identified some of the reasons for multiplicity, but did not investigate the prevalence of the problem nor its impact on power. Cafri, Kromrey, and Brannick (2010) studied the extent of Type ~~I~~I and Type II errors through a systematic review of meta-analyses in psychology. The authors randomly selected a relatively small subset of studies (*n* = 13) for analysis. The results indicated that only 14% of statistical tests had power less than .80 (pg. 252), and up to 78% of the studies' conclusions could be the product of Type ~~I~~I errors. While the authors investigated both Type I and Type II errors in meta-analyses, they did not discuss these issues simultaneously. Though reviewers could make corrections for Type I error rates across a meta-analysis, few researchers have acknowledged the simultaneous impact on power when adjusting for Type I error rates. The prevalence of statistical tests in meta-analysis and their consequences for both Type I and Type II error is the concern of this paper.

We have three aims in this paper. First, the paper reviews the kinds of tests used in meta-analysis, and discusses the problems with Type I and Type II error rates when multiple tests are conducted. Second, the paper updates the review of meta-analyses by Cafri et al. (2010) to examine the prevalence of multiplicity in both psychology and education reviews. This larger sample of meta-analyses in these two areas allows the examination of the potential for Type I errors, and the subsequent impact on power when adjusting for Type I error rate. Third, given the prevalence of statistical testing in meta-analyses, we suggest a framework for categorizing families of tests in order to plan for optimal levels of Type I and Type II errors. We believe this paper will elucidate further the issue of significance testing in meta-analyses as well as serve to inform the community of current and past practices so as to produce more meaningful methodological policy.

Background

Statistical Significance Testing in Meta-Analysis

In some methods of meta-analysis (Borenstein, Hedges, Higgins, & Rothstein, 2009), ~~A number of~~ statistical significance tests coincide with each quantitative synthesis phase ~~(Borenstein, Hedges, Higgins, & Rothstein, 2009)~~. Each of these tests follows the traditional null hypothesis testing framework from a frequentist perspective (Cohen, 1994; Howell, 2006). Several test statistics pervade the meta-analytic literature as illustrated in Table 1. Notably, Hunter and Schmidt (2004) suggested confidence intervals in place of statistical significance tests.

Generally speaking, the reviewer ~~usually~~ begins by conducting a z-test to determine whether the average effect size differs significantly from zero. The meta-analyst then tests the overall distribution of effect sizes for homogeneity (the second row of Table 1). Given ~~significant~~ appropriate variation in the distribution of effect sizes, the meta-analyst has a number of choices to explain the variability: subgroup analyses, categorical moderator analyses using the *Q*-Between and *Q*-Within statistics, pairwise comparisons, and meta-regression models (the remaining rows of Table 1). Each type of analysis uses a test statistic to determine whether the effect occurred by chance. A test statistic greater than or equal to a pre-determined critical value provides information to reject the null hypothesis.

Formatted: Indent: First line: 36 pt

The framework of testing for statistical significance at each stage will result in a number of tests of statistical significance being conducted in each study. For example, a typical meta-analysis in education or psychology may examine a number of different outcomes, and thus conduct a series of tests for each of these outcomes such as a test of the overall effect size, a test of homogeneity, and a number of subgroup tests if there are sufficient numbers of studies that examine this outcome. The consequence of conducting multiple tests of statistical significance is a higher probability of Type I~~I~~ errors. Tukey (1949) deemed the error rate within an experiment (i.e., a single study) the family-wise error rate. This can be shown as:

$$\text{Family-wise error rate (FWER)} = 1 - (1 - \alpha')^{c} \qquad \text{(EQ 1)}$$

where α' is the analysts' chosen alpha level for a given comparison (e.g., α' = .05), and c is the number of statistical tests. EQ 1 represents a strong control of Type ~~I~~I error rate, or the assumption that Type ~~I~~I errors can be controlled under any combination of true and false null hypotheses (Dudoit, Shaffer, & Boldrick, 2003). In other words, in EQ 1, we assume that 100% of the null hypotheses are true, and we want our error rate across these tests to be constant. In practice, however, we may hypothesize that less than 100% of the null hypotheses are true, presumably guided by prior research. Such a scenario, where less than 100% of null hypotheses are presumed to be true, represents weak control, and is represented by an adaptation to EQ 1:

$$\text{FWER*} = 1 - (1 - \alpha')^{c*p} \qquad \text{(EQ 2)}$$

where *p* represents the proportion of true null hypotheses proposed (Dudoit et al., 2003). When *p* = 1.0 or 100% of true null hypotheses, EQ 2 simplifies to EQ 1. Figure 1 illustrates the FWER under different assumptions about *p,* the percentage of null hypotheses that are true. When 100% of the null hypotheses are true (*p*=1.0), our FWER increases steeply so that our chances of making a Type I error is .5, or 50% when conducting between 10 and 15 tests. If we assume that only 1% of our null hypotheses are true (*p*=0.1), then we are much less likely to commit a Type I error even after conducting 90 tests. Thus, our FWER depends not only on our choice of alpha (α) but also on the number of null hypotheses we consider are true. This will be an important consideration in thinking about the impact of multiplicity on power in meta-analysis.

Type I~~1~~ Error Rate Control

Many corrections, adjustments, and modifications to statistical tests have been proposed in primary studies in order to control Type I error rate for multiple hypothesis testing, and any could be proposed for use in meta-analysis (see Keselman, Miller, & Holland, 2011 for a review). The major strategy for these corrections is to decrease the critical *p*-value to a level less than the original nominal level. The corrections limit Type ~~1~~I errors by making it more difficult to reject a null hypothesis.

Hedges and Olkin (1985) hypothesized the use of two such techniques for Type I error control in meta-analysis. The first advocated by Hedges and Olkin, the Bonferroni correction, is formally written:

$$\alpha^* = \frac{\alpha}{c} \qquad \text{(EQ 3)}$$

where α* is the researcher's new alpha level, α is the researcher's alpha level for the family of comparisons, and c is the number of comparisons. The researcher simply divides the alpha (usually .05) by the number of comparisons, $\alpha^* = \frac{.05}{5} = .01$. The procedure is identical to the primary study procedure and therefore encounters similar problems (i.e., too conservative with many comparisons; Hochberg, 1988). The second method advocated by Hedges and Olkin (1985) was Scheffe's S (1959) procedure. The meta-analyst compares the squared z-statistic against a chi-square distribution with degrees of freedom equal to the number of comparisons minus 1. For example, if the researcher is interested in 6 comparisons, the researcher would use the chi-square with 5 *df* for the critical value; thus, the squared z-statistic must exceed 11.05 to reject the null hypothesis. A third suggestion comes from Borenstein et al. (2009) who advocated for decreasing the alpha level to .01. Borenstein et al. argued that increased usage of statistical tests, as well as an increase in perceived power, provided ample leverage to suggest the decrease in alpha. The logic of decreasing the alpha value to .01 follows Fisher's (1935) originally suggested values. Yet, Borenstein et al. failed to provide an explanation as to why the p-value of .01 was chosen beyond convention.

The Relationship between Type ~~1~~I Error Rate Controls and Power

When a researcher applies corrections to control for Type I error rates, there will be a corresponding impact to the power of the statistical tests. To illustrate the relationship between power and the Bonferroni correction in the context of meta-analysis, let us first delineate the equation for the power of the overall z-test for the meta-analytic main effect:

$$p = 1 - \Phi(c_{\alpha-}\lambda) \qquad \text{(EQ 4)}$$

where c_{α} represents the 100(1-α) critical value from the standard normal distribution, $\Phi(x)$ is the standard normal cumulative distribution function, and λ represents the noncentrality parameter and provides evidence against the null the hypothesis. Please not the model varies slightly given different assumptions of the distribution of effect sizes (i.e., fixed- vs. random-effects models). Valentine, Pigott, and Rothstein (2010) provide the ~~exact values for~~ formulas ~~comput~~ for computing ~~ing~~ the noncentrality parameter under these scenarios~~a number of scenarios~~.

As can be readily seen from EQ 4, power is closely related to the critical value, c_{α}, chosen. ~~Traditionally, p~~Power analyses traditionally keep constant this parameter and vary the noncentrality parameter to measure the impact of effect size magnitude or variance on power. However, ~~O~~one may also vary the critical value as a function of the number of statistical tests, as is done using the Bonferroni correction. Figure~~s~~ 2 ~~and 3~~illustrates ~~show~~ how the power of the overall z-test varies as function of the number of tests conducted~~, and the number of studies~~

~~included in the meta-analysis~~. Each graph illustrates the power to detect an effect size of .2 and a within-study sample size of 50 for the treatment and control groups. We varied the size of the meta-analyses to be 10, 20, and 40 studies. The upper quadrants represent a fixed effect analysis. The lower quadrants represent a random-effects analysis, where heterogeneity is moderate (Pigott, 2012; I^2 = 75%). The right-hand quadrants use alpha values of .05, while the left-hand quadrants use values of .01. ~~Figure 2 uses a critical values of .05 and Figure 3 uses the .01 critical value.~~ We ~~then~~ adjusted these critical values using the Bonferroni procedure, dividing the critical value by the number of statistical tests (e.g., 1-20).

As seen in Figure 2, the power to detect an effect size of .2 is quite large given a moderate number of studies (k = 20 ~~and above~~ and 40). For example, using the Bonferroni corrected critical value of .0025 for 20 significance tests (.05/20), under the fixed-effect framework, would still result in power of .95. Using the smaller original critical value of .01 ~~as in Figure 3 would~~ results in a much lower critical value given 20 tests of statistical significance (α* = .01/20 = .0005); however, the resulting power to detect a main effect only decreases to .88, above what is generally considered the nominal level of .80. With a smaller number of studies (k=10), however, the power is much reduced. Indeed the power of the test using the .0025 critical value in Figure 2 drops to .64 and decreases substantially to .45 in Figure 3 using the critical value of .0005. The power to detect significant effects under a random-effects model, intuitively, is always lower.

Comment [JRP1]: if we change the graphs, this would need to be changed to reflect type 2 errors instead of power

The above illustration shows the interplay between power and Type I error corrections in the context of meta-analysis for the overall z-test of the mean effect size. As seen in the next section of the paper, there are a number of other tests that are conducted routinely in meta-analysis that are generally less powerful than the test of the mean effect size, and thus the power of those tests will be more impacted by corrections for Type I error rate. The next section presents results from a review of reviews designed to examine the prevalence of statistical tests conducted, the typical rates of Type I error, and the potential impact on power of adjusting for those error rates.

The Prevalence of Statistical Tests in Meta-Analyses in Psychology and Education

Cafri, Kromey & Brannick (2010) conducted what they called a meta-meta-analysis or a meta-review or umbrella review (Cooper & Koenka, 2012) in meta-analyses in psychology to illustrate the typical statistical power, Type I error rates, and other practices used. In this section, we examine both psychology and education meta-analyses to understand the number and types of statistical tests conducted. Knowing the prevalence of particular types of statistical tests can guide the development of a framework for planning meta-analyses for optimal levels of both Type I error and power.

Sampling Frame

Published meta-analyses served as the primary observational unit. Included reviews were published from 1986-2011 in either Psychological Bulletin (PB) or the Review of Educational Research (RER). The date parameters were chosen because they represented the era when meta-analysis grew in popularity and the field established a primary set of recommendations and procedures. These journals were chosen purposively for a number of reasons. First, these journals published reviews that were relevant to psychology and education, our main substantive interests. Second, the journals offered high quality studies that underwent an intensive peer-review process.

Inclusion/Exclusion Criteria

The reviews must have met several criteria to be included in the sample. First, the reviews must have been a quantitative synthesis (i.e., meta-analysis) and not merely a systematic review. Studies that reviewed other meta-analyses (e.g., meta-reviews, umbrella reviews, etc.) were also eliminated. Second, the review must have utilized a technique that combines effect sizes to calculate point estimates and confidence intervals (or have the potential to calculate confidence intervals). This eliminated vote counting procedures or descriptive studies. Third, the studies must have presented the results of the synthesis in a quantitative manner.

The selection process followed two stages. The first author screened all titles and abstracts for potential reviews. Citations that appeared to meet the inclusion criteria were downloaded and fully screened. The studies that met the inclusion criteria listed above were included. Due to the large number of studies included, we found it necessary to sample from this included population. To ensure that the sample of studies represented the populations, we utilized a stratified-random sampling technique. The date of publication and publication outlet served as the two main strata. This sampling design constituted a 2x2 sampling frame. Proportional allocation (Lohr, 1999) was used to maintain the unbalanced samples across the strata. We sampled meta-analyses, rather than to select all meta-analyses, to reduce the burden of coding.

Coding

A codebook was constructed to extract study-specific information in a standardized manner. Four major sections guided the coding process. The first section coded basic information about the review. The second section detailed the study's meta-analytic characteristics. This section extracted specific information on the number of statistical tests and syntheses. In order to be counted as testing a hypothesis, the review authors must have presented *p*-values or its equivalent (i.e., an asterisk to represent a *p*-value). Hypotheses where the review author reported only a confidence interval were not counted as conducting a significance test (e.g., Hunter & Schmidt, 2004, meta-analyses). In addition, each independent synthesis was coded separately. For instance, if the review authors split the effect sizes into two groups (e.g., by the outcome), then each synthesis was coded separately. For each synthesis, then, the first author coded the type and number of significance tests.

Analysis

We conducted several analytic procedures to answer the research questions sufficiently. To understand the reliance and prevalence of statistical tests, we first present the descriptive results conditional on the publication outlet. These results included the number of statistical test utilized by the type of test. In addition, an examination of the use of multiplicity corrections was provided. Based on the number of statistical tests conducted, we also evaluated the probability of committing a Type ~~I~~I error under differing assumption of the proportion of true null hypotheses (100% and 10%). The calculations for this statistic were provided in Equations 1 & 2. The descriptive analyses were conducted using SPSS version 21 (IBM Corp., 2012); the figures were drawn using EXCEL.

Results

We scanned and screened every article from every issue of PB and RER from 1986 – 2011. PB citations represented 74.2% of the total and therefore were selected with the same proportionality. To ensure proportionality by publication date, we selected the greatest proportion of studies from 2000-2011 within the PB strata (42.6%). The second highest proportion derived from PB within the 1986-1999 year range (31.6%). Only 10.4% and 15.4% of the total citations selected derived from RER during the years of 1986-1999 and 2000-2011, respectively. A total of 130 articles were included in the review (list available upon request from the authors).

Descriptive Overview

Of the 130 articles reviewed, 96 (73.8%) were published in PB and 34 (36.2%) in RER (Table 2). Half of the reviews indicated that they received at least partial financial contributions (PB = 53.1%, RER = 47.1%). The two journals differed with respect to the types of reviews published. PB included a substantial portion of observational reviews (i.e., correlational; 74.0%) while RER published more experimental studies (i.e., intervention reviews; 70.6%). More reviews in RER were considered updates (44.1%) relative to PB (21.9%).

Several other relevant findings related to the methodological quality of reviews are important to note. While most studies that reported a model specification utilized a random-effects model (PB = 29.2%, RER = 32.4%), many studies failed to report the underlying model (PB = 10.4%, RER = 47.1%). Very few reviews conducted or discussed any type of power analysis (PB = 3.1%, RER = 0%). Only a small proportion reported the p-values for all statistical analyses (PB = 22.9%, RER = 14.7%), and an even smaller portion of studies discussed the possibility of multiplicity, Type ~~1~~I error rates, or statistical significance tests (PB = 9.4%, RER = 0%).

Finally, only a small portion of reviews utilized a multiplicity correction or adjustment (PB = 18.8%, RER = 20.6%). The most popular way to combat Type ~~1~~I errors, when a correction was utilized, was to simply adjust the alpha level to something less than the nominal .05 (PB = 55.6%, RER = 57.1%). In PB, one-third of studies that used a correction used the Bonferroni, a larger portion of studies relative to its use in RER (14.3%). Scheffe's correction and the False Discovery Rate were also utilized once by each publication. Overall, the review authors were unlikely to utilize a correction and when they did, it was often the nominal step of adjusting the alpha level.

Statistical Test Usage

We examine statistical test usage following the organization of Table 1. Statistical significance tests were relied upon heavily across both sources (Table 3). On average, reviews published in PB conducted 67.94 tests of significance (SD = 105.51), while reviews in RER conducted 51.85 (SD = 49.17). The average meta-analysis in PB included 7.17 tests of the statistical significance of the effect size (SD = 14.00); similar results were revealed for RER (M = 7.09, SD = 30.32). When reviewers use a test of the mean effect size, this typically indicates a "split," or an independent meta-analysis within a larger systematic review. Thus, in both PB and RER, reviewers conducted an average of seven meta-analyses. This might occur, for example, by analyzing outcomes types separately such as splitting academic measures from attitudinal ones. Relatedly, reviews in PB conducted 11.43 tests of effect size homogeneity (SD = 16.71), while RER conducted only 2.69 (SD = 3.34).

Of particular interest is the number of moderator tests conducted as seen in the later rows of Table 3. Reviewers in PB and RER each conducted a relatively large number of *Q*-Between significance tests (PB = 9.11, RER = 11.00), which are used to test differences between mean effect sizes in an ANOVA model. With regard to the number of *Q*-Within tests conducted, PB meta-analyses reported a higher number (M = 16.79, SD = 33.68) relative to RER studies (M = 4.24, SD = 10.94). Although differences may well exist, an alternative hypothesis is that RER authors fail to report the results of these significance tests. The results revealed differences less severe with regard to subgroup comparisons (PB = 12.05, RER = 17.32) and meta-regression analyses (PB = 11.39, RER = 9.53).

Table 4 provides the family-wise error rates (FWER) under two assumptions about the proportion of true null hypotheses for the number of tests in Table 3. The table presents the average FWER under two assumptions of the proportion of true null hypotheses: p=1.0 and p=0.1 (EQs 2 & 3), and confidence intervals for these average FWER values using bootstrap techniques. The probability of committing a ~~le~~ Type I~~1~~ error~~ rate~~, given the assumption of 100% true null hypotheses, was high for both PB and RER meta-analyses, not surprisingly, because of the large number of statistical tests conducted (Table 4). The FWER for PB reviews ranged from a low of .20 (95% C.I. = .15, .25) for the overall effect sizes to a high of .30 (95% C.I. = .25, .37) for the overall homogeneity calculations. For RER reviews, the FWER rate ranged from a low of .11 (95% C.I. = .06, .19) for the overall effect sizes to a high of .31 for Q-Between tests (95% C.I. = .21, .42). The FWER decreases significantly when the assumption of true null hypotheses decreases to only 10% (EQ 3 with p=0.1). For PB and RER, the FWER decreases to only 3% for the overall tests of the average effect.

A Proposed Framework for Multiple Hypothesis Testing and Power in Meta-analysis

Clearly, review authors rely on statistical significance testing at each stage of the quantitative synthesis. What is not clear is how to handle the subsequent Type ~~1~~I error rates while maintaining sufficient power for meta-analysis. Power for any correction for multiple hypothesis testing will depend not only on the usual parameters (within-study sample size, alpha level, effect size of interest) but also on the assumptions we make about the number of true null hypotheses. Although it is tempting to posit that only 10% of the null hypotheses are in fact true, this may not be a fair assumption for some types of tests. Of course, should we allow for 100% of true null hypotheses and decide instead to use a correction procedure, we must be cautious not to decrease power beyond a reasonable level. Given the myriad decisions, we believe that these decisions should be dictated *a priori*, similar to the ways in which an author provides inclusion/exclusion criteria.

To help guide reviewers, we propose a preliminary framework for planning for the power of multiple hypothesis tests in meta-analysis. In order to think about power for multiple hypothesis tests, we need to group the tests so that we can compute both the FWER and the impact on power of any potential correction for multiple hypothesis testing. Our proposed grouping is informed by our review of meta-analyses in PB and RER. One preliminary suggestion might be to correct for multiple hypothesis testing by correcting for all possible statistical tests in a given meta-analysis. As seen in Table 3, we could potentially be applying corrections for a median of 50 tests. Another strategy, and one that we will describe here, is to group the tests in terms of the type of test, as outlined in Table 1. Using the groupings in Table 1 allows us to make different kinds of assumptions about the proportions of null hypotheses that are true for different types of tests. We propose a framework that groups sets of statistical tests based on the rows of Table 1. Each of these sets of tests would be considered a family of tests, so that, for example, we correct for Type I error across all tests of the statistical significance of the mean effect size, apply another Type I error correction across all tests of homogeneity, etc.

Within each family of tests, we would need to think about our optimal FWER. This consideration should include both the number of true null hypotheses we expect and the consequences of committing a Type I error. Let us take the case of a family of tests for the statistical significance of the mean effect size. In many cases, especially in intervention meta-analyses, we might expect that many effect sizes are indeed statistically different from zero. Lipsey and Wilson (1993) surveyed 302 meta-analyses and found that only 6 of the 302 yielded average effect sizes less than 0 (i.e., harmful treatment effects). Based on this information, it may be reasonable to propose that a proportion of null hypotheses are indeed false. Thus, we might want to decide, given the likely outcomes in the studies, which effect sizes are different from zero and have null hypotheses that are probably false. In this case, we may plan for a FWER that allows for less than 100% of null hypotheses as true.

In the case of families of tests of *Q*-between, for example, we might make a different choice. We might have a number of potential moderators that we wish to test within each set of outcomes. As seen in Table 3, reviews in PB had an average of about 10 *Q*-between tests. We might not have any idea if any of the moderators are related to effect size variation, and thus may want to be more conservative in our statistical testing. Here we might want to have strong control on Type I error by assuming all null hypotheses are true. In this case, we would choose strong control in our FWER.

Table 5 shows a range of FWER for different assumptions about the number of true null hypotheses using a study in our sample by Archer (2000). Archer (2000) synthesized 82 studies that measured the gender difference in aggression between heterosexual partners. When we assume strong control, i.e., when we assume 100% true null hypotheses, all of the FWER for each of the four groups of tests is larger than the nominal 5% error rate. As we relax our assumptions about the number of true null hypotheses, our FWER is reduced, but is not close to the 5% error rate until we assume only 1% of true null hypotheses. If we are interested in maintaining a Type I error rate of 5%, we would need to apply a multiplicity correction particularly in the cases where we assume 100% or 50% true null hypotheses.

The type of multiplicity correction applied will impact the overall power of our tests. To illustrate the impact of multiplicity corrections on power, we again use the data from Archer (2000). In the Archer study, all five of the outcomes were statistically significant using the $p <$

.05 critical value. Table 5 also shows the impact on average power for the four types of statistic tests when we use two different multiplicity correction procedures: a) reduce the level of α to 0.01 as suggested by Borenstein et al. (2010), and b) Bonferroni corrections to ensure that the FWER is either 0.05 or 0.01. Looking first at the test for the average effect size, we see that each multiplicity correction procedure would result in one less effect size that is considered statistically significant. In addition, the average power across the five tests of average effect size remains high. The conclusions reached for the five tests of homogeneity are not changed using any of these procedures, and all have sufficient average power. The situation is different, however, for the tests used for moderator analyses. While 22 out of 37 Q-between tests were statistically significant at the 0.05 level, this number decreases to 19 if we use α=0.01, and to 14 and 13 when we use the Bonferroni correction to keep the FWER to 0.05 or 0.01, respectively. The average power of the Q-between tests under the correction procedures is not close to nominal levels, with the highest average power occurring with decreasing alpha to 0.01. Similar results occur with Q-within tests, though the power of these tests is closer to acceptable levels.

In this example, applying multiplicity corrections has different impacts on both the number of tests considered statistically significant as well as the power to conduct these tests. While tests of the average effect and overall homogeneity may be less prone to Type I and Type II errors, tests such as the Q-between and Q-within associated with moderator analyses are more susceptible to Type I and Type II errors due to the number of tests conducted. The overall lower power and Type I error problems with moderator analyses typically occur due to the large number of these tests that are conducted in a typical meta-analysis, and to the smaller number of studies usually available to conduct these tests. There may also be cases where we are not as concerned about our FWER, such as when we assume that only a small proportion of our null hypotheses are true. When we assume that a large proportion of the null hypotheses for the overall average effect are likely to be false, we may not need to apply multiplicity corrections.

Discussion

Our review of reviews along with Cafri et al.'s (2010) review clearly demonstrate the large number of statistical tests that are conducted in meta-analyses in high quality journals in psychology and education. Despite the number of tests conducted, less than 20% of all studies utilized a correction technique, and even fewer considered the impact of Type ~~1~~I errors on the results (6.9%). The average review in PB and RER conducted a high number of statistical significance tests (M: PB = 67.94, RER = 51.85). The resulting family-wise error rate, across all tests, was large enough to warrant caution of the findings. The remarkable irony of this conclusion is that meta-analysis is often lauded as an ends to remove all statistical significance testing. Indeed, Schmidt (1996) believed that meta-analysis would render the use of statistical significance testing null. The use of meta-analysis, Schmidt said, "…reveals more clearly than ever before the extent to which reliance on significance testing has retarded the growth of cumulative knowledge in psychology" (pg. 116). Paradoxically, the increased use of meta-analysis may have inadvertently *increased* the use of statistical significance testing.

Reviewers should not apply multiplicity corrections, however, without attention to both assumptions about the optimal FWER for a set of tests as well as the potential impact on the power of the tests. As illustrated in our example, reviewers may assume that most of the null hypotheses for the overall effect size and overall homogeneity may in fact be false, and thus do not warrant multiplicity corrections. Tests for moderator analyses may be more susceptible both to Type I errors and lower power; researchers need to be cautious in their use and interpretation of these tests. We have proposed that reviewers treat the types of tests in a meta-analysis as a family, and make decisions about optimal FWER as well as multiplicity corrections in relation to these groups of tests.

One suggestion for the practice of meta-analysis is to reduce the use of statistical tests by planning a priori the analyses that will be conducted. Though reviewers can never be sure about the numbers of studies that will be eligible for the review or the characteristics of those studies, understanding the nature of the research conducted in a given field can inform decisions about optimal levels of FWER and about the kinds of multiplicity corrections that would be needed. The recently developed quality appraisal tool for meta-analyses, indeed, asks reviewers to consider the ramifications of multiplicity specifically (Higgins et al., 2013).

We also advocate for a standardized means to test moderators. From the standpoint of decreasing multiplicity in meta-analysis, meta-regression is the logical choice. The model decreases statistical testing because it simultaneously conducts multiple tests. Moreover, the model inherently controls for all other predictors in the model, and therefore provides a more precise result. We should caution readers, however, that meta-regression has several egregious issues; and these should be weighed against the probability of committing Type ~~1~~I or ~~2~~II errors. As a result, we might also suggest two-way ANOVA models to test for the presence of interactions.

Finally, we should mention that avoiding statistical significance testing altogether is a worthy alternative. The journal *Psychological Science,* in fact, initiated a new policy to reduce and eventually eliminate the use of null hypothesis significance testing (Cumming, 2014). Hunter and Schmidt (2004) advocated for the use of confidence intervals, in lieu of *p*-values, as well as a reliance on the clinical significance of effect sizes. Moderator analyses are handled through a purely descriptive approach, where differences between subgroup are not tested for statistical significance. Alternatively, Sutton and Abrams (2001) espoused the Bayesian methodological approach. Using this framework completely eliminates the need for significance testing.

Limitations

A number of limitations about this project should be considered. For instance, it is feasible that PB and RER attract meta-analyses where the literature is developed and plentiful, therefore leading review authors to assume it is reasonable to conduct a high rate of significance testing. This is partially confirmed by the fact that reviews that included more than the average number of studies tended to conduct more significance tests. Similarly, review authors might choose to publish their findings in PB or RER because they included a large number of studies. Given these concerns, one should limit generalizability hypotheses.

One critical limitation to consider is whether Type ~~1~~I errors are a problem in meta-analysis at all. At least with regard to the overall average effect size, the goal is to collect and synthesize the population of effect sizes across every available resource. As such, it is not clear whether the distribution of effect sizes is subject to traditional frequentist logic inherent in primary study inference testing. What is not in question, however, is the Type ~~1~~I errors associated with conducting moderator analyses. The conclusions drawn from multiple one-way ANOVAs should be considered highly suspect under the condition that many such tests are conducted.

Conclusions

The conceit of this project was to investigate whether statistical significance testing is a problem worth addressing in meta-analysis. It seems clear, given the high rate of null hypothesis significance testing coupled with the egregious lack of correction, Type ~~1~~I errors can and will impact the validity of meta-analytic results. Instead of blindly correcting for Type ~~1~~I errors, however, we must consider the impact of these corrections on power and balance both these concerns. As with any statistical analysis, an *a priori* protocol of the tests to be conducted as well as a power analysis is advised. Meta-analysis methodologists and practitioners simply cannot afford to ignore these issues while simultaneously promoting the promise of the paradigm's results. We must take action to prevent any doubt about the results of meta-analysis.

References

Archer, J. (2000). Sex differences in aggression between heterosexual partners: a meta-analytic review. *Psychological bulletin*, *126*(5), 651.

Bender, R., Bunce, C., Clarke, M., Gates, S., Lange, S., Pace, N. L., & Thorlund, K. (2008). Attention should be given to multiplicity issues in systematic reviews. *Journal of Clinical Epidemiology, 61*, 857-865. doi: 10.1016/j.jclinepi.2008.03.004

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. New York, NY: John Wiley & Sons.

Cafri, G., Kromrey, J. D., & Brannick, M. T. (2010). A meta-meta-analysis: Empirical review of statistical power, type I error rates, effect sizes, and model selection of meta-analyses published in psychology. *Multivariate Behavioral Research, 45*, 239-270. doi: 10.1080/00273171003680187

Cohen, J. (1994). The earth is round (p<.05). *American Psychologist, 49*, 997-1001. doi: 10.1037//0003-066X.49.12.997

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design & analysis issues for field settings*. New York, NY: Rand McNally.

Cooper, H., & Koenka, A. C. (2012). The overview of reviews: Unique challenges and opportunities when research syntheses are the principal elements of new integrative scholarship. *American Psychologist*, *67*, 446-462. doi: 10.1037/a0027119

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7-29.

Dudoit, S., Shaffer, J. P., & Boldrick, J. C. (2003). Multiple hypothesis testing in microarray experiments. *Statistical Science*, 71-103.

Fisher, R. A. (1935). The design of experiments. Oxford, England: Oliver & Boyd.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. London, England: Academic Press Inc.

Hedges, L.V., & Pigott, T.D. (2001). The power of statistical tests in meta-analysis. *Psychological Methods, 6,* 203-217.

Hedges, L. V., & Pigott, T. D. (2004). The power of statistical tests for moderators in meta-analysis. *Psychological Methods*, *9*(4), 426-445.

Higgins, J. P.T., Lane, P. W., Anagnostelis, B., Anzures-Cabrera, J., Baker, N. F., Cappelleri, J. C., Haughie, S., Hollis, S., Lewis, S. C., Moneuse, P. and Whitehead, A. (2013). A tool to assess the quality of a meta-analysis. *Res. Synth. Method*. doi: 10.1002/jrsm.1092

Hochberg, Y., & Tamhane, A. C. (1987). *Multiple comparison procedures*. New York, NY: John Wiley & Sons.

Howell, D. C. (2006). *Statistical methods for psychology*. Independence, KY: Cengage Learning Tools.

IBM Corp. (2012). *IBM SPSS Statistics for Windows*, Version 21.0 [Software]. Armonk, NY: IBM Corp.

Keselman, H. J., Cribbie, R., & Holland, B. (1999). The pairwise multiple comparison multiplicity problem: An alternative approach to familywise and comparison wise Type I error control. *Psychological Methods, 4*, 58-69. doi: 10.1037//1082-989X.4.1.58

Keselman, H. J., Miller, C. W., & Holland, B. (2011). Many test of significance: New methods for controlling type ~~1~~I errors. *Psychological Methods, 16*, 420-431. doi: 10.1037/a0025810

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: confirmation from meta-analysis.*American psychologist*, *48*, 1181.

Lohr, S. (1999). *Sampling: Design and analysis*. Independence, KY: Cengage Learning.

Matt, G. E., & Cook, T. D. (2009). Threats to the validity of generalized inferences (pp.537-660). In H. Cooper, L. V. Hedges & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis*. New York, NY: The Russell Sage Foundation.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference:Part I. *Biometrika, 20A*, 175-240. doi: 10.2307/2331945

Scheffe, H. (1959). *The analysis of variance*. New York, NY: Wiley & Sons, Ltd.

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*, 115-129. doi: 10.1037//1082-989X.1.2.115

Sutton, A. J., & Higgins, J. (2008). Recent developments in meta-analysis. *Statistics in Medicine*, *27*, 625-650.

Tukey, J. W. (1949). *Comparing individual means in the analysis of variance*. Unpublished Doctoral Dissertation. Princeton University. New Jersey.

Tendal, B., Nüesch, E., Higgins, J. P. T., Jüni, P., & Gøtzsche, P. C. (2011). Multiplicity of data in trial reports and the reliability of meta-analyses: Empirical study. *BMJ, 343*, 1-13. doi: 10.1136/bmj.d4829

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, *35*, 215-247.

Wickham, H. (2009). *ggplot2: Elegant graphics for data analysis*. New York, NY: Springer.

Williams, R. T. (2012). *Using robust standard errors to combine multiple estimates with meta-analysis*. Unpublished Doctoral Dissertation. Loyola University Chicago. Chicago, IL.

Table 1. Types of Statistical Tests in Meta-Analysis

Type	Description	Test Distribution
Overall Effect Size	H_0: ES = 0; Tests whether the effect size differs from zero	Z
Overall Homogeneity	H_0: $Q = 0$; Tests whether the homogeneity of effect sizes differs greater than from sampling error alone	Chi-square
Subgroup Effect Size	H_0: ES = 0; Tests whether the subgroup's effect size differs from zero (Can be independent or dependent subgroup)	Z
Subgroup Homogeneity	H_0: $Q = 0$; Tests whether the subgroup's homogeneity of effect sizes differs greater than from sampling error alone (Can be independent or dependent subgroup)	Chi-square
Subgroup comparisons	H_0: $ES_1 – ES_2 = 0$; Tests for the difference between subgroups	Z
Moderator analysis: Q-Between	H_0: $QB = 0$; Tests whether the variance of the subgroups is greater than sampling error alone	Chi-square
Moderator analysis: Q-Within	H_0: $QW = 0$; Tests whether the variance within each subgroup is greater than sampling error alone	Chi-square
Meta-regression: Model fit statistics	H_0: $Q = 0$; Tests whether at least one variable has a significant relationship with the outcome	F
Meta-regression: Parameter estimates	H_0: $\beta = 0$; Tests whether the slope of the regression line is greater than sampling error	Z or t

Table 2. Characteristics of Included Reviews

	Total		Psychological Bulletin		Review of Educational Research	
	N	%	N	%	N	%
N	130		96		34	
Funded	67	51.5	51	53.1	16	47.1
Type						
Experimental	43	33.1	19	19.8	24	70.6
Observational	80	61.5	71	74.0	9	26.5
Other	7	5.4	6	6.3	1	2.9
Update						
Yes	34	26.2	21	21.9	15	44.1
No	79	60.8	64	66.7	13	38.2
Partial	17	13.1	11	11.5	6	17.6
Model specification						
None	26	20.0	10	10.4	16	47.1
Fixed and Random	10	7.7	8	8.3	2	5.9
Fixed	15	11.5	10	10.4	5	14.7
Random	39	30.0	28	29.2	11	32.4
Power analysis	3	2.3	3	3.1	0	0
Report all p-values	27	21.0	22	22.9	5	14.7
Discuss multiplicity	9	6.9	9	9.4	0	0
~~Multiplicity~~ Type I error correction used	25	19.2	18	18.8	7	20.6
Technique						
Adjust alpha level	14	56.0	10	55.6	4	57.1
Bonferroni	7	28.0	6	33.3	1	14.3
Scheffe	2	8.0	1	5.6	1	14.3
False Discovery Rate	2	8.0	1	5.6	1	14.3
Number of primary studies	32	53.5	33	65	30	29.5

Notes: Percentages may not add to 100% due to rounding; Number of primary studies represented by~~s~~ the median and interquartile range~~e~~.

Formatted Table

Table 3. Number of Statistical Significance Tests by Publication Source

	Psychological Bulletin			Review of Educational Research		
	Mean (S.D.)	95% C.I.	Median	Mean (S.D.)	95% C.I.	Median
Overall ES	7.17 (14.00)	4.34, 10.01	1.50	7.09 (30.32)	0, 17.67	1.00
Overall Homogeneity	11.43 (16.71)	8.04, 14.82	4.00	2.69 (3.34)	1.51, 3.84	1.50
Q-Between	9.11 (18.38)	5.39, 12.84	0.00	11.00 (16.16)	5.36, 16.63	0.00
Q-Within	16.79 (33.68)	9.96, 23.62	0.00	4.24 (10.94)	.41, 8.05	0.00
Subgroup Comparisons	12.05 (24.71)	7.16, 16.94	0.00	17.32 (35.49)	4.94, 29.71	0.00
Meta-Regression	11.39 (29.71)	5.36, 17.41	0.00	9.53 (20.50)	2.37, 16.68	0.00
Total	67.94 (105.51)	46.42, 88.64	50.5	51.85 (49.17)	35.32, 68.37	36.5

Notes: Numbers in parentheses represent standard deviations.

Table 4. Average Family-Wise Error Rate for Each Type of Statistical Test

	Psychological Bulletin		Review of Educational Research	
True Null Hypotheses	100%	10%	100%	10%
Overall ES	.20 (.15, .25)	.03 (.02, .05)	.11 (.06, .19)	.03 (.01, .07)
Overall Homogeneity	.30 (.25, .37)	.05 (.04, .07)	.12 (.08, .16)	.01 (.01, .02)
Q-Between	.23 (.17, .29)	.04 (.03, .06)	.31 (.21, .42)	.05 (.03, .08)
Q-Within	.28 (.20, .36)	.07 (.05, .10)	.12 (.04, .21)	.02 (.01, .04)
Subgroup Comparisons	.24 (.17, .32)	.05 (.03, .08)	.26 (.14, .40)	.07 (.03, .12)
Meta-Regression	.22 (.16, .29)	.05 (.03, .07)	.20 (.09, .31)	.04 (.02, .07)

*Notes: N*s: Psychological Bulletin = 96, Review of Educational Research = 34; Numbers in parentheses represent bootstrap confidence intervals.

Table 5. Use of Proposed Guidelines for Multiplicity on Archer's (2000) Review

	Overall ES	Homogeneity	Q-Between	Q-Within
Number of Tests	5	5	37	47
Author Stated Significant (p < .05)	5	5	22	27
FWER (100%)	22.6	22.6	85.0	91.0
FWER (50%)	12.1	12.1	61.3	70.1
FWER (1%)	.25	.25	1.88	2.38
Significant at .01	4	5	19	23
Significant at .05 with Bonferroni	4	5	14	23
Significant at .01 with Bonferroni	4	5	13	16
Average Power at .01	84.1	99.9	52.7	70.2
Average Power at .05 with Bonferroni	84.1	99.9	43.6	61.6
Average Power at .01 with Bonferroni	79.7	99.7	33.2	50.5

Notes: N experiment-wise statistical significance tests = 94; Author set p < .05 as critical value for all tests; FWER = Family-wise Error Rate; FWER calculation based on the assumption true null hypotheses represented in the parentheses; ~~Author set p < .05 as critical value for all tests;~~ Power calculated by finding power for each test within the group and taking the average.

Figure 1. Type I Error Rate for Different Proportions of True Null Hypotheses

Notes: Lines represent assumptions of the percentage of true null hypotheses; FWER = Family-wise error rate.

Figure 2. Power of the Test of Average Effect Size under Different Model Assumptions Using the Bonferroni Correction

Notes: All analyses represent two-sided hypothesis; Lines represent numbers of studies (k = 10, 20, 40); Upper quadrants represent Fixed-effect analysis; Lower quadrants represent Random-effects analysis ($I^2 = 75\%$); Left quadrants represent Alpha = .01; Right quadrants represent Alpha = .05; $d = .20$; Sample size (Treatment = 50, Control = 50).