

Technical Details of WWC-Conducted Computations

(4-16-2007)

To assist in the interpretation of study findings and facilitate comparisons of findings across studies, the WWC computes the effect sizes (ES) and the improvement indices associated with study findings on outcome measures relevant to the WWC's review. In general, the WWC focuses on ESs based on student-level findings regardless of the unit of assignment or the unit of intervention. Focusing on student-level findings not only improves the comparability of ES estimates across studies, but also allows us to draw upon existing conventions among the research community to establish the criterion for "substantively important" effects for intervention rating purposes. In addition to ESs and improvement indices, the WWC also computes the levels of statistical significance of student-level findings corrected for clustering and/or multiple comparisons where necessary.

The purpose of this document is to provide the technical details about the various types of computations conducted by the WWC as part of its review process, which will allow readers to better understand the findings that we report and the conclusions that we draw regarding the effectiveness of the educational interventions reviewed by the WWC.¹ Specifically, the technical details of the following types of WWC-conducted computations are presented:

- I. Effect Size Computation for Continuous Outcomes
 - ES as Standardized Mean Difference (Hedges's g)
 - ES Computation Based on Results from Student-Level T-Tests or ANOVA
 - ES Computation Based on Results from Student-Level ANCOVA
 - ES Computation Based on Results from Cluster-Level Analyses
 - ES Computation Based on Results from HLM Analysis in Studies with Cluster-Level Assignment
- II. Effect Size Computation for Dichotomous Outcomes
- III. Computation of the Improvement Index
- IV. Clustering Correction of the Statistical Significance of Effects Estimated with Mismatched Analyses
- V. Benjamini-Hochberg Correction of the Statistical Significance of Effects Estimated with Multiple Comparisons

In addition to computational procedures, this document presents the rationale for the specific computations conducted and their underlying assumptions. These procedures are currently used to compute effect sizes and make corrections for study designs and reporting practices most commonly encountered during WWC's review process. It is not meant to serve as a comprehensive compendium of an exhaustive list of ES computation methods that have ever been developed in the field.

¹ The WWC regularly updates WWC technical standards and their application to take account of new considerations brought forth by experts and users. Such changes may result in re-appraisals of studies and/or interventions previously reviewed and rated. Current WWC standards offer guidance for those planning or carrying out studies, not only in the design considerations but the analysis and reporting stages as well. WWC standards, however, may not pertain to every situation, context, or purpose of a study and will evolve.

I. Effect Size Computation for Continuous Outcomes

ES as Standardized Mean Difference (Hedges's g)

Different types of ES indices have been developed for different types of outcome measures, given their distinct statistical properties. For continuous outcomes, the WWC has adopted the most commonly-used ES index—the standardized mean difference, which is defined as the difference between the mean outcome of the intervention group and the mean outcome of the comparison group divided by the pooled within-group standard deviation (SD) on that outcome measure. Given that the WWC generally focuses on student-level findings, the default SD used in ES computation is the student-level SD.

The basic formula for computing standardized mean difference is as follows:

$$\text{Standardized mean difference} = (X_1 - X_2) / S_{\text{pooled}}, \quad (1)$$

where X_1 and X_2 are the means of the outcome for the intervention group and the comparison group respectively, and S_{pooled} is the pooled within-group SD of the outcome at the student level. Formulaically,

$$S_{\text{pooled}} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}, \text{ and} \quad (2)$$

$$S_{\text{pooled}} = \text{sqrt}\{[(n_1-1)S_1^2 + (n_2-1)S_2^2]/(n_1+n_2-2)\}$$
$$\text{Standardized mean difference (g)} = \frac{X_1 - X_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}} \quad (3)$$

$$(g) = (X_1 - X_2) / \text{sqrt}\{[(n_1-1)S_1^2 + (n_2-1)S_2^2]/(n_1+n_2-2)\}$$

where n_1 and n_2 are the student sample sizes, and S_1 and S_2 the student-level SDs, for the intervention group and the comparison group respectively.

The ES index thus computed is referred to as Hedges's g .² This index, however, has been shown to be upwardly biased when the sample size is small. Therefore, we have applied a simple correction for this bias developed by Hedges (1981), which produces an unbiased ES estimate by multiplying the Hedges's g by a factor of $[1-3/(4N-9)]$, with N being the total sample size. Unless otherwise noted, Hedges's g corrected for small-sample bias is the default ES measure for continuous outcomes used in the WWC's review.

² The Hedges' g index differs from the Cohen's d index in that Hedges's g uses the square root of degrees of freedom ($\text{sqrt}(N-k)$ for k groups) for the denominator of the pooled within-group SD (S_{pooled}), whereas Cohen's d uses the square root of sample size ($\text{sqrt}(N)$) to compute S_{pooled} (Rosenthal, 1994; Rosnow, Rosenthal, & Rubin, 2000).

In certain situations, however, the WWC may present study findings using ES measures other than Hedges’s g . If, for instance, the SD of the intervention group differs substantially from that of the comparison group, the PIs and review teams may choose to use the SD of the comparison group instead of the pooled within-group SD as the denominator of the standardized mean difference, and compute the ES as Glass’s Δ instead of Hedges’s g . The justification for doing so is that when the intervention and comparison groups have unequal variances, as in the case where the variance of the outcome is affected by the intervention, the comparison group variance is likely to be a better estimate of the population variance than the pooled within-group variance (Cooper, 1998; Lipsey & Wilson, 2001). The WWC may also use Glass’s Δ , or other ES measures used by the study authors, to present study findings—if there is not enough information available for computing Hedges’s g . These deviations from the default will be clearly documented in the WWC’s review process.

The sections to follow focus on the WWC’s default approach to computing student-level ESs for continuous outcomes. We describe procedures for computing Hedges’s g based on results from different types of statistical analysis most commonly encountered in the WWC reviews.

ES Computation Based on Results from Student-Level T-Tests or ANOVA

For randomized controlled trials, study authors may assess an intervention’s effects based on student-level t-tests or analyses of variance (ANOVA) without adjustment for pretest or other covariates, assuming group equivalence on pre-intervention measures achieved through random assignment. If the study authors reported posttest means and SD as well as sample sizes for both the intervention group and the comparison group, the computation of ESs will be straightforward using the standard formula for Hedges’s g (see Equation (3)).

Where the study authors did not report the posttest mean, SD, or sample size for each study group, the WWC computes Hedges’s g based on t-test or ANOVA F-test results, if they were reported along with sample sizes for both the intervention group (n_1) and the comparison group (n_2). For ESs based on t-test results,

$$\begin{aligned} \text{Hedges's } g &= t \sqrt{\frac{n_1 + n_2}{n_1 n_2}}, \\ \text{Hedges's } g &= t * \text{sqrt} [(n_1 + n_2)/n_1 n_2] \end{aligned} \tag{4}$$

For ESs based on ANOVA F-test results,

$$\begin{aligned} \text{Hedges's } g &= \sqrt{\frac{F(n_1 + n_2)}{n_1 n_2}}, \\ \text{Hedges's } g &= \text{sqrt} [(F(n_1 + n_2)/n_1 n_2)] \end{aligned} \tag{5}$$

ES Computation Based on Results from Student-Level ANCOVA

Analysis of covariance (ANCOVA) is a commonly used analytic method for quasi-experimental designs. It assesses the effects of an intervention while controlling for important covariates,

particular pretest, that might confound the effects of the intervention. ANCOVA is also used to analyze data from randomized controlled trials so that greater statistical precision of parameter estimates can be achieved through covariate adjustment.

For study findings based on student-level ANCOVA, the WWC computes Hedges’s g as *covariate adjusted mean difference* divided by *unadjusted pooled within-group SD*. The use of adjusted mean difference as the numerator of ES ensures that the ES estimate is adjusted for covariate difference between the intervention and the comparison groups that might otherwise bias the result. The use of unadjusted pooled within-group SD as the denominator of ES allows comparisons of ES estimates across studies by using a common metric to standardize group mean differences, i.e., the population SD as estimated by the unadjusted pooled within-group SD.

Specifically, when sample sizes, and adjusted means and unadjusted SDs of the posttest from an ANCOVA are available for both the intervention and the comparison groups, the WWC computes Hedges’s g as follows:

$$\text{Hedges's } g = \frac{X_1' - X_2'}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}}, \quad (6)$$

$$\text{Hedges's } g = (X_1' - X_2') / \sqrt{\{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\} / (n_1 + n_2 - 2)}$$

where X_1' and X_2' are adjusted posttest means, n_1 and n_2 the student sample sizes, and S_1 and S_2 the student-level unadjusted posttest SD, for the intervention group and the comparison group respectively,

It is not uncommon, however, for study authors to report unadjusted group means on both pretest and posttest, but not adjusted group means or adjusted group mean differences on the posttest. Absent information on the correlation between the pretest and the posttest, as is typically the case, the WWC’s default approach is to compute the numerator of ES—the adjusted mean difference—as the difference between the pretest-posttest mean difference for the intervention group and the pretest-posttest mean difference for the comparison group. Specifically,

$$\text{Hedges's } g = \frac{(X_1 - X_{1-pre}) - (X_2 - X_{2-pre})}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}}, \quad (7)$$

$$\text{Hedges's } g = [(X_1 - X_{1-pre}) - (X_2 - X_{2-pre})] / \sqrt{\{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\} / (n_1 + n_2 - 2)}$$

where X_1 and X_2 are unadjusted posttest means, X_{1-pre} and X_{2-pre} unadjusted pretest means, n_1 and n_2 the student sample sizes, and S_1 and S_2 the student-level unadjusted posttest SD, for the intervention group and the comparison group respectively,

This “difference-in-differences” approach to estimating an intervention’s effects while taking into account group difference in pretest is not necessarily optimal, as it is likely to either

overestimate or underestimate the adjusted group mean difference, depending on which group performed better on the pretest.³ Moreover, this approach does not provide a means for adjusting the statistic significance of the adjusted mean difference to reflect the covariance between the pretest and the posttest. Nevertheless, it yields a reasonable estimate of the adjusted group mean difference, which is equivalent to what would have been obtained from a commonly used alternative to the covariate adjustment-based approach to testing an intervention’s effect—the analysis of gain scores.

Another limitation of the “difference-in-differences” approach is that it assumes the pretest and the posttest are the same test. Otherwise, the means on the two types of tests might not be comparable, and hence it might not be appropriate to compute the pretest-posttest difference for each group. In cases where different pretest and posttests were used, and only unadjusted means on pretest and posttest were reported, the Principal Investigators (PIs) will need to consult with the WWC Technical Review Team to determine whether it is reasonable to use the difference-in-differences approach to compute the ESs.

The difference-in-differences approach presented above also assumes that the pretest-posttest correlation is unknown. In some areas of educational research, however, empirical data on the relationships between pretest and posttest may be available. If such data are dependable, the WWC PIs and the review team in a given topic area may choose to use the empirical relationship to estimate the adjusted group mean difference that is unavailable from the study report or study authors, rather than using the default difference-in-differences approach. The advantage of doing so is that, if indeed the empirical relationship between pretest and posttest is dependable, the covariate-adjusted estimates of the intervention’s effects will be less biased than those based on the difference-in-differences (gain score) approach. If the PIs and review teams choose to compute ESs using an empirical pretest-posttest relationship, they will need to provide an explicit justification for their choice as well as evidence on the credibility of the empirical relationship.

Computationally, if the pretest and posttest has a correlation of r , then

$$\text{Hedges's } g = \frac{(X_1 - X_2) - r(X_{1-pre} - X_{2-pre})}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}}, \quad (8)$$

$$\text{Hedges's } g = [(X_1 - X_2) - r(X_{1-pre} - X_{2-pre})] / \sqrt{\{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2\} / (n_1 + n_2 - 2)}$$

where all the other terms are the same as those in Equation (7).

A final note about ANCOVA-based ES computation is that Hedges’s g cannot be computed based on the F-statistic from an ANCOVA using Equation (5). Unlike the F-statistic from an ANOVA, which is based on unadjusted within-group variance, the F-statistic from an ANCOVA is based on covariate-adjusted within-group variance. Hedges’s g , however, requires

³ If the intervention group had a higher average pretest score than the comparison group, the difference-in-difference approach is likely to underestimate the adjusted group mean difference. Otherwise, it is likely to overestimate the adjusted group mean difference.

the use of unadjusted within-group SD. Therefore, we cannot compute Hedges's g with the F -statistic from an ANCOVA in the same way as we compute g with the F -statistic from an ANOVA. If the pretest-posttest correlation is known, however, we could derive Hedges's g from the ANCOVA F -statistic as follows:

$$\text{Hedges's } g = \sqrt{\frac{F(n_1 + n_2)(1 - r^2)}{n_1 n_2}}, \quad (9)$$

$$\text{Hedges's } g = \text{sqrt}[F(n_1 + n_2)(1 - r^2)/n_1 n_2]$$

where r is the pretest-posttest correlation, and n_1 and n_2 are the sample sizes for the intervention group and the comparison group respectively.

ES Computation Based on Results from Cluster-Level Analyses

The ES computation methods described above are all based on student-level analyses, which are appropriate analytic approaches for studies with student-level assignment. The case is more complicated, however, for studies with assignment at the cluster level (e.g., assignment of teachers, classrooms, or schools to conditions), where data may have been analyzed at the student level, the cluster level, or through multilevel analyses. Although there has been a consensus in the field that multilevel analysis should be used to analyze clustered data (e.g., Bloom, Bos, & Lee, 1999; Donner & Klar, 2000; Flay & Collins, 2005; Murray, 1998; and Snijders & Bosker, 1999), student-level analyses and cluster-level analyses of such data still frequently appear in the research literature despite their problems.

The main problem with student-level analyses in studies with cluster-level assignment is that they violate the independence of observations assumption underlying traditional hypothesis tests and result in underestimated standard errors and inflated statistical significance (see Section IV for details about how to correct for such bias). The estimate of the group mean difference in such analyses, however, is unbiased and therefore can be appropriately used to compute the student-level ES using methods explained in the previous sections.

For studies with cluster-level assignment, analyses at the cluster level, or aggregated analyses, are also problematic. Other than the loss of power and increased Type II error, potential problems with aggregated analysis include shift of meaning and ecological fallacy (i.e., relationships between aggregated variables cannot be used to make assertions about the relationships between individual-level variables), among others (Aitkin & Longford, 1986; Snijders & Bosker, 1999). Such analyses also pose special challenges to ES computation during WWC reviews. In the remainder of this section, we discuss these challenges and describe WWC's approach to handling them during WWC reviews.

How to compute student-level ESs for studies with cluster-level analyses

For studies that only reported findings from cluster-level analyses, it might be tempting to compute ESs using cluster-level means and SDs. This, however, is not appropriate for the purpose of the WWC reviews for at least two reasons. First, because cluster-level SDs are

typically much smaller than student-level SDs,⁴ ESs based on cluster-level SDs will be much larger than, and therefore incomparable with, student-level ESs that are the focus of WWC reviews. Second, the criterion for “substantively important” effects in the WWC Intervention Rating Scheme (ES of at least 0.25) was established specifically for student-level ESs, and does not apply to cluster-level ESs. Moreover, there is not enough knowledge in the field as yet for judging the magnitude of cluster-level effects. A criterion of “substantively important” effects for cluster-level ESs, therefore, cannot be developed for intervention rating purposes. An intervention rating of potentially positive effects based on a cluster-level ES of 0.25 or greater (i.e., the criterion for student-level ESs) would be misleading.

In order to compute the student-level ESs, we need to use the student-level means and SDs on the findings. This information, however, is often not reported in studies with cluster-level analyses. If the study authors could not provide student-level means, the review team may use cluster-level means (i.e., mean of cluster means) to compute the group mean difference for the numerator of student-level ESs if: (1) the clusters were of equal or similar sizes, (2) the cluster means were similar across clusters, or (3) it is reasonable to assume that cluster size was unrelated to cluster means. If any of the above conditions holds, then group means based on cluster-level data would be similar to group means based on student-level data, and hence could be used for computing student-level ESs. If none of the above conditions holds, however, the review team will have to obtain the group means based on student-level data in order to compute the student-level ESs.

While it is possible to compute the numerator (i.e., group mean difference) for student-level ESs based on cluster-level findings for most studies, it is generally much less feasible to compute the denominator (i.e., pooled SD) for student-level ESs based on cluster-level data. If the student-level SDs are not available, we could compute them based on the cluster-level SDs and the actual intra-class correlation (ICC) (student-level SD = (cluster-level SD)/sqrt(ICC)). Unfortunately, the actual ICCs for the data observed are rarely provided in study reports. Without knowledge about the actual ICC, one might consider using a default ICC, which, however, is not appropriate, because the resulting ES estimate would be highly sensitive to the value of the default ICC and might be seriously biased even if the difference between the default ICC and the actual ICC is not large.

Another reason that the formula for deriving student-level SDs (student-level SD = (cluster-level SD)/sqrt(ICC)) is unlikely to be useful is that the cluster-level SD required for the computation was often not reported either. Note that the cluster-level SD associated with the ICC is not exactly the same as the observed SD of cluster means that were often reported in studies with cluster-level analyses, because the latter reflects not only the true cluster-level variance, but also part of the random variance within clusters (Raudenbush & Liu, 2000; Snijder & Bosker, 1999).

It is clear from the above discussion that in most cases, requesting student-level data, particularly student-level SDs, from the study authors will be the only way that allows us to compute the student-level ESs for studies only reporting cluster-level findings. If the study authors could not provide the student-level data needed, then we would not be able to compute

⁴ Cluster-level SD = (student-level SD)*sqrt(ICC).

the student-level ESs. Nevertheless, such studies will not be automatically excluded from the WWC reviews, but could still potentially contribute to intervention ratings as explained below.

How to handle studies with cluster-level analyses in intervention ratings if the student-level ESs could not be computed

A study's contribution to the effectiveness rating of an intervention depends mainly on three factors: the quality of the study design, the statistical significance of the findings, and the size of the effects. For studies that only reported cluster-level findings, the quality of their design is not affected by whether student-level ESs could be computed or not. Such studies could still meet WWC evidence standards with or without reservations and be included in intervention reports even if student-level ESs were not available.

While cluster-level ESs cannot be used in intervention ratings, the statistical significance of cluster-level findings could contribute to intervention ratings. Cluster-level analyses tend to be underpowered, hence estimates of the statistical significance of findings from such analyses tend to be conservative. Therefore, significant findings from cluster-level analyses would remain significant had the data been analyzed using appropriate multilevel models, and should be taken into account in intervention ratings. The size of the effects based on cluster-level analyses, however, could not be considered in determining "substantively important" effects in intervention ratings for reasons described above. In WWC's intervention reports, cluster-level ESs will be excluded from the computation of domain average ESs and improvement indices, both of which will be based exclusively on student-level findings.

ES Computation Based on Results from HLM Analyses in Studies with Cluster-Level Assignment

As explained in the previous section, multilevel analysis is generally considered the preferred method for analyzing data of from studies with cluster-level assignment. With recent methodological advances, multilevel analysis has gained increased popularity in education and other social science fields. More and more researchers have begun to employ the hierarchical linear modeling (HLM) method to analyze data of a nested nature (e.g., students nested within classes and classes nested within schools) (Raudenbush & Bryk, 2002)⁵. Similar to student-level ANCOVA, HLM can also adjust for important covariates such as pretest when estimating an intervention's effect. Unlike student-level ANCOVA that assumes independence of observations, however, HLM explicitly takes into account the dependence among members within the same higher-level unit (e.g., the dependence among students within the same class). Therefore, the parameter estimates, particularly the standard errors, generated from HLM are less biased than those generated from ANCOVA when the data have a multilevel structure.

Hedges's g for intervention effects estimated from HLM analyses is defined in a similar way to that based on student-level ANCOVA: adjusted group mean difference divided by unadjusted pooled within-group SD. Specifically,

⁵ Multilevel analysis can also be conducted using other approaches, such as the SAS PROC MIXED procedure. Although different approaches to multilevel analysis may differ in their technical details, they are all based on similar ideas and underlying assumptions.

$$\text{Hedges's } g = \frac{\gamma}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}}, \quad (10)$$

$$\text{Hedges's } g = \gamma / \sqrt{\frac{[(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2]}{(n_1 + n_2 - 2)}}$$

Where γ is the HLM coefficient for the intervention's effect, which represents the group mean difference adjusted for both level-1 and level-2 covariates, if any;⁶ n_1 and n_2 are the student sample sizes, and S_1 and S_2 the unadjusted student-level SDs for the intervention group and the comparison group respectively.

One thing to note about the denominator of Hedges's g based on HLM results is that the level-1 variance, also called "within-group variance," estimated from a typical two-level HLM analysis is not the same as the conventional unadjusted pooled within-group variance that should be used in ES computation. The within-group variance from an HLM model that incorporates level-1 covariates has been adjusted for these covariates. Even if the within-group variance is based on an HLM model that does not contain any covariates (i.e., a fully-unconditional model), it is still not appropriate for ES computation, because it does not include the variance between level-2 units within each study condition that is part of the unadjusted pooled within-group variance. Therefore, the level-1 within-group variance estimated from an HLM analysis tends to be smaller than the conventional unadjusted pooled within-group variance, and would thus lead to an overestimate of the ES if used in the denominator of the ES.

The ES computations for continuous outcomes explained above pertain to individual findings within a given outcome domain examined in a given study. If the study authors assessed the intervention's effects on multiple outcome measures within a given domain, the WWC computes a domain average ES as a simple average of the ESs across all individual findings within the domain.

II. Effect Size Computation for Dichotomous Outcomes

Although not as common as continuous outcomes, dichotomous outcomes are sometimes used in studies of educational interventions. Examples include dropout vs. stay in school; grade promotion vs. retention; and pass vs. fail a test. Group mean differences, in this case, appear as differences in proportions or differences in the probability of the occurrence of an event. The ES measure of choice for dichotomous outcomes is odds ratio, which has many statistical and practical advantages over alternative ES measures such as the difference between two probabilities, the ratio of two probabilities, and the phi coefficient (Fleiss, 1994; Lipsey & Wilson, 2001).

⁶ The level-2 coefficients are adjusted for the level-1 covariates under the condition that the level-1 covariates are either uncentered or grand-mean centered, which are the most common centering options in an HLM analysis (Raudenbush & Bryk, 2002). The level-2 coefficients are not adjusted for the level-1 covariates if the level-1 covariates are group-mean centered. For simplicity purposes, the discussion here is based on a two-level framework (i.e., students nested with clusters). The idea could easily be extended to a three-level model (e.g., students nested with teachers who were in turn nested within schools).

The measure of odds ratio builds on the notion of odds. For a given study group, the odds for the occurrence of an event are defined as follows:

$$\text{Odds} = \frac{p}{1-p}, \quad (11)$$

$$\text{Odds} = p/(1-p)$$

where p is the probability of the occurrence of an event within the group. Odds ratio (OR) is simply the ratio between the odds for the two groups compared:

$$\text{OR} = \frac{\text{Odds}_1}{\text{Odds}_2} = \frac{p_1(1-p_2)}{p_2(1-p_1)}, \quad (12)$$

$$\text{OR} = \text{Odds}_1/\text{Odds}_2 = [p_1(1-p_2)]/[p_2(1-p_1)]$$

where p_1 and p_2 are the probabilities of the occurrence of an event for the intervention group and the comparison group respectively.

As is the case with ES computation for continuous variables, the WWC computes ESs for dichotomous outcomes based on student-level data in preference to aggregate-level data for studies that had a multi-level data structure. The probabilities (p_1 and p_2) used in calculating the odds ratio represent the proportions of students demonstrating a certain outcome among students across all teachers/classrooms or schools in each study condition, which are likely to differ from the probabilities based on aggregate-level data (e.g., means of school-specific probabilities) unless the classrooms or schools in the sample were of similar sizes.

Following conventional practice, the WWC transforms odds ratio calculated based on Equation (12) to logged odds ratio (LOR) (i.e., the natural log of the odds ratio) to simplify statistical analyses:

$$\text{LOR} = \ln(\text{OR}) \quad (13)$$

The logged odds ratio has a convenient distribution form, which is approximately normal with a mean of 0 and a SD of $\pi/\sqrt{3}$, or 1.81.

The logged odds ratio can also be expressed as the difference between the logged odds, or logits, for the two groups compared. Equivalent to Equation (13),

$$\text{LOR} = \ln(\text{Odds}_1) - \ln(\text{Odds}_2), \quad (14)$$

$$\text{LOR} = \ln(\text{Odds}_1) - \ln(\text{Odds}_2)$$

which shows more clearly the connection between the logged odds ratio index and the standardized mean difference index (Hedges's g) for ESs. To make logged odds ratio comparable to standardized mean difference and thus facilitate the synthesis of research findings based on different types of outcomes, researchers have proposed a variety of methods for "standardizing" logged odds ratio. Based on a Monte Carlo simulation study of seven different types of ES indices for dichotomous outcomes, Sanchez-Meca, Marin-Martinez, and Chacon-Moscoso

(2003) concluded that the ES index proposed by Cox (1970) is the least biased estimator of the population standardized mean difference, assuming an underlying normal distribution of the outcome. The WWC, therefore, has adopted the Cox index as the default ES measure for dichotomous outcomes. The computation of the Cox index is straightforward:

$$\text{LOR}_{\text{Cox}} = \text{LOR}/1.65 \quad (15)$$

The above index yields ES values very similar to the values of Hedges's *g* that one would obtain if group means, SDs, and sample sizes were available—assuming that the dichotomous outcome measure is based on an underlying normal distribution. Although the assumption may not always hold, as Sanchez-Meca and his colleagues (2003) note, primary studies in social and behavioral sciences routinely apply parametric statistical tests that imply normality. Therefore, the assumption of normal distribution is a reasonable conventional default.

III. Computation of the Improvement Index

In order to help readers judge the practical importance of an intervention's effect, the WWC translates ES into an "improvement index." The improvement index represents the difference between the percentile rank corresponding to the intervention group mean and the percentile rank corresponding to the comparison group mean (i.e., 50th percentile) in the comparison group distribution. Alternatively, the improvement index can be interpreted as the expected change in percentile rank for an average comparison group student if the student had received the intervention.

As an example, if an intervention produced a positive impact on students' reading achievement with an effect size of 0.25, the effect size could be translated to an improvement index of 10 percentile points. We could then conclude that the intervention would have led to a 10% increase in percentile rank for an average student in the comparison group, and that 60% (10% + 50%=60%) of the students in the intervention group scored above the comparison group mean.

Specifically, the improvement index is computed as follows:

(1) Convert the ES (Hedges's *g*) to Cohen's U3 index.

The U3 index represents the percentile rank of a comparison group student who performed at the level of an average intervention group student. An effect size of 0.25, for example, would correspond to a U3 of 60%, which means that an average intervention group student would rank at the 60th percentile in the comparison group. Equivalently, an average intervention group student would rank 10 percentile points higher than an average comparison group student, who, by definition, ranks at the 50th percentile.

Mechanically, the conversion of an effect size to a U3 index entails looking up on a table that lists the proportion of area under the standard normal curve for different values of *z*-scores, which can be found in the appendices of most statistics textbooks. For a given effect size, U3 has a value equal to the proportion of area under the normal curve below the value of the effect

size—under the assumptions that the outcome is normally distributed and that the variance of the outcome is similar for the intervention group and the comparison group.

(2) Compute:

$$\text{Improvement index} = U3 - 50\% \quad (16)$$

Given that $U3$ represents the percentile rank of an average intervention group student in the comparison group distribution, and that the percentile rank of an average comparison group student is 50%, the improvement index, defined as $(U3 - 50\%)$, would represent the difference in percentile rank between an average intervention group student and an average comparison group student in the comparison group distribution.

In addition to the improvement index for each individual finding, the WWC also computes a domain average improvement index for each study as well as a domain average improvement index across studies for each outcome domain. The domain average improvement index for each study is computed based on the domain average effect size for that study rather than as the average of the improvement indices for individual findings within that study. Similarly, the domain average improvement index across studies is computed based on the domain average effect size across studies, with the latter computed as the average of the domain average effect sizes for individual studies.

IV. Clustering Correction of the Statistical Significance of Effects Estimated with Mismatched Analyses

In order to adequately assess an intervention's effects, it is important to know not only the magnitude of the effects as indicated by ES, but also the statistical significance of the effects. The correct statistical significance of findings, however, is not always readily available, particularly in studies where the unit of assignment does not match the unit of analysis. The most common "mismatch" problem occurs when assignment was carried out at the cluster level (e.g., classroom or school level), whereas the analysis was conducted at the student level, ignoring the dependence among students within the same clusters. Although the point estimates of the intervention's effects based on such mismatched analyses are unbiased, the standard errors of the effect estimates are likely to be underestimated, which would lead to inflated Type I error and overestimated statistical significance.

In order to present a fair judgment about an intervention's effects, the WWC computes clustering-corrected statistical significance for effects estimated from mismatched analyses and the corresponding domain average effects based on Hedges' (2005) most recent work. As clustering correction will decrease the statistical significance (or increase the p-value) of the findings, non-significant findings from a mismatched analysis will remain non-significant after the correction. Therefore, the WWC only applies the correction to findings reported to be statistically significant by the study authors.

The basic approach to clustering correction is to first compute the t-statistic corresponding to the ES that ignores clustering, and then correct both the t-statistic and the

associated degrees of freedom for clustering based on sample sizes, number of clusters, and the intra-class correlation. The statistic significance corrected for clustering could then be obtained from the t-distribution with the corrected t-statistic and degrees of freedom. In the remainder of this section, we detail each step of the process.

(1) Compute the t-statistic for the ES ignoring clustering

This is essentially the reverse of Equation (4) that computes Hedges’s g based on t:

$$t = g \sqrt{\frac{n_1 n_2}{n_1 + n_2}},$$

$$t = g * \text{sqrt} [n_1 n_2 / (n_1 + n_2)] \tag{17}$$

where g is the ES that ignores clustering, and n₁ and n₂ are the sample sizes for the intervention group and the comparison group respectively for a given outcome. For domain average ESs, n₁ and n₂ are the average sample sizes for the intervention and comparison groups respectively across all outcomes within the domain

(2) Correct the above t-statistic for clustering

$$t_A = t \sqrt{\frac{(N - 2) - 2\left(\frac{N}{m} - 1\right)\rho}{(N - 2)\left[1 + \left(\frac{N}{m} - 1\right)\rho\right]}}, \tag{18}$$

$$t_A = t * \text{sqrt} \{ [(N-2) - 2(N/m-1)\rho] / [(N-2)(1+(N/m-1)\rho)] \}$$

where N is the total sample size at the student level (N = n₁ + n₂), m is the total number of clusters in the intervention and comparison groups (m = m₁ + m₂, m₁ and m₂ are the number of clusters in each of the two groups), and ρ is the intra-class correlation (ICC) for a given outcome.

The value of ICC, however, is often not available from the study reports. Based on empirical literature in the field of education, the WWC has adopted a default ICC value of .20 for achievement outcomes and .10 for behavioral and attitudinal outcomes. The PIs and review teams may set different defaults with explicit justification in terms of the nature of the research circumstances or the outcome domain.

For domain average ESs, the ICC used in Equation (18) is the average ICC across all outcomes within the domain. If the number of clusters in the intervention and comparison groups differs across outcomes within a given domain, the total number of clusters (m) used for computing the corrected t-statistic will be based on the largest number of clusters in both groups across outcomes within the domain (i.e., largest m₁ and m₂ across outcomes). This gives the study the benefit of the doubt by crediting the measure with the most statistical power, so that the WWC’s rating of interventions will not be unduly conservative.

(3) Compute the degrees of freedom associated with the t-statistics corrected for clustering:

$$h = \frac{\left[(N - 2) - 2\left(\frac{N}{m} - 1\right)\rho \right]^2}{(N - 2)(1 - \rho)^2 + \frac{N}{m}\left(N - 2\frac{N}{m}\right)\rho^2 + 2\left(N - 2\frac{N}{m}\right)\rho(1 - \rho)} \quad (19)$$

$$h = [(N-2)-2(N/m-1)\rho]^2 / [(N-2)(1-\rho)^2 + (N/m)(N-2N/m)\rho^2 + 2(N-2N/m)\rho(1-\rho)]$$

(4) Obtain the statistical significance of the effect corrected for clustering

The clustering-corrected statistical significance (p-value) is determined based on the t-distribution with corrected t-statistic (t_A) and the corrected degrees of freedom (h). This p-value can either be looked up in a t-distribution table that can be found in the appendices of most statistical textbooks, or computed using the t-distribution function in Excel: $p = \text{TDIST}(t_A, h, 2)$.

Further information on this topic is available in the WWC's technical papers on the WWC Tutorial on Mismatch Between Unit of Assignment and Unit of Analysis and the WWC Intervention Rating Scheme.

V. Benjamini-Hochberg Correction of the Statistical Significance of Effects Estimated with Multiple Comparisons

In addition to clustering, another factor that may inflate Type I error and the statistical significance of findings is when study authors perform multiple hypothesis tests simultaneously. The traditional approach to addressing the problem is the Bonferroni method, which lowers the critical p-value for individual comparisons by a factor of $1/m$, with m being the total number of comparisons made. The Bonferroni method, however, has been shown to be unnecessarily stringent for many practical situations; therefore the WWC has adopted a more recently developed method to correct for multiple comparisons or multiplicity—the Benjamini-Hochberg (BH) method (Benjamini & Hochberg, 1995). The BH method adjusts for multiple comparisons by controlling false discovery rate (FDR) instead of familywise error rate (FWER). It is less conservative than the traditional Bonferroni method, yet still provides adequate protection against Type I error in a wide range of applications. Since its conception in the 1990's, there has been growing evidence showing that the FDR-based BH method may be the best solution to the multiple comparisons problem in many practical situations (Williams, Jones, & Tukey, 1999)

As is the case with clustering correction, the WWC only applies the BH correction to statistically significant findings, because non-significant findings will remain non-significant after correction. For findings based on analyses where the unit of analysis was properly aligned with the unit of assignment, we use the p-values reported in the study for the BH correction. If the exact p-values were not available, but the ESs could be computed, we will convert the ESs to t-statistics (see Equation (4)), and then obtain the corresponding p-values.⁷ For findings based on

⁷ The p-values corresponding to the t-statistics can either be looked up in a t-distribution table, or computed using the t-distribution function in Excel: $p = \text{TDIST}(t, df, 2)$, where df is the degrees of freedom, or the total sample size minus 2 for findings from properly aligned analyses.

mismatched analyses, we first correct the author-reported p-values for clustering, and then use the clustering-corrected p-values for the BH correction.

Although the BH correction procedure described above was originally developed under the assumption of independent test statistics (Benjamini & Hochberg, 1995), Benjamini and Yekutieli (2001) point out that it also applies to situations where the test statistics have positive dependency, and that the condition for positive dependency is general enough to cover many problems of practical interest. For other forms of dependency, a modification of the original BH procedure could be made, which, however, is “very often not needed, and yields too conservative a procedure” (p. 1183).⁸ Therefore, the WWC has chosen to use the original BH procedure rather than its more conservative modified version as the default approach to correcting for multiple comparisons. In the remainder of this section, we describe the specific procedures for applying the BH correction in three types of situations: studies that tested multiple outcome measures in the same outcome domain with a single comparison group, studies that tested a given outcome measure with multiple comparison groups, and studies that tested multiple outcome measures in the same outcome domain with multiple comparison groups.

Benjamini-Hochberg Correction of the Statistical Significance of Effects on Multiple Outcome Measures within the Same Outcome Domain Tested with a Single Comparison Groups

The most straightforward situation that may require the BH correction is when the study authors assessed an intervention’s effect on multiple outcome measures within the same outcome domain using a single comparison group. For such studies, the review team needs to first check whether the study authors’ analyses already took into account multiple comparisons (e.g., through a proper multivariate analysis). If so, obviously no further correction is necessary. If the authors did not address the multiple comparison problem in their analyses, then the review team will need to correct the statistical significance of the authors’ findings using the BH method. For studies that examined measures in multiple outcome domains, the BH correction will be applied to the set of findings within the same domain rather than across different domains. Assuming that the BH correction is needed, the review team will apply the BH correction to multiple findings within a given outcome domain tested with a single comparison group as follows:

(1) Rank order statistically significant findings within the domain in ascending order of the p-values, such that: $p_1 \leq p_2 \leq p_3 \leq \dots \leq p_m$, with m being the number of significant findings within the domain.

(2) For each p-value (p_i), compute:

$$p_i' = \frac{i * \alpha}{M}, \tag{20}$$

$$[p_i' = i * \alpha / M]$$

⁸ The modified version of the BH procedure uses α over the sum of the inverse of the p-value ranks across the m comparisons (i.e., $\alpha / \sum_{i=1}^m \frac{1}{i}$) instead of α in Equation (20).

where i is the rank for p_i , with $i = 1, 2, \dots, m$; M is the total number of findings within the domain reported by the WWC; and α is the target level of statistical significance.

Note that the M in the denominator of Equation (20) may be less than the number of outcomes that the study authors actually examined in their study for two reasons: (1) the authors may not have reported findings from the complete set of comparisons that they had made, and (2) certain outcomes assessed by the study authors may be deemed irrelevant to the WWC’s review. The target level of statistical significance, α , in the numerator of Equation (20) allows us to identify findings that are significant at this level after correction for multiple comparisons. The WWC’s default value of α is 0.05, although other values of α could also be specified. If, for instance, α is set at 0.01 instead of 0.05, then the results of the BH correction would indicate which individual findings are statistically significant at the 0.01 level instead of the 0.05 level after taking multiple comparisons into account.

(3) Identify the largest i —denoted by k —that satisfies the condition: $p_i \leq p_i'$. This establishes the cut-off point, and allows us to conclude that all findings with p-values smaller than or equal to p_k are statistically significant, and findings with p-values greater than p_k are not significant at the pre-specified level of significance ($\alpha = 0.05$ by default) after correction for multiple comparisons.

One thing to note is that, unlike clustering correction, which produces a new p-value for each corrected finding, the BH correction does not generate a new p-value for each finding, but rather only indicates whether the finding is significant or not at the pre-specified level of statistical significance after the correction. As an illustration, suppose a researcher compared the performance of the intervention group and the comparison group on eight measures in a given outcome domain, and reported six statistically significant effects and two non-significant effects based on properly aligned analyses. To correct the significance of the findings for multiple comparisons, we would first rank-order the p-values of the six author-reported significant findings in the first column of Table 1, and list the p-value ranks in the second column. We then compute $p_i' = i * \alpha / M$, using Equation (20) with $M = 8$ and $\alpha = 0.05$, and record the values in the third column. Next, we identify k , the largest i , that meets the condition: $p_i \leq p_i'$. In this example, $k = 4$, and $p_k = 0.014$. Thus, we can claim that the four finding associated with a p-value of 0.014 or smaller are statistically significant at the 0.05 level after correction for multiple comparisons. The other two findings, although reported as being statistically significant, are no longer significant after the correction.

Table 1. An Illustration of Applying the Benjamini-Hochberg Correction for Multiple Comparisons

Author-reported or clustering-corrected p-value (p_i)	P-value rank (i)	$p_i' = \frac{i * (0.05)}{8}$ $p_i' = i * (0.05)/8$	$p_i \leq p_i'?$	Statistical significance after BH correction ($\alpha = .05$)
0.002	1	0.006	Yes	significant
0.009	2	0.013	Yes	significant
0.011	3	0.019	Yes	significant
0.014	4	0.025	Yes	significant

0.034	5	0.031	No	n.s.
0.041	6	0.038	No	n.s.

Note. n.s.: not statistically significant.

Benjamini-Hochberg Correction of the Statistical Significance of Effects on a Given Outcome Tested with Multiple Comparison Groups

The above discussion pertains to the multiple comparisons problem when the study authors tested multiple outcomes within the same domain with a single comparison group. Another type of multiple comparisons problem occurs when the study authors tested an intervention's effect on a given outcome by comparing the intervention group with multiple comparison groups. The WWC's recommendation for handling such studies is as follows:

1. In consultation with the PI and the study authors if needed, the review team selects a single comparison group that best represented the "business as usual" condition or that is considered most relevant to the WWC's review. Only findings based on comparisons between the intervention group and this particular comparison group will be included in the WWC's review. Findings involving the other comparison groups will be ignored, and the multiplicity due to one intervention group being compared with multiple comparison groups could also be ignored.
2. If the PI and the review team believe that it is appropriate to combine the multiple comparison groups, and if adequate data are available for deriving the means and SDs of the combined group, the team may present the findings based on comparisons of the intervention group and the combined comparison group instead of findings based on comparisons of the intervention group and each individual comparison group. The kind of multiplicity due to one intervention group being compared with multiple comparison groups will no longer be an issue in this approach.

The PI and the review team may judge the appropriateness of combining multiple comparison groups by considering whether there was enough common ground among the different comparison groups that warrant such a combination; and particularly, whether the study authors themselves conducted combined analyses or indicated the appropriateness, or the lack thereof, of combined analyses. In cases where the study authors did not conduct or suggest combined analyses, it is advisable for the review team to check with the study authors before combining the data from different comparison groups.

3. If the PI and the review team believe that neither of the above two options is appropriate for a particular study, and that findings from comparisons of the intervention group and each individual comparison group should be presented, they need to make sure that the findings presented in the WWC's intervention report are corrected for multiplicity due to multiple comparison groups if necessary. The review team needs to first check the study report or check with the study authors whether the comparisons of the multiple groups were based on a proper statistical test that already took multiplicity into account (e.g., Dunnett's test (Dunnett, 1955), the Bonferroni method (Bonferroni, 1935), Scheffe's test (1953), and Tukey's HSD test (1949)). If so, then there would be no need for further corrections.

It is also advisable for the team to check with the study authors regarding the appropriateness of correcting their findings for multiplicity due to multiple comparison groups, as the authors might have theoretical or empirical reasons for considering the findings from comparisons of the intervention group and a given comparison group without consideration of other comparisons made within the same study. If the team decides that multiplicity correction is necessary, they will apply such correction using the BH method in the same way as they would apply it to findings on multiple outcomes within the same domain tested with a single comparison group as described in the previous section.

Benjamini-Hochberg Correction of the Statistical Significance of Effects on Multiple Outcome Measures in the Same Outcome Domain Tested with Multiple Comparison Groups

A more complicated multiple comparison problem arises when a study tested an intervention's effect on multiple outcome measures in a given domain with multiple comparison groups. The multiplicity problem thus may originate from two sources. Assuming that both types of multiplicity need to be corrected, the review team will apply the BH correction in accordance with the following three scenarios.

Scenario 1: The study authors's findings did not take into account either type of multiplicity.

In this case, the BH correction will be based on the total number of comparisons made. For example, if a study compared one intervention group with two comparison groups on five outcomes in the same domain without taking multiplicity into account, then the BH correction will be applied to the 10 individual findings based on a total of 10 comparisons.

Scenario 2: The study authors's findings took into account the multiplicity due to multiple comparisons, but not the multiplicity due to multiple outcomes.

In some studies, the authors may have performed a proper multiple comparison test (e.g., Dunnett's test) on each individual outcome that took into account the multiplicity due to multiple comparison groups. For such studies, the WWC will only need to correct the findings for the multiplicity due to multiple outcomes. Specifically, separate BH corrections will be made to the findings based on comparisons involving different comparison groups. With two comparison groups, for instance, the review team will apply the BH correction to the two sets of findings separately—one set of findings (one finding for each outcome) for each comparison group.

Scenario 3: The study authors's findings took into account the multiplicity due to multiple outcomes, but not the multiplicity due to multiple comparison groups.

Although this scenario may be relatively rare, it is possible that the study authors performed a proper multivariate test (e.g., MANOVA or MANCOVA) to compare the intervention group with a given comparison group that took into account the multiplicity due to multiple outcomes, and performed separate multivariate tests for different comparison groups. For such studies, the review team will only need to correct the findings for multiplicity due to

multiple comparison groups. Specifically, separate BH corrections will be made to the findings on different outcomes. With five outcomes and two comparison groups, for instance, the review team will apply the BH correction to the five sets of findings separately—one set of findings (one finding for each comparison group) for each outcome measure.

The decision rules for the three scenarios described above are summarized in the table below.

Table 2. Decision Rules for Correcting the Significance Levels of Findings from Studies That had a Multiple Comparison Problem due to Multiple Outcomes in a Given Domain and/or Multiple Comparison Groups, by Scenario

Authors' Analyses	Benjamini-Hochberg Correction
1. Did not correct for multiplicity from any source	<ul style="list-style-type: none"> • BH correction to all 10 individual findings
2. Corrected for multiplicity due to multiple comparison groups only	<ul style="list-style-type: none"> • BH correction to the 5 findings based on T vs. C1 comparisons • BH correction to the 5 findings based on T vs. C2 comparisons
3. Corrected for multiplicity due to multiple outcomes only	<ul style="list-style-type: none"> • BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O1 • BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O2 • BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O3 • BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O4 • BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O5

Note. T: treatment (intervention) group;
 C1 and C2: comparison groups 1 and 2;
 O1, O2, O3, O4, and O5: five outcome measures within a given outcome domain.

On a final note, although the BH corrections are applied in different ways to the individual study findings in different scenarios, such differences do not affect the way in which the intervention rating is determined. In all three scenarios of the above example, the 10 findings will be presented in a single outcome domain, and the characterization of the intervention's effects for this domain in this study will be based on the corrected statistical significance of each individual finding as well as the magnitude and statistical significance of the average effect size across of the 10 individual findings within the domain.

References

Aitkin, M., & Longford, N. (1986). Statistical modeling issues in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society, A*(149), 1-43.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57(1), 289-300.

Benjamini, Y., & Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4), 1165-1188.

Bloom, H. S., Bos, J.M., & Lee, S.W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review*, 234:445- 69.

- Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*. Rome: Italy, pp. 13–60.
- Cooper, H. (1998). *Synthesizing research: A guide for literature review*. Thousand Oaks, CA: Sage Publications.
- Cox, D.R. (1970). *Analysis of binary data*. New York: Chapman & Hall/CRC.
- Donner, A. and Klar, N. (2000) Design and Analysis of Cluster Randomized Trials in Health Research. London: Arnold Publishing.
- Dunnett, C. (1955). A multiple comparisons procedure for comparing several treatments with a control. *Journal of American Statistical Association*, 50, 1096–1121.
- Flay, B. R., & Collins, L. M. (2005). Historical review of school-based randomized trials for evaluating problem behavior prevention programs. *The Annals of the American Academy of Political and Social Science*, 599, 147-175.
- Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245-260). New York: Russel Sage Foundation.
- Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, 6, 107–128.
- Hedges, L. V. (2005). Correcting a significance test for clustering. Unpublished manuscript.
- Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage Publications.
- Murray, D. M. (1998). *Design and analysis of group-randomized trials*. (Vol. 27). New York: Oxford University Press.
- Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, 5(2), 199-213.
- Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science*, 11(6), 446–453.
- Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.
- Sanchez-Meca, J., Marin-Martinez, F., & Chacon-Moscoso, S. (2003). Effect-size indices for dichotomous outcomes in meta-analysis. *Psychological Methods*, 8(4), 448–467.

Scheffe, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, 40, 87–104.

Snijders, T., & Bosker, R. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London: Sage Publications.

Tukey, J. (1949). Comparing individual means in the analysis of variance. *Biometrika*, 5, 99–114.

Williams, V. S. L., Jones, L. V., & Tukey, J. W. (1999). Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, 24(1), 42-69.