What Works Clearinghouse™

Procedures Handbook

Version 4.0, 7/14/2017 Draft

CONTENTS

TABLES

FIGURES

I. INTRODUCTION

It is critical that education decision makers have access to the best evidence about the effectiveness of education products, programs, policies, and practices. However, it can be difficult, time-consuming, and costly for decision makers to access and draw conclusions from relevant studies about the effectiveness of these interventions. The What Works Clearinghouse (WWC) addresses the need for credible, succinct information by identifying existing research on education interventions, assessing the quality of this research, and summarizing and disseminating the evidence from studies that meet WWC standards.

The WWC is an initiative of the U.S. Department of Education's Institute of Education Sciences (IES), which was established under the Education Sciences Reform Act of 2002. It is an important part of IES's strategy to use rigorous and relevant research, evaluation, and statistics to improve our nation's education system. The mission of the WWC is to be a **central and trusted source of scientific evidence for what works in education.** The WWC examines research about interventions that focus on improving educationally relevant outcomes, including those for students and educators.

The WWC systematic review process is the basis of many of its products, enabling the WWC to use consistent, objective, and transparent standards and procedures in its reviews, while also ensuring comprehensive coverage of the relevant literature. The WWC systematic review process consists of five steps:

1. *Developing the review protocol.* A formal review protocol is developed for each review effort, including one for each WWC topic area (e.g., adolescent literacy, primary mathematics, or charter schools) to define the parameters for the research to be included within the scope of the review (e.g., population characteristics and types of interventions); the literature search (e.g., search terms and databases); and any topic-specific applications of the standards (e.g., acceptable thresholds for sample attrition and characteristics for group equivalence).
2. *Identifying relevant literature.* Studies are gathered through a comprehensive search of published and unpublished publicly available research literature. The search uses electronic databases, outreach efforts, and public submissions.
3. *Screening studies.* Manuscripts initially are screened for eligibility to determine whether they report on original research, provide potentially credible evidence of an intervention's effectiveness, and fall within the scope of the review protocol.
4. *Reviewing studies.* Every eligible study is reviewed against WWC standards. The WWC uses a structured review process to assess the causal validity of findings reported in education effectiveness research. The WWC standards focus on the causal validity within the study sample (*internal* validity) rather than the extent to which the findings might be replicated in other settings (*external* validity).
5. *Reporting on findings.* The details of the review and its findings are summarized on the WWC website, and often in a WWC publication. For many of its products, the WWC combines findings from individual studies into summary measures of effectiveness, including the magnitude of findings and the extent of evidence.

In addition, the WWC reviews some studies outside of the systematic review process. These reviews are also guided by a review protocol and use the same WWC standards and reporting procedures.

This *What Works Clearinghouse Procedures Handbook (Version 4.0)* provides a detailed description of the procedures used by the WWC in the systematic review process—specifically, Steps 1–3 and Step 5 described above. A separate *What Works Clearinghouse Standards Handbook* describes Step 4, including the standards used by the WWC to review studies and assign one of the following three study ratings indicating the credibility of evidence from the study: *Meets WWC Design Standards Without Reservations*, *Meets WWC Design Standards With Reservations*, or *Does Not Meet WWC Design Standards*. Taken together, these two documents replace the single document used since March 2014, the *What Works Clearinghouse Procedures and Standards Handbook (Version 3.0)*.

In general, this new *Procedures Handbook* contains the same procedures that were included in the *Procedures and Standards Handbook (Version 3.0)*. However, in addition to changes to the organization, the following substantive updates were made:

- **The *Handbook* includes additional clarification of procedures.** The additional clarification of the procedures is intended to support consistency across reviews, and includes new discussion of how the WWC defines studies and conducts its correction for multiple comparisons within studies.
- **The *Handbook* includes updated formulas for calculating statistical significance of findings.** The formula for continuous outcomes includes a new small sample size adjustment, and a separate formula is provided for dichotomous outcomes.
- **The *Handbook* includes procedures for reporting findings from randomized controlled trials that present complier average causal effects.** A new appendix describes how the WWC reports findings and statistical significance from studies with complier average causal effect estimates.

The remainder of the document is organized as follows: Chapter II describes the steps that the WWC uses to develop a review protocol. Chapter III describes how the WWC identifies the relevant literature. Chapter IV describes the screening process to determine if a study is eligible for review, and Chapter V describes the procedures used to review eligible studies. Chapter VI describes how the WWC summarizes evidence of effectiveness. Organizational procedures used by the WWC to ensure an independent, systematic, and objective review are described in the appendices.

As the WWC uses and applies the procedures in this *Procedures Handbook*, reviewers may occasionally need additional guidance. If necessary, the WWC will produce guidance documents for reviewers to provide clarification and interpretation of procedures and support consistency across reviews. This WWC reviewer guidance will clarify how these procedures should be implemented in situations where the current *Procedures Handbook* is not sufficiently specific to ensure consistent reviews.

As the WWC continues to refine and develop procedures, the *Procedures Handbook* will be revised to reflect these changes. Readers who want to provide feedback on the *Procedures Handbook,* or the WWC in general, may contact us at http://ies.ed.gov/ncee/wwc/help.

II. DEVELOPING THE REVIEW PROTOCOL

Prior to conducting a systematic review or other review effort, the WWC develops a formal review protocol to guide the review. The WWC develops a review protocol after a new topic area has been prioritized for review (see Appendix A). Because research on education covers a wide range of topics, interventions, and outcomes, a review protocol must describe what studies are eligible for review, how the WWC will search for them, and how they will be reviewed. The protocol defines the types of interventions that fall within the scope of the review, the population on which the review focuses, the keyword search terms, the parameters of the literature search, and any review-specific applications of the standards. Specifically, WWC protocols include guidance on the following issues:

- *Purpose statement.* All WWC review protocols begin with a description of the general purpose of the product. Protocols for some review efforts also provide background on the topic of focus and describe the goals of the review.
- *Key definitions.* Protocols define key terms and concepts that are specific to the substance and goal of the review.
- *Procedures for conducting the literature search.* Each protocol includes a list of the keywords and related terms that will be used in searching the literature and a list of the databases to search (see Appendix B for a sample list of keywords and search terms). A protocol also may provide special instructions regarding searching of the "gray literature," including public submissions to the WWC through the website or staff, research conducted and disseminated by distributors/developers of interventions, unpublished literature identified through prior WWC and non-WWC reviews and syntheses, unpublished research identified through listservs, and studies posted on organizational websites.
- *Eligibility criteria.* Protocols for all WWC products specify the criteria for determining whether a study is eligible for inclusion in the review. The review team leadership (lead methodologist and content expert, described further in Appendix C) makes decisions about key parameters, such as eligible population groups, types of interventions, study characteristics, and outcomes of interest. Examples of review-specific parameters commonly defined in the review protocols include the following:
  - *Eligible populations*. Protocols specify grade or age ranges (e.g., grades 1–4) and sample characteristics (e.g., over half the sample consist of students who are not native English speakers) for eligible student populations, along with subgroups of interest to the review.
  - *Eligible interventions.* Protocols provide descriptions of the types of interventions that fall within the bounds of the review, including the nature of the intervention (e.g., textbook-based literacy programs); the settings in which the intervention is delivered (e.g., regular classrooms or as a supplement to the regular school day); and whether the intervention is a "branded" product.
  - *Eligible research.* Protocols define the scope of research eligible to be included in the review based on characteristics such as time frame, language, and location.

  - *Eligible outcomes.* Protocols describe a set of domains containing outcomes of interest for the review (e.g., mathematics achievement or problem behavior).

- *Evidence standards.* The WWC uses the same design standards to review all eligible studies, as detailed in the *Standards Handbook*. However, within those standards, some parameters vary across reviews and must be specified in the protocol. These include the choice of boundary separating acceptable and unacceptable levels of sample attrition, the measures on which studies must demonstrate baseline equivalence, and some parameters related to cluster-assignment studies. Each of the items specified must be applied consistently for all studies that fall within the scope of the protocol.

III. IDENTIFYING RELEVANT LITERATURE

After a review protocol has been developed, and a topic for the systematic review has been prioritized (see Appendix A), the next step in the systematic review process is to conduct a *systematic and comprehensive search* for relevant literature. A literature search is *systematic* when it uses well-specified search terms and processes in order to identify studies that may be relevant, and it is *comprehensive* when a wide range of available databases, websites, and other sources is searched for studies on the effects of an intervention.

After a review protocol is established for a WWC systematic review, studies are gathered through a comprehensive search of published and unpublished research literature, including submissions from intervention distributors/developers, researchers, and the public to the WWC Help Desk. Only studies written in English that are publicly available (accessible on the web or available through a publication, such as a journal) at the time of the literature search are eligible for WWC review. Additionally, Masters' and Education Specialists' theses are ineligible for review. The WWC also reviews some individual studies outside of the systematic review process (see Appendix A for more detail).

Trained WWC staff use the keywords defined in the review protocol to search a large set of electronic databases and organizational websites (see Appendix B). Full citations and, where available, abstracts and full texts for studies identified through these searches are catalogued for subsequent eligibility screening. In addition, the WWC contacts intervention developers and distributors to identify other research.

All citations gathered through the search process undergo a preliminary screening to determine whether the study meets the criteria established in the review protocol. This screening process is described in Chapter IV.

The WWC also requires review teams to identify studies that have been previously reviewed by the WWC, perhaps for another product or under a previous version of the standards. Those studies will be re-reviewed using the eligibility criteria of the specific protocol used and up-to-date standards, as necessary.

IV. SCREENING STUDIES

Studies gathered during the literature search are screened against the parameters specified in the review protocol in order to identify a set of studies eligible for WWC review. The initial screening for eligibility is conducted by a WWC staff member who has been certified as a screener. Studies may be designated as *Ineligible for WWC Review* for any of the following reasons:

- *The study does not use an eligible design.* An eligible design is one for which the WWC has pilot or final design standards, and that uses primary analysis to examine the effectiveness of an intervention.
    - *Eligible designs*. The WWC includes findings from randomized controlled trials (RCTs), quasi-experimental designs (QEDs), and regression discontinuity designs (RDDs). The WWC also has pilot standards for single-case designs (SCDs), which may also be reviewed and described in reports if specified in the review protocol. Studies using other study designs are not eligible for review.
    - *Primary analysis of the effectiveness of an intervention*. Additionally, some studies are not primary studies of an intervention's impacts or effectiveness. For example, studies of how well an intervention was implemented, literature reviews, or meta-analyses are not eligible to be included in a WWC review (but may still be used as additional sources for the review of an eligible study).
- *The study does not use a sample aligned with the protocol.* Characteristics of study samples that are eligible for review will be listed in the protocol and may include age, grade range, gender, or English learner status.
- *The study is outside the scope of the protocol.* Each protocol identifies the characteristics of studies that are eligible for review, including outcome measures, time frame for publication, setting of the study, and types of interventions.
    - *Outcome measures*. Studies eligible for review must include at least one outcome that falls within the domains identified in the review protocol.
    - *Time frame for publication*. When the WWC begins the review of studies for a new topic, a cutoff date is established for research to be included. Typically, this cutoff is set at 20 years prior to the start of the WWC review of the topic. This time frame generally encompasses research that adequately represents the current status of the field and avoids inclusion of research conducted with populations and in contexts that may be very different from those existing today.
    - *Study setting*. Review protocols might limit eligible studies to those that take place in certain geographic areas (such as in the United States), or in certain types of schools or classrooms.
    - *Interventions*. Review protocols describe the interventions that are eligible for review and any requirements for eligibility, such as replicability (i.e., those that can be reproduced). In addition to meeting the specific requirements in the review protocol, to be eligible, the intervention must be aligned with the focus of the review. For example, each intervention report is focused on a specific intervention, and only

studies that examine the effectiveness of that intervention are eligible for review. Additionally, an intervention report is focused on a single intervention, whereas a practice guide may focus on a wider range of interventions. When developing the review protocol, it is not possible to anticipate the specific focus of all future review efforts that it could guide, so all relevant eligibility criteria cannot be specified in the review protocol. In particular, if the focus of a review is a specific intervention, but the intervention is always offered in combination with a second intervention, the study is ineligible for review. However, if the focus of the review is not specific to one of the two interventions, and both interventions are individually eligible for review under the same review protocol, the WWC will view the combination as a single, eligible intervention.

V. REVIEWING STUDIES

A. Definition of a Study

The core of the systematic review process is the assessment of eligible studies against WWC design standards. The definition of a study is important, given how the WWC reports on and summarizes evidence. Both the level of evidence in practice guides and the summary of findings in an intervention report depend on the number of studies that meet WWC design standards. For example, a rating of *positive effects* requires at least two studies that meet WWC design standards.

A study is not necessarily equivalent to a manuscript, such as a journal article, book chapter, or report. A single study can be described in multiple manuscripts (e.g., a 5-year study of an intervention may release interim annual reports). Alternatively, a manuscript can include multiple studies (e.g., many articles include several separate experiments). In the case of multiple manuscripts that report on one study, the WWC selects the earliest manuscript with relevant findings as the lead citation used throughout the product and lists other manuscripts that describe the study as related sources.

The critical issue in defining a study as distinct from a related analysis is whether it provides a separate test of the intervention. That is, does it provide new evidence that is distinct from existing evidence? When analyses of the same intervention share certain characteristics, there may be a concern that they do not provide independent tests of the intervention.

Frequently, the question of whether there is more than one study arises from the separate presentation of findings that share one or more characteristics. When two findings share certain characteristics, the WWC may consider them parts of the same study. These characteristics include:

- **Sample members, such as teachers or students.** Findings from analyses that include some or all of the same teachers or students may be related.
- **Group formation procedures, such as the methods used to conduct random assignment or matching.** When authors use identical (or nearly identical) methods to form the groups used in multiple analyses, or a single procedure was used to form the groups, the results may not provide independent tests of the intervention.
- **Data collection and analysis procedures.** Similar to group formation, when authors use identical (or nearly identical) procedures to collect and analyze data, the findings may be related. Sharing data collection and analysis procedures means collecting the same measures from the same data sources, preparing the data for analysis using the same rules, and using the same analytic methods with the same control variables.
- **Research team.** When manuscripts share one or more authors, the reported findings in those manuscripts may be related.

The WWC considers findings on the effectiveness of the same intervention to be a single study if they share at least three of these four characteristics in total (see Appendix D for examples). In particular, when two findings meet this condition, they demonstrate:

1. *Similarity or continuity in the intervention and comparison groups used to produce the findings.* They either share sample members or use the same group formation procedures, and
2. *Similarity or continuity in the procedures used to produce the findings.* They either share the same data collection and analysis procedures or share research team members.

When is it unclear whether findings meet the criteria described above, the review team leadership (lead methodologist and content expert, described further in Appendix C) has the discretion to determine what constitutes a single study or multiple studies, and the decision is clearly noted in the WWC product that includes the review.

B. The WWC Study Review Process

In 2011, after an evaluation of the process of using two reviewers to review every study (see *Handbook,* version 2.1, p. 11), the WWC implemented a streamlined review process for randomized controlled trials and quasi-experimental studies. This section describes the steps of the study review process. After eligible studies are identified through the comprehensive literature search (described in Appendix B), all studies adhere to the following review process (also depicted in Figure V.1).

Each study receives a **first review**, documented in a study review guide (SRG). The SRG and instructions can be accessed at http://ies.ed.gov/ncee/wwc/StudyReviewGuide.aspx.

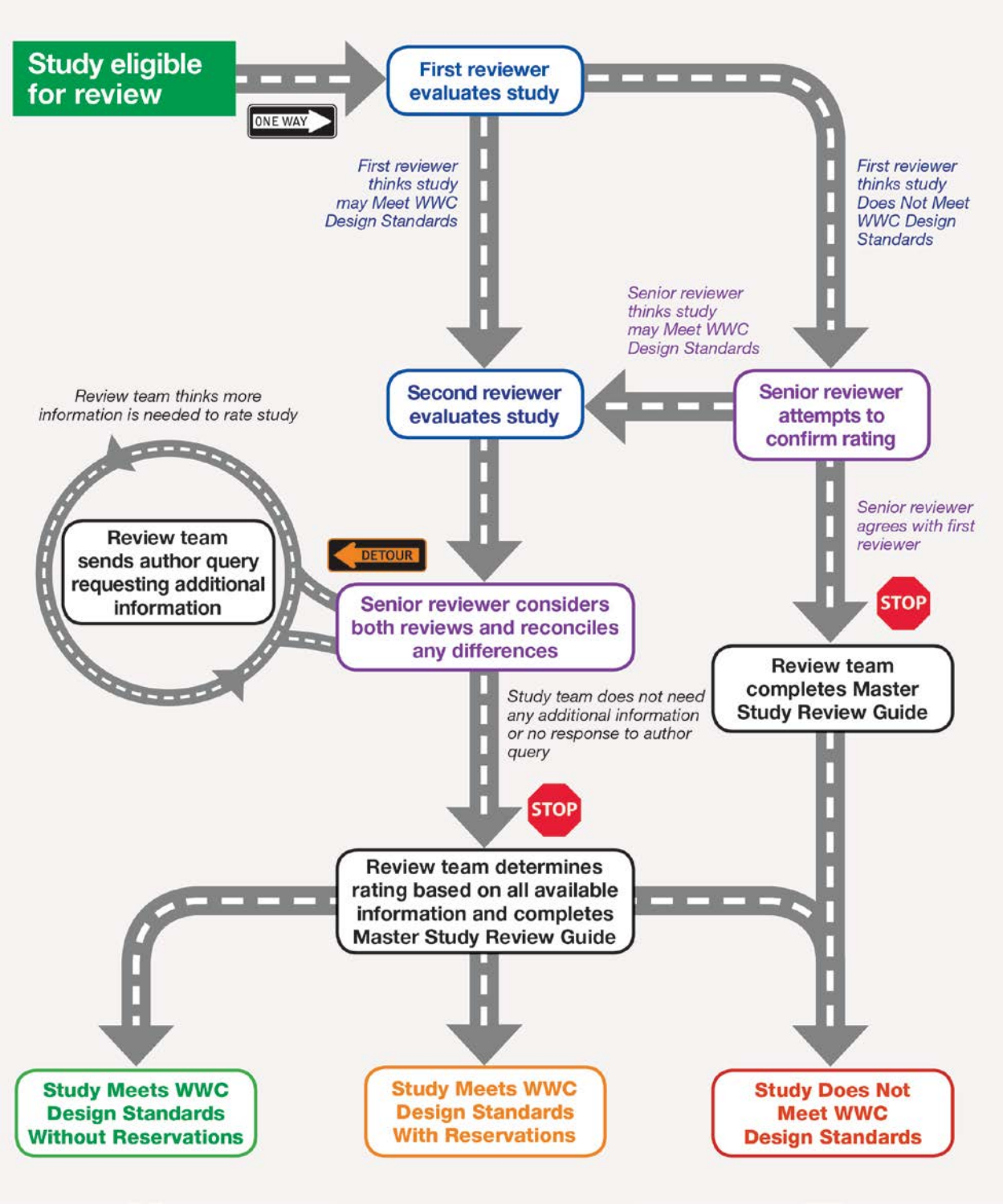
- If the first reviewer determines that the study does not meet WWC standards, a senior reviewer for the topic area team examines the study and determines whether the reason for not meeting standards indicated by the first reviewer is correct.
  - If the senior reviewer *agrees* with the first reviewer's assessment, the master SRG is created and completed.
  - If the senior reviewer *disagrees*, the study receives a second review.
- If the first reviewer determines that the study meets WWC standards, or could meet standards with more data provided by the study author, the study receives a second review.

If a study receives a **second review**, it is conducted without knowledge of the previous review or rating so that it cannot be influenced by previous findings. After the second review is complete, the coordinator asks the second reviewer to compare his or her assessment with that of the first reviewer (or senior reviewer, in the event that he or she did not verify the first reviewer's assessment).

- If the second reviewer and first (or senior) reviewer agree on their assessment of the study rating and the key components of the review, then a master SRG is created. Key components include the level of attrition, establishment of equivalence, which measures to include, and effect sizes. Minor discrepancies, such as those involving sample sizes, can be resolved without involvement of the topic area team leadership.
- If the reviewers disagree on the final study rating, the reason for the rating, or other key components of the review, discrepancies or uncertainties are brought to the lead methodologist for the team for resolution before a master SRG is created.

When necessary, the WWC contacts study authors to obtain information to complete the master SRG. This **author query** may ask for information related to sample characteristics, sample sizes, baseline statistics; outcome statistics; or other information on group formation, confounding factors, and outcome measures. The WWC may also ask for information about analyses referenced in the article but not presented, including summary data such as means, standard deviations, and simple correlations between measures, but the WWC does not ask for new analyses to be conducted. All information received through an author query that is used in a report is made available to the public and is documented in the final report.

Figure V.1. Roadmap of the study review process for group design studies



VI. REPORTING ON FINDINGS

To the extent possible, the WWC reports the magnitude and statistical significance of study-reported estimates of the effectiveness of interventions, using common metrics and applying corrections (e.g., clustering and multiple comparisons) that may affect the study-reported results. Next, a heuristic is applied to characterize study findings in a way that incorporates the direction, magnitude, and statistical precision of the impact estimates. Finally, in some of its products (e.g., intervention reports and practice guides), the WWC combines findings from individual studies into summary measures of effectiveness, including aggregate numerical estimates of the size of impacts, overall ratings of effectiveness, and a rating for the extent of evidence.

A. Finding from an Individual Analysis

The WWC defines an individual finding as the measured effect of the intervention relative to a specific comparison condition on an outcome for a sample at a certain point in time relative to the introduction of the intervention.

1. Magnitude of a Finding

The WWC reports the magnitude of study findings in two ways: (a) effect sizes (i.e., standardized mean differences) and (b) a WWC-calculated "improvement index."

Effect Sizes

For all studies, the WWC records the study findings in the units reported by the study authors. In addition, the WWC computes and records the **effect size** associated with study findings on relevant outcome measures. In general, to improve the comparability of effect size estimates across studies, the WWC uses student-level standard deviations when computing effect sizes, regardless of the unit of assignment or the unit of intervention. For effect size measures used in other situations, such as those based on student-level *t*-tests or cluster-level assignment, see Appendix E.

For **continuous outcomes**, the WWC has adopted the most commonly used effect size index, the standardized mean difference known as Hedges' *g*, with an adjustment for small samples. It is defined as the difference between the mean outcome for the intervention group and the mean outcome for the comparison group, divided by the pooled within-group standard deviation of the outcome measure. Defining y_i and y_c as the means of the outcome for students in the intervention and comparison groups, n_i and n_c as the student sample sizes, s_i and s_c as the student-level standard deviations, and ω as the small sample size correction (see Appendix E), the effect size is given by

$$g = \frac{\omega\left(y_i - y_c\right)}{\sqrt{\frac{\left(n_i - 1\right)s_i^2 + \left(n_c - 1\right)s_c^2}{n_i + n_c - 2}}}.$$

For **dichotomous outcomes,** the difference in group means is calculated as the difference in the probability of the occurrence of an event. The effect size measure of choice for dichotomous outcomes is the Cox index, which yields effect size values similar to the values of Hedges' *g* that

one would obtain if group means, standard deviations, and sample sizes were available, assuming the dichotomous outcome measure is based on an underlying normal distribution. Defining p_i and p_c as the probability of an outcome for students in the intervention and comparison groups, the effect size is given by

$$d_{Cox} = \omega \left[ln\left(\frac{p_i}{1-p_i} \right) - ln\left(\frac{p_c}{1-p_c} \right) \right] / 1.65 .$$

The WWC also follows these additional guidelines when calculating effect sizes:

- If a study reports both unadjusted and adjusted post-intervention means, the WWC reports the adjusted means and unadjusted standard deviations and uses these in computing effect sizes.
- When only unadjusted group means are reported, and information about the correlation between the tests is not available, the WWC computes the effect size of the difference between the two groups on the pretest and the effect size of the difference between the two groups on the posttest separately, with the final effect size given by their difference. The WWC considers this *post hoc* adjustment an acceptable statistical adjustment for baseline differences if the pretest and posttest are sufficiently related based on the requirements described in Chapter II.A of the *Standards Handbook*.
- When the WWC makes a difference-in-differences adjustment to findings provided by the study author, the WWC reports statistical significance levels for the adjusted differences that reflect the adjustment in the effect size. For example, consider a pre-intervention difference of 0.2 on an achievement test. If the post-intervention difference were 0.3, the difference-in-differences adjusted effect would be 0.1. Subsequently, the statistical significance reported by the WWC would be based on the adjusted finding of 0.1, rather than the unadjusted finding of 0.3.
- When the author-reported and WWC-calculated effect sizes differ, the WWC attempts to identify the source of the difference and explains the reason for the discrepancy in a table note. In general, when this occurs, the WWC will report the WWC-calculated effect size because its computation can be verified and using the WWC-calculated measures supports comparability across outcomes and studies. However, the WWC will report an author-reported effect size that is comparable to Hedges' *g* if it adjusts for baseline differences and the WWC-calculated effect size does not or is based on the *post hoc* adjustment described above. For the WWC, effect sizes of 0.25 standard deviations or larger are considered to be **substantively important**. Effect sizes at least this large are interpreted as a qualified positive (or negative) effect, even though they may not reach statistical significance in a given study.

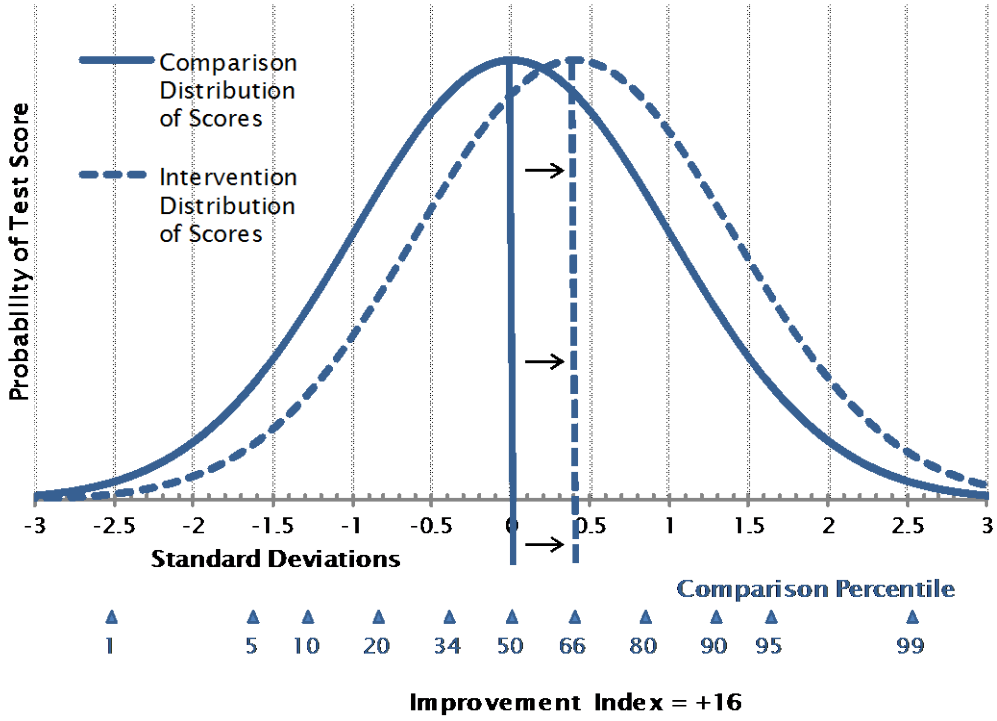
Improvement Index

In order to help readers judge the practical importance of an intervention's effect, the WWC translates effect sizes into "improvement index" values. The improvement index for an individual study finding represents the difference between the percentile rank corresponding to

the mean value of the outcome for the intervention group and the percentile rank corresponding to the mean value of the outcome for the comparison group distribution (details on the computation of the improvement index are presented in Appendix E). 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.

Figure IV.1 illustrates the interpretation of the improvement index. In this example, the estimated average impact of the intervention is an improvement of 0.4 standard deviations in reading test scores. Thus, on average, a student in the comparison group who scores at the 50th percentile for the study sample would be expected to have scored 0.4 standard deviations above the mean if he or she had received the intervention, or at the 66th percentile of students. The resulting improvement index is +16, corresponding to moving performance for the average student from the 50th to the 66th percentile of the comparison group distribution. For more details, see Appendix E.

Figure IV.1. Computation of the WWC Improvement Index



2. Statistical Significance of a Finding

To adequately assess the effects of an intervention, it is important to know the statistical significance of the estimates of the effects in addition to the mean difference, effect size, or improvement index, as described above. For the WWC, a **statistically significant** estimate of an effect is one for which the probability of observing an effect that is at least as large as the measured effect under the view that the intervention had no impact is less than one in 20 (using a two-tailed *t*-test with $p = 0.05$), assuming there is a single measure or mean effect within each domain.

The WWC generally accepts the statistical significance levels reported by the author(s) of the study. However, the WWC will compute the statistical significance levels if the study does not include statistical significance estimates. The WWC calculates statistical significance for findings based on continuous outcome measures by computing the *t*-statistic:

$$t = g \Big/ \sqrt{\frac{n_i + n_c}{n_i n_c} + \frac{g^2}{2\left(n_i + n_c\right)}},$$

where g is the effect size, and n_i and n_c are the average sample sizes for the intervention and comparison groups, respectively, for a set of findings. The second term under the radical applies a correction for small sample sizes (Hedges & Olkin, 1985).

Additionally, the *t*-statistic for findings based on dichotomous outcome measures is calculated using:

$$t = 1.65\, d_{Cox} \Big/ \left(\sqrt{\frac{1}{p_i n_i} + \frac{1}{(1-p_i) n_i} + \frac{1}{p_c n_c} + \frac{1}{(1-p_c) n_c}} \right),$$

where d_{Cox} is the effect size based on the Cox index, and p_i and p_c are the probabilities of a positive outcome for students in the intervention and comparison groups, respectively (Sanchez-Meca, Marin-Martinez, & Chacon-Moscoso, 2003).

Also, the WWC will make adjustments to statistical significance levels reported in the study if they do not account for clustering when there is a mismatch between the unit of assignment and unit of analysis. These WWC-calculated or recalculated estimates appear in WWC products with a note describing the source of the calculations and noting any difference between WWC and author-reported findings.

Clustering Correction for "Mismatched" Analyses

A "mismatch" problem occurs when assignment is carried out at the cluster level (e.g., classroom or school level), while the analysis is conducted at the individual level (e.g., student level). When the analysis ignores the correlation between outcomes among individuals within the same clusters when computing the standard errors of the impact estimates, this approach leads to underestimated standard errors and overestimated statistical significance. The point estimates of the intervention's effects (including the effect size and improvement index) obtained from mismatched analyses are not affected by this feature of the study sample.

To assess the statistical significance of an intervention's effects in cases where study authors have not corrected for the clustering, the WWC computes clustering-corrected statistical significance estimates based on guidance in Hedges (2007). The basic approach to the clustering correction is first to compute the *t*-statistic corresponding to the effect size 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 an estimate of the intra-class correlation coefficient (ICC). As defaults, the WWC uses the ICC values of 0.20 for achievement outcomes and 0.10 for all other outcomes, but will use study-reported ICC values when available. If a deviation from these defaults is warranted, it will be stated in the review protocol. The statistical significance estimate corrected for clustering is then obtained from the *t*-

distribution using the corrected *t*-statistic and degrees of freedom. Each step of the process is specified in Appendix F.

B. Findings from Multiple Analyses

Studies often present several findings obtained from analyses that vary the comparison condition, outcome measure, sample, or point in time. For example, analyses may include all participants in the study or subsets of the population. Similarly, analyses may include multiple outcome measures in the same domain, a single outcome measured at multiple points in time, and a composite measure and its components.

For a study with multiple analyses, all eligible analyses as defined by the protocol are reviewed (as described in Chapter II, protocols define eligible interventions, populations, and outcomes). Author queries are conducted as needed to evaluate all eligible analyses presented in the study. The study rating is specified as the highest rating obtained across all eligible analyses.

1. Presentation of Findings from Multiple Analyses

The WWC reports findings from all eligible analyses (as defined in the applicable review protocol) that meet standards, split into main and supplemental findings. The study's characterization of findings (described in Section C) is based on the main findings. For each outcome measure, and among those findings that meet WWC design standards, the WWC uses the following criteria to designate one finding or set of findings as the main finding: (1) includes the full sample; (2) uses the most aggregate measure of the outcome measure (rather than individual subscales); and (3) is measured at a time specified by the protocol (e.g., latest follow-up period, earliest follow-up period after conclusion of the intervention, or after 1 year of exposure).

The following rules guide the distinction of findings from eligible analyses that meet WWC design standards, as illustrated by an example of a study with two cohorts of eighth-grade students. In this example, the study includes eligible analyses both of a pooled sample of students and of each cohort analyzed separately. Which analyses will be presented as main findings and which will be presented as supplemental findings depends on which analyses meet WWC design standards.

- **All eligible analyses meet standards.** The pooled analysis is presented as a main finding, while the other analyses (separate cohorts) are presented as supplemental findings.
- **The pooled analysis meets standards, and one of the cohort-specific analyses meets standards.** The pooled analysis is presented as a main finding, while only the other analysis that meets standards (one of the two separate cohorts) is presented as a supplemental finding.
- **The pooled analysis meets standards, and neither of the cohort-specific analyses meet standards.** The pooled analysis is presented as a main finding, with no supplemental findings.
- **The pooled analysis does not meet standards, but both of the cohort-specific analyses meet standards.** Because the cohort-specific analyses each separately meet

standards and in combination cover the entire sample, the WWC creates a pooled sample from the cohorts as the main finding using the formulas provided below in Section 2. The findings from the analyses for separate cohorts are presented as supplemental findings. However, if the only findings meeting standards in this example were instead findings for separate subscales of a composite measure, both based on the entire sample, the WWC would report the findings for each subscale separately as main findings (along with the unweighted domain average that aggregates the findings, also described in Section 2).

- **The pooled analysis does not meet standards, and only one of the cohort-specific analyses meets standards.** Because there is no set of analyses meeting standards that cover the entire sample, the cohort-specific analysis that meets standards is presented as a main finding, with no supplemental findings. However, reviewers should also assess whether the WWC-calculated finding based on pooling across both cohorts can meet WWC design standards, and report this pooled finding as the main finding if it does.[1] If the only finding meeting standards in this example were instead for a separate subscale of a composite measure, the WWC would report the finding for the subscale that meets WWC design standards as the main finding.

These rules allow the WWC to characterize a study's findings based on the most comprehensive information available. However, not all studies will report a single finding or set of findings that meets the criteria described above (comprehensive sample, aggregate outcome measure, and preferred time period) that the WWC can designate as the main finding. When applying these rules is not straightforward because of incomplete information about findings, overlapping samples, or other complications, the review team leadership has discretion for a study or group of studies under review to identify main and supplemental findings (among those that meet WWC design standards) in a way that best balances the goals of characterizing each study's findings based on the criteria above and presenting the findings in a clear and straightforward manner, while avoiding overlap in the samples and subscales in the main findings. Additionally, when a study reports multiple findings for the same outcome measure by comparing the intervention group with multiple comparison groups, the review team leadership has discretion to choose one as the main finding from the study, or to create a pooled comparison group from multiple groups. See Appendix F for detail about assessing statistical significance in reviews of studies with multiple comparison groups.

Some WWC review efforts are designated as *expedited* by IES. For expedited reviews, the review team leadership also has discretion to focus each study review on eligible findings only from the full sample (rather than on subgroups), only on composite measures (rather than subscales), only on the most relevant time period as specified in the applicable review protocol, and only using the most relevant comparison group. Other eligible findings within a study may not be reviewed when the WWC conducts an expedited review.

The WWC makes an exception to these rules when an author reports a set of sensitivity analyses that focus on the same or very similar samples, but apply different analytic methods to

[1] It is possible for the WWC-calculated finding to meet WWC design standards even when the author-reported findings from the pooled analysis and one of the cohorts do not. For example, the author-reported analysis might include an endogenous covariate, while the findings used to form the WWC-calculated pooled finding do not adjust for the endogenous covariate. Also, the WWC-calculated pooled finding might have low attrition, while only one of the author reported cohort-specific findings has low attrition.

obtain each finding. In these cases, among those that meet WWC design standards, the WWC designates as the main finding the one that receives the highest rating (if the rating differs across findings), accounts for the baseline measure(s) specified in the review protocol (if any of the sensitivity analyses do this), uses the most comprehensive sample, and is most robust to threats to internal validity based on the judgment of the authors (as reported in the study) or the review team leadership. The topic area leadership have discretion to select a finding when these specifications do not distinguish a single finding. The remaining sensitivity analyses are not reported as supplemental findings, but instead are noted in the WWC product that includes the review.

See Appendix G for procedures for reporting findings from studies that report findings from both intent-to-treat (ITT) and complier average causal effects (CACE) analyses.

Finally, the WWC adjusts for multiple comparisons (described below) among all main findings, but not supplemental findings.

2. Magnitude of Findings

The WWC combines findings in three situations: across subsamples for a single outcome measure within a study, across outcome measures within a study, and across studies.

Some studies present findings separately for several subsamples of subjects without presenting an aggregate result (or the aggregate result may not meet WWC design standards). Examples include a middle school math study that presents the effects separately for sixth-, seventh-, and eighth-grade students; an adolescent literacy study that examines high- and low-risk students; and a beginning reading study that considers low-, medium-, and high-proficiency students. When the study presents findings separately for portions of the sample without presenting a full sample result, the WWC queries authors to see if they conducted an analysis on the full sample (although, the review team has discretion to skip this query when there is reason to believe a full sample analysis was not conducted). The study's analysis is preferred, as it may be more precise than the WWC's computation. If the WWC is unable to obtain aggregate results from the author, or the aggregate result does not meet WWC design standards, the WWC averages **across subsamples for a single outcome measure within a study.**

More concretely, if a study provides findings for G mutually exclusive subsamples that make up the entire sample, but no overall finding, the WWC computes a sample-weighted average of the separate impacts. For continuous outcomes, defining n_g, m_g, and s_g as the size, impact, and standard deviation for subsample g, respectively, the average estimate of the impact (M) across all subsamples and the standard deviation (S) for the average estimate of the impact are given by

$$M = \frac{\sum_{g=1}^{G} n_g m_g}{\sum_{g=1}^{G} n_g} \quad \text{and} \quad S = \sqrt{\frac{\sum_{g=1}^{G}\left[\left(n_g - 1\right)s_g^2 + n_g\left(M - m_g\right)^2\right]}{\sum_{g=1}^{G} n_g - 1}}.$$

The effect size g is then given by $\omega M / S$.

For dichotomous outcomes, defining p_{gi} and p_{gc} as the probabilities of the occurrence of a positive outcome for the intervention and the comparison groups for subsample g, respectively, the WWC first calculates the average probabilities across subsamples P_i and P_c using:

$$P_i = \frac{\sum_{g=1}^{G} n_{gi} p_{gi}}{\sum_{g=1}^{G} n_{gi}} \quad \text{and} \quad P_c = \frac{\sum_{g=1}^{G} n_{gc} p_{gc}}{\sum_{g=1}^{G} n_{gc}}.$$

Then, the effect size is given by the Cox index using P_i and P_c (see Appendix E).

If a study reports findings that meet WWC design standards for more than one outcome measure in a domain, the effect sizes for all of that study's outcomes are combined into a **study average effect size** using the simple, unweighted average of the individual effect sizes. The **study average improvement index** is computed directly from the study average effect size.

For systematic reviews that include more than one study, if more than one study has outcomes in a domain, the study average effect sizes for all of those studies are combined into a **domain average effect size** using the simple, unweighted average of the study average effect sizes. The **domain average improvement index** is computed directly from the domain average effect size.

3. Statistical Significance of Findings

As a second component in summarizing findings from multiple analyses, the WWC assesses statistical significance using the same *t*-statistic formulas given in Section A. For study average effect sizes based on continuous outcome measures, g is the average effect size across findings, and n_i and n_c are the average sample sizes for the intervention and comparison groups, respectively, for a set of findings. For study average effect sizes based on dichotomous outcome measures, d_{Cox} is the average effect size based on the Cox index across findings, and p_i and p_c are the average probabilities of a positive outcome for students in the intervention and comparison groups, respectively.

For WWC-aggregated findings for the sample outcome measure across subsamples, the *t*-statistic is given by one of the following formulas for continuous and dichotomous outcome measures, respectively:

$$t = g \Big/ \sqrt{\frac{N_i + N_c}{N_i N_c} + \frac{g^2}{2(N_i + N_c)}}, \quad \text{or}$$

$$t = 1.65\, d_{Cox} \Big/ \left(\sqrt{\frac{1}{P_i N_i} + \frac{1}{(1-P_i)N_i} + \frac{1}{P_c N_c} + \frac{1}{(1-P_c)N_c}} \right),$$

where g is the effect size based on M and S as defined above, d_{Cox} is the effect size based on the Cox index using P_i and P_c as defined above, and N_i and N_c are the total sample sizes across the subsamples for the intervention and comparison groups, respectively.

Benjamini-Hochberg Correction for Multiple Comparisons

The WWC has adopted the Benjamini-Hochberg (BH) correction to account for multiple comparisons or "multiplicity," which can lead to inflated estimates of the statistical significance of findings (Benjamini & Hochberg, 1995). Repeated tests of highly correlated outcomes will lead to a greater likelihood of mistakenly concluding that the differences in means for outcomes of interest between the intervention and comparison groups are significantly different from zero (called Type I error in hypothesis testing). Thus, the WWC uses the BH correction to reduce the possibility of making this type of error.

If the exact *p*-values are not available but effect sizes are available, the WWC converts the effect size to *t*-statistics and then obtains the corresponding *p*-values. For findings based on analyses in which the unit of analysis was aligned with the unit of assignment, or where study authors conducted their analysis in such a way that their *p*-values were adjusted to account for the mismatch between the level of assignment and analysis, the *p*-values reported by the study authors are used for the BH correction. For findings based on mismatched analyses that have not generated *p*-values that account for the sample clustering, the WWC uses the clustering-corrected *p*-values for the BH correction. For more detail, see Appendix F.

C. Qualitative Summaries of Findings

WWC products, including practice guides and intervention reports, provide qualitative summaries of evidence from individual studies and across multiple studies in systematic reviews. These qualitative summaries indicate the direction (e.g., positive, negative, or indeterminate) and strength (e.g., magnitude and statistical significance) of findings. The summaries are based on findings that meet WWC design standards and are designated by the WWC as the main findings in the study.

1. Summary of Evidence for an Individual Study

Using the estimated effect size and statistical significance level (accounting for clustering and multiple comparisons when necessary), the WWC characterizes study findings within each outcome domain in one of five categories: (a) statistically significant positive (favorable) effect, (b) substantively important positive effect, (c) indeterminate effect, (d) substantively important negative (unfavorable) effect, and (e) statistically significant negative effect. For findings based on a single outcome measure, the rules in Table IV.1 are used to determine which of the five categories applies.

Table IV.1. WWC Characterization of Findings of an Effect Based on a *Single Outcome Measure* Within a Domain

Statistically significant positive effect	The estimated effect is positive and statistically significant (correcting for clustering when not properly aligned).
Substantively important positive effect	The estimated effect is positive and not statistically significant but is substantively important.
Indeterminate effect	The estimated effect is neither statistically significant nor substantively important.
Substantively important negative effect	The estimated effect is negative and not statistically significant but is substantively important.

Statistically significant negative effect	The estimated effect is negative and statistically significant (correcting for clustering when not properly aligned).

Note: A statistically significant estimate of an effect is one for which the probability of observing an effect that is at least as large as the measured effect under the view that the intervention had no impact is less than one in 20 (using a two-tailed *t*-test with $p = 0.05$). A properly aligned analysis is one for which the unit of assignment and unit of analysis are the same, or that accounts for the correlation between outcomes among individuals within the same clusters. An effect size of 0.25 standard deviations or larger is considered to be substantively important.

If the effect is based on multiple outcome measures within a domain, the rules in Table IV.2 apply.

Table IV.2. WWC Characterization of Findings of an Effect Based on *Multiple Outcome Measures* Within a Domain

Statistically significant positive effect	When any of the following is true: 1. At least one finding is positive and statistically significant, and none are negative and statistically significant based on univariate statistical tests, accounting for multiple comparisons (and correcting for clustering when not properly aligned). 2. The WWC-calculated study average effect size for the multiple outcome measures is positive and statistically significant (correcting for clustering when not properly aligned). 3. The study reports that the omnibus effect for all outcome measures together is positive and statistically significant on the basis of a multivariate statistical test in a properly aligned analysis.
Substantively important positive effect	The WWC-calculated study average effect size is positive and not statistically significant but is substantively important.
Indeterminate effect	The WWC-calculated study average effect size is neither statistically significant nor substantively important.
Substantively important negative effect	The WWC-calculated study average effect size is negative and not statistically significant but is substantively important.
Statistically significant negative effect	When any of the following is true: 1. At least one finding is negative and statistically significant, and none are positive and statistically significant based on univariate statistical tests, accounting for multiple comparisons (and correcting for clustering when not properly aligned). 2. The WWC-calculated study average effect size for the multiple outcome measures is negative and statistically significant (correcting for clustering when not properly aligned). 3. The study reports that the omnibus effect for all outcome measures together is negative and statistically significant on the basis of a multivariate statistical test in a properly aligned analysis.

Note: A statistically significant estimate of an effect is one for which the probability of observing such a result an effect that is at least as large as the measured effect under the view that the intervention had no impact is less than one in 20 (using a two-tailed *t*-test with $p = 0.05$). A properly aligned analysis is one for which the unit of assignment and unit of analysis are the same, or that accounts for the correlation between outcomes among individuals within the same clusters. An effect size of 0.25 standard deviations or larger is considered to be substantively important.

Because they are not directly comparable to individual-level (e.g., student-level) effect sizes, results based on the analysis of cluster-level data, such as school level outcomes, cannot be considered in determining substantively important effects in intervention ratings. (However, cluster-level means can be used to calculate effect sizes that are comparable to student-level effect sizes, so long as the calculation uses a standard deviation based on individual-level data.) Therefore, in intervention reports, **cluster-level effect sizes** are excluded from the computation of domain average effect sizes and improvement indices. However, the statistical significance and direction (positive or negative) of cluster-level findings is taken into account in determining the characterization of study findings.

In addition to characterizing study findings as described above, in Find What Works on the WWC website, the WWC also identifies studies that include one or more findings that are statistically significant and positive. This designation, reported on the web page for a study, is based on examining all findings in the study that meet WWC design standards, including supplementary findings that did not contribute to an intervention rating in an intervention report. A study receives this designation if it includes at least one main or supplemental finding that meets WWC design standards and is positive and statistically significant after WWC adjustments for clustering and multiple comparisons, if necessary. The determination is assessed separately for each review of the study (for example, some findings may be eligible for review under one protocol, but not under another, which could affect how the multiple comparisons adjustment is applied).

Identifying studies that have a statistically significant and positive finding points decision makers to studies that provide some evidence of an intervention's effectiveness. However, this determination is not based on a systematic review of the evidence, so other studies, or even other findings within the same study, could provide contradictory evidence. In contrast, the WWC's characterization of study findings described above provides a summary across all of the evidence for an intervention within a study that met standards, and the characterization of findings across studies described below provides a summary across all of the evidence for an intervention that met standards from a systematic review.

2. Summary of Evidence for an Intervention Report

In intervention reports, the WWC provides a rating for the intervention's effects within each outcome domain, and characterizes the extent of evidence for that rating.

Intervention Rating

The WWC combines findings of effectiveness across multiple studies of an intervention to determine an intervention rating. The WWC uses a set of guidelines to determine the rating for an intervention just as it uses guidelines to determine the characterization of findings for an individual study (Table IV.3). The criteria in Table IV.3 are not mutually exclusive. If multiple criteria are met, the intervention is assigned the highest intervention rating for which it is eligible.

Table IV.3. Criteria Used to Determine the WWC Rating of Effectiveness for an Intervention

Positive effects: Strong evidence of a positive effect with no overriding contrary evidence.	• Two or more studies show statistically significant positive effects, at least one of which meets WWC group design standards without reservations, AND • No studies show statistically significant or substantively important negative effects.
Potentially positive effects: Evidence of a positive effect with no overriding contrary evidence.	• At least one study shows statistically significant or substantively important positive effects, AND • Fewer or the same number of studies show indeterminate effects than show statistically significant or substantively important positive effects, AND • No studies show statistically significant or substantively important negative effects.
No discernible effects: No affirmative evidence of effects.	• None of the studies show statistically significant or substantively important effects, either positive or negative.
Mixed effects: Evidence of inconsistent effects.	EITHER both of the following: • At least one study shows statistically significant or substantively important positive effects, AND • At least one study shows statistically significant or substantively important negative effects, BUT no more such studies than the number showing statistically significant or substantively important positive effects. OR both of the following: • At least one study shows statistically significant or substantively important effects, AND • More studies show an indeterminate effect than show statistically significant or substantively important effects.
Potentially negative effects: Evidence of a negative effect with no overriding contrary evidence.	EITHER both of the following: • One study shows statistically significant or substantively important negative effects, AND • No studies show statistically significant or substantively important positive effects. OR both of the following: • Two or more studies show statistically significant or substantively important negative effects, at least one study shows statistically significant or substantively important positive effects, AND • More studies show statistically significant or substantively important negative effects than show statistically significant or substantively important positive effects.
Negative effects: Strong evidence of a negative effect with no overriding contrary evidence.	• Two or more studies show statistically significant negative effects, at least one of which meets WWC group design standards without reservations, AND • No studies show statistically significant or substantively important positive effects.

Note: A statistically significant estimate of an effect is one for which the probability of observing such a result an effect that is at least as large as the measured effect under the view that the intervention had no impact is less than one in 20 (using a two-tailed *t*-test with $p = 0.05$). An effect size of 0.25 standard

deviations or larger is considered to be substantively important. An indeterminate effect is one for which the single or mean effect is neither statistically significant nor substantively important.

Extent of Evidence Characterization

The final step in combining findings of effectiveness across multiple studies of an intervention is to report on the extent of the evidence used to determine the intervention rating. The extent of evidence categorization was developed to inform readers about how much evidence was used to determine the intervention rating, using the number and sizes of studies. This scheme has two categories: (a) medium to large and (b) small (Table IV.4).

Table IV.4. Criteria Used to Determine the WWC Extent of Evidence for an Intervention

Medium to large	• The domain includes more than one study, AND • The domain includes more than one setting, AND • The domain findings are based on a total sample of at least 350 students, OR, assuming 25 students in a class, a total of at least 14 classrooms across studies.
Small	• The domain includes only one study, OR • The domain includes only one setting, OR • The domain findings are based on a total sample size of fewer than 350 students, AND, assuming 25 students in a class, a total of fewer than 14 classrooms across studies.

With only one study, the possibility exists that some characteristics of the study—for example, the outcome measures or the timing of the intervention—might have affected the findings. Multiple studies reduce potential bias due to sampling error. Therefore, the WWC considers the extent of evidence to be small when the findings are based on only one study.

Similarly, with only one setting (e.g., school), the possibility exists that some characteristics of the setting—for example, the principal or student demographics within a school—might have affected the findings or were intertwined or confounded with the findings. Therefore, the WWC considers the extent of evidence to be small when the findings are based on only a single setting.

The sample size of 350 was selected because it is generally the smallest sample size needed to have adequate statistical power (e.g., 80% probability of rejecting the null hypothesis when it is false and no more than a 5% probability of mistakenly concluding there is an impact) to detect impacts that are meaningful in size (e.g., 0.3 standard deviations or larger) for a simple RCT (e.g., students are randomized to the intervention or comparison conditions in equal proportions) with no covariates used in the analysis.

3. Summary of Evidence for a Practice Guide

In combining the evidence for each recommendation, the expert panel and WWC review staff consider the following:

- The number of studies
- The quality of the studies
- Whether the studies represent the range of participants, settings, and comparisons on which the recommendation is focused

- Whether findings from the studies can be attributed to the recommended practice
- Whether findings in the studies are consistently positive

Practice guide panels rely on a set of definitions to determine the level of evidence supporting their recommendations (Table IV.5).

Table IV.5. Levels of Evidence for Practice Guides

Criteria	Strong Evidence Base	Moderate Evidence Base	Minimal Evidence Base
Validity	The research has high internal validity and high external validity based on studies that meet standards.	The research has high internal validity but moderate external validity, or high external validity but moderate internal validity.	The research may include evidence from studies that do not meet the criteria for moderate or strong evidence.
Effects on relevant outcomes	The research shows consistent positive effects without contradictory evidence in studies with high internal validity.	The research shows a preponderance of evidence of positive effects. Contradictory evidence must be discussed and considered with regard to relevance to the scope of the guide and the intensity of the recommendation as a component of the intervention evaluated.	There may be weak or contradictory evidence of effects.
Relevance to scope	The research has direct relevance to scope—relevant context, sample, comparison, and outcomes evaluated.	Relevance to scope may vary. At least some research is directly relevant to scope.	The research may be out of the scope of the practice guide.
Relationship between research and recommendations	Direct test of the recommendation in the studies, or the recommendation is a major component of the intervention tested in the studies.	Intensity of the recommendation as a component of the interventions evaluated in the studies may vary.	Studies for which the intensity of the recommendation as a component of the interventions evaluated in the studies is low, and/or the recommendation reflects expert opinion based on reasonable extrapolations from research.
Panel confidence	Panel has a high degree of confidence that this practice is effective.	The panel determines that the research does not rise to the level of strong but is more compelling than a minimal level of evidence. Panel may not be confident about whether the research has effectively controlled for other explanations or whether the practice would be effective in most or all contexts.	In the panel's opinion, the recommendation must be addressed as part of the practice guide; however, the panel cannot point to a body of research that rises to the level of moderate or strong.
Role of expert opinion	Not applicable.	Not applicable.	Expert opinion based on defensible interpretation of theory.
When assessment is the focus of the recommendation	Assessments meet the standards of *The Standards for Educational and Psychological Testing.*	For assessments, evidence of reliability meets *The Standards for Educational and Psychological Testing* but with evidence of validity from samples not adequately representative of the population	Not applicable.

	on which the recommendation is focused.

REFERENCES

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.

Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics, 32*(2), 151–179.

Hedges, L. V., & Olkin, I. (1985). Statistical methods for meta-analysis. Orlando, FL: Academic Press.

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.

APPENDIX A: POLICIES FOR PRIORITIZING STUDIES FOR REVIEW

Because of the large amount of research literature in the field of education, the WWC must prioritize topic areas for review and, within topic areas, prioritize the order in which interventions will be reviewed. Similarly, the WWC must determine whether studies are eligible to be reviewed as quick reviews and single study reviews and which topics will be investigated in the practice guide format. The purpose of this appendix is to describe the current policies and practices that govern decisions regarding what education interventions will be reviewed, what single studies will be reviewed and in what order, and what topics should be the focus of WWC practice guides.

A. Prioritizing Reviews for Intervention Reports

The WWC conducts reviews of interventions and generates intervention reports in areas determined by the Institute of Education Sciences (IES) to be of highest priority for informing the national education policy agenda. IES establishes its priorities based on nominations received from the public to the WWC Help Desk; input from meetings and presentations sponsored by the WWC; suggestions presented to IES or the WWC by senior members of education associations; input from state and federal policymakers; and literature scans to determine how much evidence on the effectiveness of interventions exists in various topic areas.

In consultation with the WWC contractors, IES determines the topic areas within which the WWC will conduct intervention reviews. To date, focal topic areas include those that have applicability to a broad range of students or to particularly important subpopulations; broad policy relevance; and at least a moderate volume of studies examining the effectiveness of specific, identifiable interventions.

In order to get new topic area reviews up and running quickly, a review team may conduct a quick start search, which focuses on a limited number of interventions. These interventions are identified by content expert recommendations of interventions with a large body of causal evidence likely to be of interest to educators, supplemented by interventions from key literature reviews and/or other topic areas meeting the same criteria.

After the initial search, a review team conducts a broad topic search to assess the literature related to a review topic. The goal is to identify all interventions that have been used to address the research questions of the review. Broad topic searches utilize a larger list of sources and a broader set of search parameters than those used in a quick start search. The review team, in collaboration with the content expert, develops a list of sources to be searched, as well as search parameters.

A review team will conduct an **intervention-specific search** to go "deep" into the literature of a particular intervention. The goal is to identify all publications on a particular intervention. Even if the review team has conducted a broad topic search, it must conduct an intervention-specific search before drafting a report on a given intervention.

The process for prioritizing interventions for review is based on a standard scoring system and is conducted annually for ongoing reviews and when new topic areas are established. Using information in the title and the abstract or introduction, the review coordinator scores the study based on research design and sample size. Only studies that relate to the review protocol of the

topic area (those that include the correct age range, achievement outcome measured, etc.) are considered eligible and are included in the ranking process. The scores of all the studies are combined for each intervention with a weighting factor based on whether there is an existing intervention report and its release date. Interventions with the highest scores are prioritized for review. The scoring criteria are presented below.

Study Criterion 1: Internal Validity

- Randomized controlled trial—3 points
- Regression discontinuity design or single-case design—2 points
- Quasi-experimental design—1 point
- None of the above—0 points

Study Criterion 2: Size

- The study receives one additional point if the study is large, defined as greater than 250 children or 10 classrooms. If the study size is not clear, the study does not receive any additional points.

After summing the scores across all studies within an intervention, the resulting score is multiplied by an intervention weight. The intervention weight is based on whether there is an existing intervention report and its release date.

- An intervention with no prior report gets a weight of 3.
- An intervention with a prior report gets a weight of $[1 + 0.1*(\text{current year} - \text{report release date})]^2$. For example, an intervention being prioritized in 2011 that had a report released in 2009 would get a weight of $[1 + 0.1*(2011 - 2009)]^2 = 1.44$.

Interventions are then prioritized for review based on the final scoring, with higher scores receiving higher prioritization.

The WWC also examines the "Google trend" for the top 10 interventions identified through the scoring process. Determining which interventions are being searched using the Google search engine provides a sense of the interventions of interest to the general public. The Google trends data can be considered when selecting an intervention for a systematic review, but does not contribute formally to the prioritization score.

B. Prioritizing Topics for Practice Guides

Practice guide topics are selected based on their potential to improve important student outcomes, their applicability to a broad range of students or to particularly important subpopulations, their policy relevance, the perceived demand within the education community, and the availability of rigorous research to support recommendations. In addition, IES may request that the WWC produce a practice guide on a particular issue. Suggestions for practice guide topics are welcomed. To suggest a topic, visit https://ies.ed.gov/ncee/wwc/ContactUs.aspx.

C. Prioritizing Studies for Single Study Reviews

Single study reviews are generally initiated in three ways: (a) IES requests a WWC review of a particular study, (b) a study is prioritized from public submissions to the Help Desk or from the IES-funded study list, or (c) a study meets the WWC criteria for a quick review.

First, IES may request that one of the WWC contractors complete a single study review for a variety of reasons. For example, IES may decide that a recently completed study is of sufficient importance that it warrants a review.

A second method by which studies become single study reviews is through regular review of "prioritization" lists of studies not under review by topic areas or practice guides. The prioritization lists include studies submitted through the Help Desk, IES-funded studies, and reviews of studies completed by external WWC-certified reviewers. The studies that are at the top of the prioritization list are selected for review as single study reviews.

Finally, a study can be selected for a single study review if it meets the following eligibility criteria for a WWC quick review:

- *The study must be released recently and reported on in a major national news source or a major education news publication.*

- *The study must examine the effectiveness of an intervention intended to directly or indirectly improve student academic and/or nonacademic outcomes.* Studies that do not examine the effectiveness of an intervention but that have been portrayed to do so in the media may still be eligible for a quick review.

Studies meeting the quick review eligibility criteria are forwarded to IES for a decision regarding whether the study warrants a WWC quick review.

APPENDIX B: POLICIES FOR SEARCHING STUDIES FOR REVIEW

Some WWC products, including intervention reports and practice guides, are the result of systematic, comprehensive searches of the literature. Using the search terms specified in the WWC review protocols (such as in Table B.1), trained staff identify and screen potentially relevant literature.

Table B.1. Sample Search Terms for WWC Literature Searches

Keywords	Related Search Terms
Intervention	Approach, curricul*, educational therapy, homework, improvement, instruct*, practice, program, remedial, school*, strategy, success*, teach*, treatment
Outcomes	Alphabetics, aural learning, comprehension, fluency, language, letter identification, lexicography, literacy, phonemic, phonetics, phonics, phonological, print awareness, print knowledge, readability, reading, verbal development, vocabulary, vocalization, word recognition
Population	Adolescent*, eighth grade, elementary school, eleventh grade, fifth grade, fourth grade, grade 4, grade 5, grade 6, grade 7, grade 8, grade 9, grade 10, grade 11, grade 12, high school, junior high, K–12, middle grades, middle school, ninth grade, seventh grade, sixth grade, student*, summer school, tenth grade, twelfth grade
Study design	ABAB design, affect*, assignment, causal, comparison group, control*, counterfactual, effect*, efficacy, evaluation*, experiment*, impact*, matched group, meta analysis, meta-analysis, posttest, post-test, pretest, pre-test, QED QES, quasi-experimental, quasiexperimental, random*, RCT, RDD, regression discontinuity, simultaneous treatment, SCD, single case, single subject, treatment, reversal design, withdrawal design

Note: This illustrative table is drawn from the Adolescent Literacy Review Protocol, version 3.0, found at https://ies.ed.gov/ncee/wwc/Document/29. The asterisk (*) in the related search term list allows the truncation of the term so that the search returns any word that begins with the specified letters.

Table B.2 displays a standard set of databases used during this process. Additional databases are listed in the topic area review protocol.

Table B.2. General Sources: Electronic Databases

Database	Description
Academic Search Premier	The multidisciplinary full text database contains peer-reviewed full text journals for more than 4,600 journals, including nearly 3,900 peer-reviewed titles and indexing and abstracts for more than 8,500 journals.
EconLit	The American Economic Association's electronic database is the world's foremost source of references to economic literature. There are more than 1.1 million records available.
Education Research Complete	The world's largest and most complete collection of full text education journals, ERC provides indexing and abstracts for more than 2,300 journals and full text for approximately 1,400 journals and 550 books and monographs.
E-Journals	The E-Journals database provides article-level access for thousands of e-journals available through EBSCOhost and EBSCO Subscription Services.
ERIC	Funded by the U.S. Department of Education, the Education Resource Information Center provides access to education literature and resources, including information from journals indexed in the Current Index of Journals in Education and Resources in Education Index. ERIC provides ready access to education literature to support the

	use of educational research and information to improve practice in learning, teaching, educational decision making, and research.
ProQuest Dissertations & Theses	Providing access to the world's most comprehensive collection of dissertations and theses, this is the database of record for graduate research, with more than 2.4 million dissertations and theses included from around the world.
PsycINFO	PsycINFO contains more than 1.8 million citations and summaries of journal articles, book chapters, books, dissertations, and technical reports, all in the field of psychology. Journal coverage includes international material selected from more than 1,700 periodicals in more than 30 languages. More than 60,000 records are added each year.
SAGE Journals Online	Provides access to the full text of articles in more than 500 leading journals published by SAGE on topics relating to psychology, early childhood, education, labor, statistics, and survey methodology.
Scopus	The world's largest abstract and citation database of peer-reviewed literature and quality web sources in the scientific, technical, medical, and social sciences, it covers more than 19,000 titles, articles in press, conference proceedings, and e-books.
SocINDEX	The world's most comprehensive and highest quality sociology research database features more than 2 million records and includes extensive indexing for books/monographs, conference papers, and other nonperiodical content sources, in addition to informative abstracts for more than 1,300 "core" coverage journals.
WorldCat	WorldCat is the world's largest network of library content and services, allowing users to simultaneously search the catalogues of more than 10,000 libraries for access to 1.5 billion books, articles, CDs, DVDs, and more.

The WWC also routinely searches websites of core and topic-relevant organizations to collect potentially relevant studies. The standard set of websites that is searched to identify studies appears in Table B.3, and a set of sources selectively targeted in relevant review efforts is listed in Table B.4. Additional websites may be listed in the review protocol.

Table B.3. General Sources: Websites

Abt Associates	Hoover Institution
Alliance for Excellent Education	Mathematica Policy Research
American Education Research Association	MDRC
American Enterprise Institute	National Association of State Boards of Education
American Institutes of Research	National Governors' Association
Best Evidence Encyclopedia	Policy Archive
Brookings Institution	Policy Study Associates
Carnegie Corporation of New York	RAND
Center for Research and Reform in Education	Regional Educational Laboratories
Congressional Research Service	SRI
Government Accountability Office	Thomas B. Fordham Institute
Grants/contracts awarded by IES	Urban Institute
Heritage Foundation	

Table B.4. Targeted Sources: Electronic Databases or Websites

After-School Alliance	Learning Disabilities Association of America
American Speech-Language-Hearing Association	Linguistic Society of America
Campbell Collaboration	Natl. Association for Bilingual Education
Carnegie Corporation for the Advancement of Teaching	Natl. Association of State Directors of Career Tech. Ed.
Center for Social Organization of Schools	Natl. Association of State Directors of Special Education
Chapin Hall Center for Children, University of Chicago	Natl. Autism Center - National Standards Project
CINAHL	Natl. Center for Learning Disabilities
Cochrane Central Register of Controlled Trials	Natl. Center on Response to Intervention
Cochrane Database of Systematic Reviews	Natl. Center on Secondary Education and Transition
Council for Exceptional Children	Natl. College Access Network
Council for Learning Disabilities	Natl. Dissemination Center for Children with Disabilities
Database of Abstracts of Reviews of Effects	Natl. Dropout Prevention Center/Network
Florida Center for Reading Research	Natl. Institute on Out-of-School Time at Wellesley
Harvard Family Research Project	NBER Working Papers
Institute for Higher Education Policy	Teachers of English to Speakers of Other Languages
Institute for Public Policy and Social Research	TA Ctr. on Social Emotional Interv. for Young Children

To determine whether new research is being mentioned in major news sources, and thus potentially eligible for a quick review, the WWC monitors major news sources, news clippings, and news aggregator services based on the Lexis Nexis list of major US newspapers. The Major US Newspapers source contains English language newspapers published in the United States that are listed in the top 50 in circulation in Editor & Publisher Year Book.

APPENDIX C: STAFFING, REVIEWER CERTIFICATION, AND QUALITY ASSURANCE

The purpose of this appendix is to describe the roles and responsibilities of WWC staff in developing WWC products, the certification of WWC reviewers, and the procedures in place for assuring WWC product quality.

A. Staffing for WWC Products

1. Intervention Reports

After an initial search, if there is enough literature to generate reviews of interventions for a topic area, methodology and content experts are identified as team leaders, and their names are submitted to the IES for approval. Once approved, if they are new to the WWC process, they receive training on substantive WWC content and operational procedures.

Together, the team leaders develop the review protocol for the topic area, provide methodological and content-specific support and guidance to the review teams working on reviews in the topic area, and play a central role in determining the content and quality of the final products. Throughout the process of reviewing studies, the lead methodologist reconciles differences between reviewers of a particular study; writes and reviews reports on interventions; makes technical decisions for the team; and serves as the point of contact for study authors, developers, and IES.

Other members of the review team include WWC-certified reviewers and review coordinators. WWC-certified reviewers are responsible for reviewing and analyzing relevant literature. Reviewers have training in research design and methodology and in conducting critical reviews of effectiveness studies; they have also passed a WWC-reviewer certification exam (see below for more details). As part of the team, these individuals review, analyze, and summarize relevant literature for evidence of effectiveness and assist in drafting intervention reports.

Coordinators support the team leaders, reviewers, and other review team members in managing the various aspects of the reviews. For example, coordinators work with library staff in overseeing the literature search process, screening the literature, organizing and maintaining communication, tracking the review process, overseeing review team staffing, and managing the production process.

2. Practice Guides

Practice guides are developed under the guidance of a panel composed of approximately six members. Each panel is chaired by a nationally recognized researcher with expertise in the topic. The panel consists of four researchers who have diverse expertise in the relevant content area and/or relevant methodological expertise, along with at least two practitioners who have backgrounds that allow them to offer guidance about implementation of the recommendations.

Working with the panel, WWC research staff develop the research protocol, review studies, and draft the guide. There are four primary roles: (a) an evidence coordinator, who ensures that the research used to support recommendations is rigorous and relevant; (b) a practice coordinator, who ensures that the discussion of how to implement each recommendation is concrete, specific, and appropriate; (c) WWC-certified reviewers, who assess whether supporting literature meets WWC standards; and (d) a panel coordinator, who arranges meetings and

manages other logistical needs or concerns. Ultimately, the practice guide is a result of the teamwork and consensus of the panel and research staff.

3. Single Study Reviews

Similar to the staffing structure for conducting reviews for intervention reports, single study reviews (including quick reviews) are conducted under the guidance of a lead methodologist as described above. When the subject of a single study falls under a topic area for which the WWC has developed a review protocol, the study is reviewed according to that protocol and with guidance from the content expert for that topic area. In other cases, the WWC identifies a content expert who has relevant expertise.

The lead methodologist for a single study review is responsible for ensuring that the study in question meets the criteria for being reviewed by the WWC. For each single study review, the team leader works with a minimum of two certified WWC reviewers in completing the requisite study review guide and preparing the report. The key responsibility of the lead methodologist in this process is to reconcile any differences in the judgments of the principal reviewers about the quality or findings of the study, resolve any technical issues or refer them to the senior WWC team for resolution, and review and ensure the quality of draft reports.

B. Reviewer Certification

All studies that are included in WWC products are systematically reviewed by WWC-certified reviewers who must successfully complete a training and certification process designed and administered by or under the supervision of the WWC. Potential reviewers are screened for appropriate and relevant expertise and experience in rigorous research design and analysis methods prior to being admitted to reviewer training. There are separate trainings and certification exams for group designs (including randomized controlled trials [RCTs] and quasi-experimental designs [QEDs]), regression discontinuity designs (RDDs), and single-case designs (SCDs). Group design trainings have entailed a 2-day interactive session that includes an overview of the WWC and its products and in-depth instruction on the WWC review standards, review tools, policies, and practices. Alternatively, the group design training can be completed using a set of video modules on the WWC website that cover the same material. Trainings for RDDs and SCDs are each 1 day. Information about WWC training and certification is posted on the website.

At the conclusion of training, participants pursuing certification are expected to take and pass a multiple-choice precertification examination. Those who pass the precertification exam are then required to complete and earn an acceptable grade on a full study review following the WWC study review guide. The review is graded by the certification team, with feedback provided to the participant. If the participant has not satisfactorily completed the review, he or she will be asked to review a second article. If the participant still has not attained a passing grade, he or she may be asked to complete a third review, as long as the second review showed improvement. If there is no apparent improvement, or the participant does not adequately complete the third review, he or she will not receive certification.

Upon the release of updated *Procedures* and *Standards Handbooks*, certified reviewers are required to attend a re-certification webinar and pass a multiple-choice recertification examination to be certified to review studies under the new standards.

C. Quality Assurance

1. Statistical, Technical, and Analysis Team

The WWC statistical, technical, and analysis team (STAT) is a group of highly experienced researchers who consider issues requiring higher-level technical skills, including revising existing standards and developing new standards. Additionally, issues that arise during the review of studies are brought to the STAT for its consideration.

2. Document Review

At each stage, reviewers examine the accuracy of the study reviews, evaluate the product for consistency and clarity, and ensure that the report conforms to WWC processes. It is only after intense review from several perspectives that a WWC product is released to the public.

After an extensive drafting and revision process with multiple layers of internal review, the completed draft is submitted to IES, which reviews the document internally and sends it out for external peer review by researchers who are knowledgeable about WWC standards and are not staff on a WWC contract. Both sets of comments are returned to the contractor's drafting team, which responds to each comment and documents all responses in a memo. The report undergoes a final review by IES staff to ensure that any issues have been addressed appropriately. Intervention reports for which no studies meet standards are subject only to IES review, not external peer review. Practice guides also undergo review by the U.S. Department of Education's Standards and Review Office.

3. Quality Review Team

The WWC Quality Review Team (QRT) addresses concerns about WWC reports raised by external inquiries through a quality review process. Inquiries must (a) be submitted in writing to the WWC Help Desk through the Contact Us page (https://ies.ed.gov/ncee/wwc/ContactUs.aspx), (b) pertain to a specific study or set of studies, and (c) identify and explain the specific issue(s) in the report that the inquirer believes to be incorrect. A QRT review is conducted by WWC staff who did not contribute to the product in question in order to determine the following:

- Whether a study that was not reviewed should have been reviewed
- Whether the rating of a study was correct
- Whether outcomes excluded from the review should have been included
- Whether the study's findings were interpreted correctly
- Whether computation procedures were implemented correctly

After an inquiry is forwarded to the QRT, a team member verifies that the inquiry meets criteria for a quality review and notifies the inquirer whether a review will be conducted. A member of the QRT is assigned to conduct an independent review of the study, examine the original review and relevant author and distributor/developer communications, notify the topic

area team leadership of the inquiry, and interview the original reviewers. When the process is complete, the QRT makes a determination on the inquiry.

If the original WWC decisions are validated, the QRT reviewer drafts a response to the inquirer explaining the steps taken and the disposition of the review. If the review concludes that the original review was flawed, a revision will be published, and the inquirer will be notified that a change was made as a result of the inquiry. These quality reviews are one of the tools used to ensure that the standards established by IES are upheld on every review conducted by the WWC.

4. Conflicts of Interest

Given the potential influence of the WWC, the U.S. Department of Education's National Center for Education Evaluation and Regional Assistance within IES has established guidelines regarding actual or perceived conflicts of interest specific to the WWC. WWC contractors administer this conflict of interest policy on behalf of the U.S. Department of Education.

Any financial or personal interests that could conflict with, appear to conflict with, or otherwise compromise the efforts of an individual because they could impair the individual's objectivity are considered potential conflicts of interest. Impaired objectivity involves situations in which a potential contractor, subcontractor, employee or consultant, or member of his or her immediate family (spouse, parent, or child) has financial or personal interests that may interfere with impartial judgment or objectivity regarding WWC activities. Impaired objectivity can arise from any situation or relationship, impeding a WWC team member from objectively assessing research on behalf of the WWC.

The intention of this process is to protect the WWC and review teams from situations in which reports and products could be reasonably questioned, discredited, or dismissed because of apparent or actual conflicts of interest and to maintain standards for high quality, unbiased policy research and analysis. All WWC product team members, including methodologists, content experts, panel chairs, panelists, coordinators, and reviewers, are required to complete and sign a form identifying whether potential conflicts of interest exist. Conflicts for all tasks must be disclosed before any work is started.

As part of the review process, the WWC occasionally will identify studies for review that have been conducted by organizations or researchers associated with the WWC. In these cases, review and reconciliation of the study are conducted by WWC-certified reviewers from organizations not directly connected to the research, and this is documented in the report.

Studies that have been conducted by the developer of an intervention do not fall under this conflict of interest policy. Therefore, the WWC does not exclude studies conducted or outcomes created by the developer of the product being reviewed. The authors of all studies are indicated in WWC reports, and the WWC indicates the source of all outcome measures that are used, including those created by the developer.

In combination with explicit review guidelines, IES review of all documents, and external peer review of all products, these conflict of interest policies achieve the WWC goal of transparency in the review process while also ensuring that WWC reviews are free from bias.

APPENDIX D: EXAMPLES OF STUDY DEFINITION

When two findings share at least three of the following characteristics, the WWC considers them parts of the same study:

- **Sample members, such as teachers or students.** Findings from analyses that include some or all of the same teachers or students may be related.

- **Group formation procedures, such as the methods used to conduct random assignment or matching.** When authors use identical (or nearly identical) methods to form the groups used in multiple analyses, or a single procedure was used to form the groups, the results may not provide independent tests of the intervention.

- **Data collection and analysis procedures.** Similar to group formation, when authors use identical (or nearly identical) procedures to collect and analyze data, the findings may be related. Sharing data collection and analysis procedures means collecting the same measures from the same data sources, preparing the data for analysis using the same rules, and using the same analytic methods with the same control variables.

- **Research team.** When manuscripts share one or more authors, the reported findings in those manuscripts may be related.

This appendix provides examples of how this rule is applied in different circumstances.

Example 1: Findings authored by the same research team. A research team presents findings on the effectiveness of an intervention using two distinct samples in the same manuscript. Because the same research team might conduct analyses that have little else in common, sharing only the research team is not sufficient for the WWC to consider the findings part of the same study. Therefore, these findings would be considered separate studies. But if the analyses in the manuscript also shared two of the remaining three characteristics, they would instead be considered the same study.

Example 2: Findings presented by gender. Within a school, authors stratified by gender and randomly assigned boys and girls to condition separately. The authors analyzed and reported findings separately by gender. The WWC would consider this to be a single study because all four of the characteristics listed above are shared by the two samples. First, the same teachers are likely present in both samples, so the sample members overlap. Next, even though boys and girls were randomly assigned to condition separately, the WWC considers strata or blocks within random assignment to be part of a single group formation process. Furthermore, the two samples likely share the same data collection and analysis procedures, and the research teams are the same. Considering this to be a single study is consistent with the goal of the WWC to provide evidence of effectiveness to a combined target population that includes both boys and girls.

Example 3: Findings presented by grade within the same schools. Within a middle school, authors randomly assigned youth to condition, separately by grade. The authors analyzed and reported findings separately by grade, but used the same procedures and data collection. The WWC would consider this to be a single study that tests the effect of an intervention for middle school students. Again, the two samples share all four characteristics.

Example 4: Findings presented by grade across different schools. Within each participating elementary and middle school, authors randomly assigned youth to condition,

separately by grade. The authors analyzed and reported findings separately for elementary and middle schools, and collected data on different outcome measures and background characteristics in the two grade spans. The WWC would consider this to be two distinct studies. The manuscripts share only two of the four characteristics: the data collection was different, and the samples do not overlap.

Example 5: Findings presented by cohort. Study authors randomly assign teachers within a school to intervention and comparison conditions. The study authors examine the impact of the intervention on achievement outcomes for third-grade students after 1 year (Cohort 1) and after 2 years (Cohort 2, same teachers but different students). The study authors report results for these two cohorts separately. The WWC would consider this to be a single study that tests the effect of an intervention on third-graders because the two samples share all four characteristics.

Example 6: Findings for the same students after re-randomization. Findings based on an initial randomization procedure and those based on re-randomizing the same units to new conditions might be considered different studies. Despite using different group formation procedures, the first condition is met because the sample members are the same. If the findings were reported by the same research team members, the second condition is also met. It is unlikely (but not impossible) that the same data collection and analysis procedures were used given the separation in time. If so, the findings share only two of the four characteristics, and the findings would be considered different studies.

Example 7: Findings reported by site separately over time. Separately for six states, study authors randomly assigned school districts within a state to intervention and comparison conditions. The same procedures were used at the same time to form the groups, and the same data elements were collected in all six states. The authors published each state's findings separately, releasing them over time. The final report used a different analytic approach from the previous reports. The authors of the reports changed, but each report shared at least one author with the original report. The WWC would consider all of these but the final report to be a single study of the intervention, because the same group formation procedures were used, the same data collection and analysis procedures were used, and the reports all shared at least one research team member with another report. However, the WWC would consider findings from the site in the final report to be a separate study; because a different analytic approach was used, the findings from the final site only share two characteristics with the findings in the earlier reports.

Example 8: Findings from related samples, based on different designs. Study authors randomly assigned students to condition and conducted an RCT analysis. Using a subsample of the randomly assigned students, the same authors also examined a QED contrast that also examined the effectiveness of the intervention. They used different analysis procedures for the two designs. The WWC would consider the QED findings as a separate study from the RCT findings because the findings share only two of the four characteristics: sample members and research team. The WWC considers matching approaches to identifying intervention and comparison groups part of the analysis procedure, so a matching analysis based on data from an RCT would be considered to use different analysis procedures from an analysis of the full randomized sample, even if the analytical models were otherwise identical.

Example 9: Findings reported for multiple contrasts. If authors compare an intervention group to two different comparison groups, the WWC would consider both contrasts to be part of

the same study. They share a research team, sample members, and the group formation process (the intervention group in both contrasts is the same). Because there are many different business-as-usual conditions, all comparisons between the intervention and a comparison group are informative and should be presented as primary findings. However, if a contrast is between two versions of the intervention, the findings should be presented as supplemental.

APPENDIX E: MAGNITUDE OF FINDINGS FOR RANDOMIZED CONTROLLED TRIALS AND QUASI-EXPERIMENTAL DESIGNS

The results of analyses can be presented in a number of ways, with varying amounts of comparability and utility. To the extent possible, the WWC attempts to report on the findings from studies in a consistent way, using a common metric and accounting for differences across analyses that may affect their results. This appendix describes WWC methods for obtaining findings, including specific formulae for computing the size of effects, that are comparable across different types of eligible designs with a comparison group.

A. Effect Sizes

To assist in the interpretation of study findings and facilitate comparisons of findings across studies, the WWC computes the effect size (ES) associated with study findings on outcome measures relevant to the area under review. In general, the WWC focuses 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 effect size estimates across studies but also allows us to draw upon existing conventions from the research community to establish the criterion for substantively important effects for intervention rating purposes. Different types of effect size indices have been developed for different types of outcome measures because of their distinct statistical properties.

1. Studies with Student-Level Assignment

The sections that follow focus on the WWC's default approach to computing student-level effect sizes, or teacher-level effect sizes when the outcome is not based on aggregating data on students (such as teacher retention). We describe procedures for computing Hedges' *g* based on results from the different types of statistical analyses that are most commonly encountered. When possible, the WWC reports on and calculates effect sizes for post-intervention means adjusted for the pre-intervention measure. If a study reports both unadjusted and adjusted post-intervention means, the WWC reports the adjusted means and unadjusted standard deviations.

a. Continuous Outcomes

Effect Sizes from Standardized Mean Difference (Hedges' g)

For continuous outcomes, the WWC has adopted the most commonly used effect size index, the standardized mean difference. It 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) of that outcome measure. Given that the WWC generally focuses on student-level findings, the default SD used in effect size computation is the student-level SD.

The basic formula for computing standardized mean difference follows:

$$g = \frac{\bar{y}_i - \bar{y}_c}{S}$$

$$S = \sqrt{\frac{(n_i - 1)s_i^2 + (n_c - 1)s_c^2}{n_i + n_c - 2}},$$

where y_i and y_c are the means of the outcome for the intervention and comparison groups, respectively; n_i and n_c are the student sample sizes; s_i and s_c are the student-level SDs; and S is the pooled within-group SD of the outcome at the student level. Combined, the resultant effect size is given by

$$g = \frac{y_i - y_c}{\sqrt{\frac{(n_i - 1)s_i^2 + (n_c - 1)s_c^2}{n_i + n_c - 2}}}.$$

The effect size index thus computed is referred to as Hedges' *g*. This index differs from the Cohen's *d* index in that Hedges' *g* uses the square root of degrees of freedom, *N–k* for *k* groups, for the denominator of the pooled within-group SD, *S*, whereas Cohen's *d* uses the square root of sample size, *N*, to compute *S* (Rosenthal, 1994; Rosnow, Rosenthal, & Rubin, 2000). 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 effect size estimate by multiplying the Hedges' *g* by a factor of $\omega = [1 - 3/(4N - 9)]$, with *N* being the total sample size. Unless otherwise noted, Hedges' *g* corrected for small-sample bias is the default effect size measure for continuous outcomes used in the WWC's review.

$$g = \frac{\omega(y_i - y_c)}{\sqrt{\frac{(n_i - 1)s_i^2 + (n_c - 1)s_c^2}{n_i + n_c - 2}}}$$

In certain situations, however, the WWC may present study findings using effect size measures other than Hedges' *g*. For example, if the SD of the intervention group differs substantially from that of the comparison group, the lead methodologist 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 effect size as Glass's Δ instead of Hedges' *g*. The justification is that when the intervention and comparison groups have unequal variances, as they do when 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 also may use Glass's Δ, Hedges' *g* without the small sample size adjustment ω, or Hedges' *g* using a qualitatively similar SD that is calculated differently than described above to present study findings if there is not enough information available for computing Hedges' *g* as described above. These deviations from the default will be clearly documented in the WWC's review process.

Effect Sizes from Student-Level t*-tests or ANOVA*

For randomized controlled trials (RCTs) with low attrition, study authors may assess an intervention's effects based on student-level *t*-tests or analyses of variance (ANOVA) without statistical adjustment for pretest or other covariates (see Chapter III). If the study authors reported posttest means and SD as well as sample sizes for both the intervention and comparison groups, the computation of effect size will be straightforward using the standard formula for Hedges' *g*.

When means or SD are not reported, the WWC can compute Hedges' *g* based on *t*-test or ANOVA *F*-test results, if they were reported along with sample sizes for both the intervention group and the comparison group. For effect sizes based on *t*-test results,

$$g = \omega t \sqrt{\frac{n_i + n_c}{n_i n_c}}.$$

For effect sizes based on ANOVA F-test results,

$$g = \omega \sqrt{\frac{F\left(n_i + n_c\right)}{n_i n_c}}.$$

Effect Sizes from Student-Level t*-tests or ANCOVA*

Analysis of covariance is a commonly used analytic method for quasi-experimental designs (QEDs). It assesses the effects of an intervention while controlling for important covariates, particularly a pretest, that might confound the effects of the intervention. ANCOVA also is used to analyze data from RCTs 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' *g* as the *covariate-adjusted* mean difference divided by the *unadjusted* pooled within-group SD:

$$g = \frac{\omega\left(y_i' - y_c'\right)}{\sqrt{\frac{\left(n_i - 1\right)s_i^2 + \left(n_c - 1\right)s_c^2}{n_i + n_c - 2}}},$$

where y_i' and y_c' are the *covariate-adjusted* posttest means of the outcome for the intervention and comparison groups, respectively.

The use of *covariate-adjusted* mean difference as the numerator of *g* ensures that the effect size estimate is adjusted for any 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 *g* allows comparisons of effect size estimates across studies by using a common metric (the population SD as estimated by the unadjusted pooled within-group SD) to standardize group mean differences.

A final note about ANCOVA-based effect size computation is that Hedges' *g* cannot be computed based on the *F*-statistic from an ANCOVA. 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' *g*, however, requires the use of unadjusted within-group SD. Therefore, we cannot compute Hedges' *g* with the *F*-statistic from an ANCOVA in the same way that we compute *g* with the *F*-statistic from an ANOVA. However, if

the correlation between pre- and posttest, *r,* is known, we can derive Hedges' *g* from the ANCOVA *F*-statistic as follows:

$$g = \omega\sqrt{\frac{F\left(n_i + n_c\right)\left(1 - r^2\right)}{n_i n_c}}.$$

Difference-in-Differences Adjustment

Study authors will occasionally report unadjusted group means on both pre- and posttest but not adjusted group means and adjusted group mean differences on the posttest. Absent information on the correlation between the baseline and outcome measures, the WWC computes the effect size of the difference between the two groups on the baseline and outcome measures separately using Hedges' *g*, with the final effect size given by their difference:

$$g = g_{post} - g_{pre}.$$

This "difference-in-differences" approach to estimating an intervention's effects, even though it takes into account the group difference on the baseline measure, is not necessarily optimal, because it is likely to either overestimate or underestimate the adjusted group mean difference, depending on which group performed better on the pretest. If the intervention group had a higher average baseline measure than the comparison group, the difference-in-differences approach is likely to underestimate the adjusted group mean difference; otherwise, it is likely to overestimate the adjusted group mean difference. Moreover, this approach does not provide a means for adjusting the statistical significance of the adjusted mean difference to reflect the relationship between the baseline and outcome measures. Nevertheless, it yields a reasonable estimate of the adjusted group mean difference, similar to what would have been obtained from an analysis of gain scores, a commonly used alternative to the covariate adjustment-based approach to testing an intervention's effect.

The difference-in-differences approach presented assumes that the correlation between the baseline and outcome measures is unknown. However, this statistic might be reported by the study authors based on the study data, or in some areas of educational research, empirical data on the relationship between the baseline and outcome measures may be available. If such data are dependable, the lead methodologist may choose to use the empirical relationship to estimate the adjusted group mean difference rather than the difference-in-differences approach. If the empirical relationship is dependable, the covariate-adjusted estimates of the intervention's effects will be less biased than those based on the difference-in-differences approach. A methodologist who chooses to compute effect size using an empirical relationship between the baseline and outcome measures must provide an explicit justification for the choice as well as evidence of the credibility of the empirical relationship. Computationally, if the baseline and outcome measures have a correlation of *r*, then

$$g = \frac{\omega\left[\left(y_i - y_{i0}\right) - r\left(y_c - y_{c0}\right)\right]}{\sqrt{\frac{\left(n_i - 1\right)s_i^2 + \left(n_c - 1\right)s_c^2}{n_i + n_c - 2}}},$$

where y_{i0} and y_{c0} are the unadjusted pretest means for the intervention and comparison groups, respectively.

The WWC also reports the adjusted group mean difference, which is calculated as

$$d = g * \sqrt{\frac{\left(n_i - 1\right)s_i^2 + \left(n_c - 1\right)s_c^2}{n_i + n_c - 2}}.$$

When the difference-in-differences adjustment is used, the statistical significance will be based on the adjusted effect. For example, consider a pre-intervention difference of 0.2 on an achievement test. If the post-intervention difference were 0.3, the difference-in-differences adjusted effect would be 0.1. Subsequently, the statistical significance would be based on the adjusted finding of 0.1 rather than the unadjusted finding of 0.3.

b. Dichotomous Outcomes

Effect Sizes from Log Odds Ratio

Although not as common as continuous outcomes, dichotomous outcomes are sometimes used in studies of educational interventions. Examples include dropping out versus staying in school, grade promotion versus retention, and passing versus failing a test. In such cases, a group mean difference appears as a difference in the probability of the occurrence of an event. The effect size measure of choice for dichotomous outcomes is the odds ratio, which has many statistical and practical advantages over alternative effect size measures, such as the difference between two probabilities, the ratio of two probabilities, and the phi coefficient (Fleiss, 1994; Lipsey & Wilson, 2001).

The odds ratio (OR) builds on the notion of odds. For a given study group, the odds for the occurrence of an event is defined as follows:

$$Odds = \frac{p}{\left(1 - p\right)},$$

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

$$OR = \frac{p_i\left(1 - p_c\right)}{p_c\left(1 - p_i\right)},$$

where p_i and p_c are the probabilities of the occurrence of an event for the intervention and the comparison groups, respectively.

As is the case with effect size computation for continuous variables, the WWC computes effect sizes for dichotomous outcomes based on student-level data in preference to aggregate-level data for studies that have a multilevel data structure. The probabilities 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 the odds ratio into a log odds ratio (LOR) to simplify statistical analyses:

$$LOR = ln\left(OR\right).$$

The LOR has a convenient distribution form, which is approximately normal with a mean of 0 and an SD of π divided by the square root of 3, or 1.81. The LOR also can be expressed as the difference between the log odds, or logits, for the two groups:

$$LOR = ln\left(Odds_i\right) - ln\left(Odds_c\right),$$

which shows more clearly the connection between the log odds ratio and the standardized mean difference (Hedges' *g*) for effect sizes.

To make the LOR comparable to the 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" the LOR. Based on a Monte Carlo simulation study of seven different types of effect size indices for dichotomous outcomes, Sanchez-Meca, Marin-Martinez, and Chacon-Moscoso (2003) concluded that the effect size index proposed by Cox (1970) is the least biased estimator of the population standardized mean difference, assuming an underlying normal distribution of the outcome. Therefore, the WWC has adopted the Cox index as the default effect size measure for dichotomous outcomes. The computation of the Cox index is straightforward:

$$d_{Cox} = \omega \frac{LOR}{1.65}.$$

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

Difference-in-Differences Adjustment

For dichotomous outcomes, the effect size of the difference between the two groups on the pretest and posttest is computed separately using Hedges' *g*, with the final effect size given by their difference:

$$g = g_{post} - g_{pre}.$$

c. Gain Scores

Some studies report only the means and standard deviations of a gain score for the two groups, which are inadequate for computing effect sizes. To be reported by the WWC, effect sizes from gain score analyses must be based on standard deviations of the outcome measure collected at the follow-up time point without adjustment for the baseline measure. Effect sizes calculated using standard deviations of gain scores or standard deviations of the outcome measure after adjusting for baseline measures are not comparable to effect sizes calculated using standard deviations of unadjusted posttest scores. The effect size based on the gain score standard deviations will generally be larger, because the standard deviation of gain scores is typically smaller than the standard deviation of unadjusted posttest scores. The WWC will not report effect sizes based on the gain score standard deviations, but gain score means can be used.

2. Studies with Cluster-level Assignment

The effect size formulae presented are based on student-level analyses, which are appropriate analytic approaches for studies with student-level assignment. However, the case is more complicated for studies with assignment at the cluster level (e.g., assignment of teachers, classrooms, or schools to conditions), when data may have been analyzed at the student level, the cluster level, or through multilevel analyses. Such analyses pose special challenges to effect size computation during WWC reviews. In the remainder of this section, we discuss these challenges and describe the WWC's approach to handling them.

a. Effect Sizes from Student-Level Analyses of Cluster-Level Assignment

The main problem with student-level analyses in studies with cluster-level assignment is that they violate the assumption of the independence of observations underlying traditional hypothesis tests and result in underestimated standard errors and inflated statistical significance (see Appendix G). However, the estimate of the group mean difference in such analyses is unbiased and can be appropriately used to compute the student-level effect sizes using methods described in previous sections.

b. Cluster-Level Effect Sizes

Studies that report findings from cluster-level analyses sometimes compute effect sizes using cluster-level means and SDs. However, the WWC will not report effect sizes based on the cluster-level SDs for two reasons. First, the intra-class correlation (ICC) yields cluster-level SDs that are typically much smaller than student-level SDs,

$$SD_{Cluster} = SD_{Student} * sqrt\left(ICC\right),$$

which subsequently results in much larger cluster-level effect sizes that are incomparable with the student-level effect sizes that are the focus of WWC reviews. Second, the criterion for "substantively important" effects (see Chapter IV) was established specifically for student-level effect sizes and does not apply to cluster-level effect sizes. Moreover, there is not enough knowledge in the field for judging the magnitude of cluster-level effects, so a criterion of "substantively important" effects for cluster-level effect sizes cannot be established.

c. Student-Level Effect Sizes from Cluster-Level Analyses

Computing student-level effect sizes requires student-level SDs, which are often unreported in studies with cluster-level analyses.

It is generally not feasible to compute the student-level SD based on cluster-level data. As seen from the relationship presented above, we could compute student-level SDs from cluster-level SDs and the ICC, but the ICC is rarely provided. Also, note that the cluster-level SD associated with the ICC is not exactly the same as the observed SD of cluster means that is 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). If the outcome is a standardized measure that has been administered to a norming sample (national or state), then the effect size may be calculated using the SD from the norming sample.

d. Student-Level Effect Sizes from Multilevel Modeling

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). 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 technical details, all are based on similar ideas and underlying assumptions.

Similar to student-level ANCOVA, HLM also can adjust for important covariates, such as a pretest, when estimating an intervention's effect. However, rather than assuming independence of observations such as ANCOVA, 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, some 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' *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,

$$g = \frac{\omega\gamma}{\sqrt{\frac{(n_i - 1)s_i^2 + (n_c - 1)s_c^2}{n_i + n_c - 2}}},$$

where γ is the HLM coefficient for the intervention's effect, representing the group mean difference adjusted for both level-1 and level-2 covariates, if any. The level-2 coefficients are adjusted for the level-1 covariates under the condition that the level-1 covariates are either not centered 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 teachers or classrooms). The idea could easily be extended to a three-level model (e.g., students nested with teachers who were, in turn, nested within schools).

3. When Student-Level Effect Sizes Cannot Be Computed

In some cases, the WWC will be unable to calculate an effect size from the data reported by the study authors that can be compared to effect sizes for other studies and outcome measures. This could occur because the data are missing, the only SD reported uses cluster-level data or is based on gain scores, or the WWC requires a statistical adjustment to satisfy the baseline equivalence requirement, but cannot calculate an appropriately adjusted effect size. Nevertheless, such studies will not be excluded from WWC reviews and may still potentially contribute to intervention ratings, as explained next.

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. The quality of design is not affected by whether a WWC-reportable effect size could be computed; therefore, such studies can still meet WWC standards and be included in intervention reports. However, to be eligible to meet WWC design standards, a study must report some information about the magnitude or direction of the impact estimate.

When WWC-reportable student-level effect sizes cannot be calculated for a finding, the WWC will exclude the finding from the computation of domain average effect sizes and improvement indices. Additionally, the finding will be not be considered substantively important.

B. Improvement Index

In order to help readers judge the practical importance of an intervention's effect, the WWC translates the effect size into an improvement index. This 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., the 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 percentage point 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 described next.

Step 1. Convert the Effect Size (Hedges' 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 using a table that lists the proportion of the 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 the 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.

Step 2. Compute Improvement Index = U3 – 50%

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 member and an average comparison group member 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.

REFERENCES

Cooper, H. (1998). *Synthesizing research: A guide for literature review*. Thousand Oaks, CA: Sage.

Cox, D. R. (1970). *Analysis of binary data*. New York, NY: Chapman & Hall/CRC.

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, NY: Russell Sage Foundation.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*(2), 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.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*(2), 199–213.

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, NY: Russell Sage Foundation.

Rosnow, R. L., Rosenthal, R., & Rubin, D. B. (2000). Contrasts and correlations in effect-size estimation. *Psychological Science, 11*(6), 446–453.

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.

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

APPENDIX F: STATISTICAL SIGNIFICANCE FOR RANDOMIZED CONTROLLED TRIALS AND QUASI-EXPERIMENTAL DESIGNS

In order to adequately assess the effects of an intervention, it is important to know not only the magnitude of the effects as indicated by the effect size or improvement index, but also the statistical significance of the effects.

A. Clustering Correction for Mismatched Analyses

However, the correct statistical significance of findings is not always readily available, particularly in studies in which 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) and 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 (2007). Because the clustering correction will decrease the statistical significance (or increase the *p*-value) of the findings, nonsignificant findings from a mismatched analysis will remain nonsignificant after the correction. Therefore, the WWC applies the correction only 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 effect size 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 statistical 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.

Step 1. Compute the t-*Statistic for the Effect Size, Ignoring Clustering*

$$t = g \Big/ \sqrt{\frac{n_i + n_c}{n_i n_c} + \frac{g^2}{2\left(n_i + n_c\right)}},$$

where g is the effect size that ignores clustering and n_i and n_c are the sample sizes for the intervention and comparison groups, respectively, for a given outcome. For domain average effect sizes, n_i and n_c are the average sample sizes for the intervention and comparison groups, respectively, across all outcomes within the domain.

Step 2. Correct the t-*Statistic for Clustering*

$$t_a = t\sqrt{\frac{(N-2)-2\left(\frac{N}{M}-1\right)\rho}{(N-2)\left[1+\left(\frac{N}{M}-1\right)\rho\right]}},$$

where N is the total sample size at the student level ($N = n_i + n_c$), M is the total number of clusters in the intervention (m_i) and comparison (m_c) groups, and ρ is the ICC for a given outcome.

If the ICC is reported by the author, it is used in the calculation above. However, the value of the ICC often is not available from the study reports. Based on empirical literature in the field of education, the WWC has adopted default ICC values of 0.20 for achievement outcomes and 0.10 for behavioral and attitudinal outcomes (Schochet, 2008). The topic area team leadership may set different defaults in the review protocol with justification.

For domain average effect sizes, the ICC used above 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. This gives the study the benefit of the doubt by crediting the measure with the most statistical power, so the WWC's rating of interventions will not be unduly conservative.

Step 3. Compute the Degrees of Freedom Associated with the t-*Statistic Corrected for Clustering*

$$df = \frac{\left[(N-2)-2\left(\frac{N}{M}-1\right)\rho\right]^2}{(N-2)(1-\rho)^2 + \frac{N}{M}\left(N-2\frac{N}{M}\right)\rho^2 + 2\left(N-2\frac{N}{M}\right)\rho(1-\rho)}$$

Step 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 (df). 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 = TDIST(t_a,df,2).

B. Benjamini-Hochberg Correction for Multiple Comparisons

Type I error and the statistical significance of findings also may be inflated when study authors perform multiple hypothesis tests simultaneously. The traditional approach to addressing the problem is the Bonferroni method (Bonferroni, 1935), which lowers the critical p-value for individual comparisons by a factor of $1/m$, with m equal to the total number of comparisons made. However, the Bonferroni method has been shown to be unnecessarily stringent for many

practical situations; therefore, the WWC has adopted the Benjamini-Hochberg (BH) method (Benjamini & Hochberg, 1995) to correct for multiple comparisons or multiplicity.

The BH method adjusts for multiple comparisons by controlling false discovery rate (FDR) instead of family-wise error rate (FWER). It is less conservative than the traditional Bonferroni method, yet it still provides adequate protection against Type I error in a wide range of applications. Since its conception in the 1990s, growing evidence has shown that the FDR-based BH method may be the best solution to the multiple comparisons problem in many practical situations (Williams, Jones, & Tukey, 1999).

The WWC applies the BH correction only to main findings, and not to supplementary findings. As is the case with clustering correction, the WWC applies the BH correction only to statistically significant findings, because nonsignificant findings will remain nonsignificant after correction (but all main findings that meet WWC design standards in the study are counted when making the correction). For findings based on analyses when 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 effect size could be computed, we convert the effect size to *t*-statistics and then obtain the corresponding *p*-values. For findings based on mismatched analyses that do not account for the correlation in outcomes for individuals within clusters, we 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 in which 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, although it is "very often not needed, and yields too conservative a procedure" (Benjamini & Yekutieli, 2001, p. 1183). The modified version of the BH procedure uses *a* over the sum of the inverse of the *p*-value ranks across the *m* comparisons instead of *a*.

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 when not accounted for in the analysis. 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.

1. Multiple Outcome Measures Tested with a Single Comparison Group

The most straightforward situation that may require the BH correction occurs when the study authors assessed the effect of an intervention on multiple outcome measures within the same outcome domain using a single comparison group. For studies that examined measures in multiple outcome domains, the BH correction is applied to the set of findings *within the same domain* rather than across different domains.

Step 1. Rank Order the Findings Based on Unadjusted Statistical Significance

Within a domain, order the *p*-values in ascending order such that

$$p_1 < p_2 < p_3 < \ldots < p_m,$$

where *m* is the number of significant findings within the domain.

Step 2. Compute Critical p-*Values for Statistical Significance*

For each *p*-value, p_x, compute the critical value, p_x':

$$p'_x = \frac{x\alpha}{M},$$

where *x* is the rank for p_x, with *x = 1, 2, ..., 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 may be less than the number of outcomes the study authors actually examined for two reasons: (a) the authors may not have reported findings from the complete set of comparisons they had made, and (b) certain outcomes assessed by the study authors may not meet the eligibility or standards requirements of the WWC review. The target level of statistical significance, α, in the numerator 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.

Step 3. Identify the Cutoff Point

Identify the largest *x,* denoted by *y,* that satisfies the condition

$$p_x \leq p'_x.$$

This establishes a cutoff point such that all findings with *p*-values smaller than or equal to p_y are statistically significant, and findings with *p*-values greater than p_y are not significant at the prespecified level of significance 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 indicates only whether the finding is significant at the prespecified 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, resulting in six statistically significant effects and two nonsignificant effects based on properly aligned analyses. To correct the significance of the findings for multiple comparisons, first rank-order the author-reported (or clustering corrected) *p*-values in the first column of Table G.1 and list the *p*-value ranks in the second column.

Then compute $p_x' = x\alpha/M$ with $M = 8$ (because there are eight outcomes in the domain) and $\alpha = 0.05$ and record the values in the third column. Next, identify y, the largest x that meets the condition $p_x \leq p_x'$; in this example, $y = 5$, $p_5 = 0.030$, and $p_5' = 0.031$. Note that for the fourth outcome, the *p*-value is greater than the new critical *p*-value. This finding is significant after correction because it has a *p*-value (0.027) lower than the highest *p*-value (0.030) to satisfy the condition.

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

Author-Reported or Clustering Corrected *p*-value (p_x)	*p*-value Rank (x)	New Critical *p*-value ($p_x' = 0.05x/8$)	Finding *p*-value ≤ New Critical *p*-value? ($p_x \leq p_x'$)	Statistical Significance after BH Correction?
0.002	1	0.006	Yes	Yes
0.009	2	0.013	Yes	Yes
0.014	3	0.019	Yes	Yes
0.027	4	0.025	No	Yes
0.030	5	0.031	Yes	Yes
0.042	6	0.038	No	No
0.052	7	0.044	No	No
0.076	8	0.050	No	No

Thus, we can claim that the five findings associated with a *p*-value of 0.030 or smaller are statistically significant at the 0.05 level after correction for multiple comparisons. The sixth finding (*p*-value = 0.042), although reported as being statistically significant, is no longer significant after the correction.

2. Single Outcome Measure Tested with Multiple Comparison Groups

Another type of multiple comparison problem occurs when the study authors tested an intervention's effect on a given outcome by comparing the intervention group with multiple comparison groups or by comparing multiple interventions.

Currently, the WWC does not have specific guidelines for studies that use multiple comparison groups. Teams have approached these studies by (a) including all comparisons they consider relevant, (b) calculating separate effect sizes for each comparison, and (c) averaging these findings together in a manner similar to multiple outcomes in a domain (see previous section). The lead methodologist should use discretion to decide the best approach for the team on a study-by-study basis.

3. When Study Authors Account Only for Some Multiplicity or Across More Findings than Required

In general, the WWC applies the BH corrections collectively to all of the main findings within a study for an outcome domain. However, a more complicated multiple comparison problem arises when the authors of a study took into account the multiplicity resulting from some findings, but not others. For example, consider a study in which authors accounted for

multiplicity resulting from multiple comparison groups, but not the multiplicity resulting from multiple outcome measures. For such a study, the WWC needs to correct only the findings for the multiplicity resulting from multiple outcomes. Specifically, BH corrections are made separately to the findings for each comparison group. For example, with two comparison groups (A and B) and three outcomes, the review team applies the BH correction separately to the three findings for A and the three findings for B. If the authors accounted for multiplicity across a subset of the main findings in a domain, but not across well-defined groups (such as an outcome measures or comparison groups), the WWC will ask the authors for the unadjusted *p*-values, and perform its own BH correction across all of the main findings.

In another scenario, the authors may have accounted for multiple comparisons across more findings than the WWC requires. In this case, the WWC will use the authors' corrected significance levels.

REFERENCES

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.

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. In *Studi in onore del Professore Salvatore Ortu Carboni* (pp. 13–60). Rome.

Hedges, L. V. (2007). Correcting a significance test for clustering. *Journal of Educational and Behavioral Statistics, 32*(2), 151–179.

Schochet P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62–87.

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.

APPENDIX G: REPORTING REQUIREMENTS FOR STUDIES THAT PRESENT A COMPLIER AVERAGE CAUSAL EFFECT

1. Reporting of Complier Average Causal Effects Estimates in WWC Products

Among RCTs, any complier average causal effects (CACE) estimate that addresses a research topic relevant to a WWC product will be reviewed, so long as it meets the eligibility criteria specified in the previous section. However, the ways in which a study's CACE estimates are reported in WWC products will vary depending on the type and focus of the product and the availability of intent-to-treat (ITT) estimates, as follows.

RCT studies that report both an ITT and CACE estimate on the same outcome. For this type of study, both the ITT and CACE estimate will be reviewed under their respective standards, and the WWC will report the estimates and their ratings as follows:

- If the study is being reviewed for a single study review, then the single study review will report both types of estimates and their ratings. The review will also make note of which estimate, if any, was identified by the study authors as the main focus of the study.
- If the study is being reviewed for an intervention report or practice guide, then only one of the two types of estimates will contribute to the intervention rating (in intervention reports) or the level of evidence (in practice guides). The lead methodologist for the intervention report, or the evidence coordinator for the practice guide, will have discretion to choose which estimate is used. For example, this choice may be based on which type of research question—effects of *being assigned* to an intervention versus effects of *receiving* an intervention—is the most common question addressed by other studies included in the WWC product. Alternatively, the choice may be based on which type of research question is deemed to be of greatest interest to decision makers. Once a particular type of estimate (ITT or CACE) is selected, the other estimate will be mentioned only in a footnote or appendix.

RCT studies that report only a CACE estimate. The WWC prefers to review both the ITT and CACE estimates and report these in WWC products as described above, but some studies may not report the ITT estimate. For this type of study, the WWC will first query the study authors to determine whether they conducted an ITT estimate. If so, the ITT estimate will be included in the review. If the authors do not provide the ITT estimate, then only the CACE estimate will be reviewed and included in intervention ratings or levels of evidence determinations.

2. Reporting Requirements for Estimated Variances of CACE Estimates

As in all study designs, the WWC relies on valid standard errors to assess the statistical significance of reported impacts. Statistical significance factors into how findings are characterized. For CACE estimates, valid standard errors need to reflect the error variance in the estimated relationships between instruments and the outcome *and* the error variance in the estimated relationships between instruments and the endogenous independent variable, as well as the covariance of these errors. Two analytic methods for estimating standard errors account for all of these sources of variance. The WWC regards standard errors estimated from the following methods as valid:

- **Two-stage least squares (2SLS) asymptotic standard errors.** These standard errors reflect all types of error discussed above. Standard statistical packages report them for 2SLS estimation.
- **Delta method.** In the case of one instrument, the 2SLS estimate is the ratio of the ITT estimate and the estimated first-stage coefficient on the instrument. The delta method (see, for example, Greene, 2000) can be used to express the variance of the CACE estimator as a function of these coefficients, the variance of the ITT estimator, the variance of the first-stage coefficient, and the covariance between the ITT estimator and the first-stage coefficient.

In all cases, when the unit of assignment differs from the unit of analysis, standard errors must account appropriately for clustering.

As in other study designs, the rating that a CACE estimate receives will not depend on whether standard errors are valid. However, if a study reports an invalid standard error, the WWC will not use the reported statistical significance of the CACE estimate in characterizing the study's findings. The CACE estimate can still be classified as substantively important if it meets the criteria for a substantively important designation.

REFERENCE

Greene, W. (2000). *Econometric analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.