What Works Clearinghouse™

Procedures Handbook,
Version 4.1

Contents

Tables

Figures

I. Introduction

It is critical that education decisionmakers have access to the best evidence about the effectiveness of education practices, products, programs, and policies. However, it can be difficult, time consuming, and costly for decisionmakers 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, such as adolescent literacy, primary mathematics, or charter schools. The protocol defines the parameters for the research to be included within the scope of the review, including population characteristics and types of interventions; the literature search terms and databases, if any; and any topic-specific applications of the standards, including acceptable thresholds for sample attrition, risk from joiners in cluster design studies, and characteristics for establishing 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 *Identifying relevant literature.* Studies are gathered through a comprehensive search of published and unpublished publicly available research literature.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.

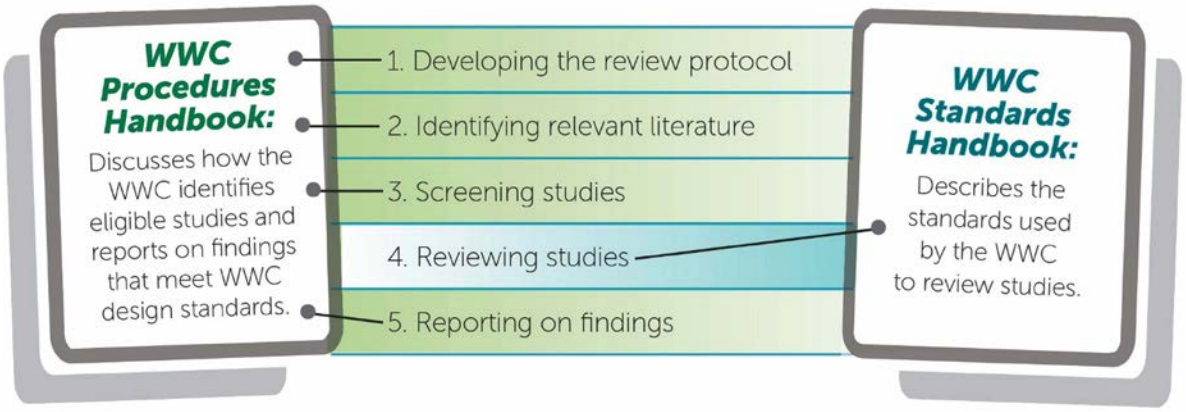
4. *Reviewing studies.* Every eligible study is reviewed against WWC standards. The WWC uses a structured review process to assess the causal validity of findings reported in education effectiveness research. The WWC standards focus on the causal validity within the study sample—that is, *internal* validity—rather than the extent to which the findings might be replicated in other settings—that is, *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, such as those that receive significant media attention. 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 4.1,* provides a detailed description of the procedures used by the WWC in the systematic review process—specifically, steps 1–3 and step 5 described previously. A separate *WWC 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*. Figure I.1 shows how the steps of the WWC systematic review process are divided between the *Standards Handbook* and the *Procedures Handbook*.

Figure I.1. Steps of the What Works Clearinghouse systematic review process and the What Works Clearinghouse handbooks



This new *Procedures Handbook* updates the previous *Version 4.0* to *Version 4.1*. The following substantive updates were made:

- **The removal of the "substantively important" designation.** In previous versions of the *Procedures Handbook*, an effect size above 0.25 was deemed "substantively important" and noted when characterizing findings. This designation has been removed in this updated version. Effect sizes are now judged only by their statistical significance and sign.
- **The addition of standard error calculations for all effect sizes.** For any study that meets WWC standards, the WWC will estimate an effect size and corresponding standard error. These estimated standard errors will be used in the estimation of the fixed-effects model.
- **The addition of single-case design (SCD) procedures for synthesizing SCD study findings using design-comparable effect sizes.** The requirement that there be at least

five SCD studies meeting WWC standards and that these be conducted by at least three nonoverlapping teams and involve at least 20 cases for the WWC to rate the effectiveness of interventions on the basis of SCD evidence (the "5-3-20" rule) no longer applies. For any study that is rated *Meets WWC SCD Standards*, the WWC will estimate an appropriate effect size if feasible and appropriate. The WWC will infer positive or potentially positive effects of the intervention in intervention reports and practice guides based on statistical hypothesis tests of the fixed-effects meta-analytic estimate of the mean effect size in each outcome domain from all studies meeting WWC standards. A description of the visual and statistical methods used to estimate design-comparable effect sizes for SCDs is provided. The WWC no longer reports effectiveness ratings using the proportion of SCD experiments demonstrating positive effects on the basis of visual analysis (that is, the approach documented in a January 2017 Handbook supplement).

- **The addition of procedures to estimate effects from regression discontinuity designs (RDDs).** The estimation of RDD effect sizes have been clarified and in some cases, added entirely.
- **Clarification of decision rules determining the use of difference-in-difference effect sizes.** The WWC estimates effect sizes that adjust for baseline differences using various calculations. The decision rules dictating specific effects estimated have been clarified.
- **The synthesis of studies within intervention reports using a fixed-effects model.** In previous versions of the *Standards Handbook* and the *Procedures Handbook*, the WWC used an unweighted average to synthesize effect sizes within intervention reports. This version replaces that procedure with a fixed-effects meta-analytic model, in which each effect size is weighted by the inverse of its variance.
- **The modification of the intervention report effectiveness rating.** In previous versions of the *Standards Handbook* and the *Procedures Handbook*, the WWC used a version of vote counting—a simple method of comparing the number of studies with positive outcomes to the number of studies with negative outcomes—to provide the intervention report effectiveness rating. The WWC will now use the fixed-effects meta-analytic average, its statistical significance, and the proportion of weight from studies that have a rating of *Meets WWC Standards Without Reservations* to determine the effectiveness rating. The fixed-effects estimation procedure weights studies by a function of their sample size, and as a result larger studies have a bigger impact, relative to small studies, on the average effect size.
- **Levels of evidence in practice guides.** When assigning levels of evidence for practice guide recommendations, the WWC will include as criteria the extent of evidence meeting WWC standards and the effectiveness ratings corresponding with each recommendation, defining these ratings in the same manner for practice guides as for intervention reports.

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 whether 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 and other WWC staff may occasionally need additional guidance. If necessary, the WWC will produce guidance documents to provide clarification and interpretation of WWC procedures.

As the WWC continues to refine and develop procedures, the *Procedures Handbook* will be revised or supplemented to reflect these changes. Any written supplements for use in combination with this *Procedures Handbook* will be specified in the protocol governing the corresponding study reviews. Readers who want to provide feedback on the *Procedures Handbook,* or the WWC in general, may contact us at https://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, as described in 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, including motivation for the reviews and the questions to be addressed by the review.
- *Handbook version.* Protocols specify which version of the *Standards Handbook* and *Procedures Handbook* will be used for the reviews.
- *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 principles for literature searches by the WWC. A protocol also may provide special instructions regarding searching additional sources of literature that may not be found in academic databases.
- *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—including a lead methodologist and content experts as described 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 and sample characteristics for eligible student populations, along with subgroups of interest to the review. For example, a protocol may specify a focus on samples of students in kindergarten through grade 8 that are at least 50 percent English learner students. The protocol may specify a minimum required sample size for the WWC to report study findings, which may depend on the population of study or the study design.
  - *Eligible interventions.* Protocols provide descriptions of the types of interventions that fall within the bounds of the review, including the nature of the intervention; the settings in which the intervention is delivered; and the minimum duration of implementation for the intervention; and whether the intervention is "branded"—that is, a commercial program or product. For an example, a protocol may focus on both "branded" literacy programs used in regular classrooms in grades K–8 and on supplemental, afterschool reading intervention practices for students in the same grades.

  - *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 main outcomes of primary interest for the review. Both student outcomes and outcomes for teachers and other educators may be eligible for WWC review. Depending on the outcome measure, the protocol may specify higher standards for assessing WWC review requirements, such as reliability, than are required in the *Standards Handbook*.

- *Evidence standard parameters.* 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, but are not limited to, the following:
  - The choice of boundary that separates acceptable and unacceptable levels of sample attrition.
  - The measures on which studies must demonstrate baseline equivalence.
  - The psychometric properties of the forcing variable in RDD studies.
  - Certain parameters related to cluster-level assignment studies, which are studies that assign groups rather than individuals to conditions.

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 as described in 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 and developers, researchers, and the public to the WWC Help Desk. Only studies written in English that are publicly available—that is, accessible on a public website – or available through a publication, such as a book or journal—at the time of the literature search are eligible for WWC review. The WWC also reviews some individual studies outside of the systematic review process, such as those that receive significant media attention (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, in accordance with the principles described in 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 and screen studies that have been previously reviewed by the WWC, such as for another product or under a previous version of the standards.

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. All abstracts identified through database searches are screened by trained WWC staff; these staff members work independently to identify abstracts that clearly do not meet the eligibility criteria specified in the protocol. 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, rather than synthesizing findings from other studies, to examine the effectiveness of an intervention.
  - *Eligible designs*. The WWC includes findings from randomized controlled trials (RCTs), quasi-experimental designs (QEDs), RDDs, and SCDs. 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 for WWC review.
  - *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, but are not limited to, age, grade range, gender, socioeconomic status, disability status, 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. Unless specified otherwise in the review protocol, 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 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 and its territories, or in certain academic settings.
  - *Interventions*. Review protocols describe the interventions that are eligible for review and any additional eligibility characteristics related to the interventions, such as information about the method of delivery, the replicability of the intervention, and the characteristics of individuals implementing the 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.[1] 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, such as a five-year study of an intervention, which may release interim annual reports. Alternatively, a manuscript can include multiple studies, as in the case of an article including several separate experiments. In the case of multiple manuscripts that report on one study, the WWC prioritizes the final, peer-reviewed manuscript that is submitted to ERIC. If a final, peer-reviewed manuscript has not been submitted to ERIC, then the preferred version is the final, published version.

The critical issue in defining a study as distinct from a related analysis is whether it provides an independent test of the intervention. That is, does it provide new evidence that is uncorrelated with 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 the following:

- **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.

[1] More information about practice guides and intervention reports is available online at https://whatworks.ed.gov.

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 (see appendix D for examples). In particular, when two findings meet this condition, they demonstrate the following:

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.

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 previously, the review team leadership—including the lead methodologist and content experts as described further in appendix C—has the discretion to determine what constitutes a single study or multiple studies. The decision is clearly noted in the WWC product that includes the review.

B. The What Works Clearinghouse study review process

This section describes the WWC's process for reviewing studies that are eligible for review. The WWC review is completed based on information available in the study report and related manuscripts and, potentially, on information obtained from study authors via an **author query**. Generally, author queries are sent to clarify information needed to arrive at study eligibility, such as the percentage of students who identify as English language learners, or a study rating, such as descriptive statistics of participants at baseline and information about the statistical model used to estimate effects. The WWC does not ask authors to conduct new analyses. Information obtained during the author query process is noted in the review documentation and becomes part of the WWC's public record of the review. Figure V.1 displays the review process for reviews of studies by the WWC.

Each study receives a first review that is documented in a study review guide (SRG). The SRG is described at https://ies.ed.gov/ncee/wwc/StudyReviewGuide. In most cases, the WWC study rating can be determined based on the information available in the study and related reports.

The following process guides studies that do not require an author query:

- The first reviewer determines that the study's rating does not meet WWC standards.
  - A senior member of the team conducts quality assurance on the review.
  - If the senior member agrees, then the master SRG—the finalized version of the SRG—is created and completed.
  - If the senior member disagrees, then the study receives a second review and uses the following steps.
- The first reviewer determines that the study's rating meets WWC standards or the senior member disagrees with the first reviewer's original decision.
  - A second review is conducted. The second review is always conducted independently of the first.
  - After the second review is complete, the reviews are reconciled into a single master review. If there are disagreements between the first and second reviewers on key components of the review—such as the level of attrition, assessment of baseline

equivalence, which outcomes to include in the review, or effect sizes—then these should be resolved with the assistance of review team leadership.
  - A senior member of the review team examines the reconciled review documentation and determines whether the rating and supporting information are correct.

In the cases where the WWC must send an author query, the process below is followed.

- The first reviewer determines that a study does not meet WWC standards based on information available in the study and related reports, but the study might meet WWC standards with additional information from the study authors.
  - A senior member of the team reviews the study and concurs with the assessment.
  - The first reviewer prepares the author query, and it is sent by the WWC. Generally, the WWC generally asks study authors to reply to an author query within two weeks in order to expedite the completion of the WWC review, although the exact timeframe will be determined by review team leadership.
  - Should review team leadership deem it necessary, a second (provisional) review may also be completed based on the information currently available to the WWC. This second review is conducted independently as above.
  - If no response is received, then the review will be completed on the basis of the information already available from the text of the study.
  - If a response is received and it changes the rating, then the reviewers are reconciled and a senior member of the review team finalizes the report as above.
- The first reviewer determines that the study meets WWC standards, but the study is missing information used to estimate effect sizes or provide additional study context:
  - A second review is completed.
  - The first reviewer prepares the author query, and it is sent by the WWC. Generally, the WWC generally asks study authors to reply to an author query within two weeks in order to expedite the completion of the WWC review, although the exact timeframe will be determined by review team leadership.
  - Regardless of response of author query, the two reviewers are reconciled and a senior member of the review team finalizes the report as above.

C. Re-reviewing individual studies

Occasionally, the WWC might need to re-review a study previously reviewed by the WWC. This occurs for two reasons. The most common reason is that the study has been identified for review using a protocol that differs from the one that guided the original review. For example, a study might have been originally reviewed for the adolescent literacy topic area and later identified for review for the secondary writing topic area. A second circumstance that might prompt the WWC to re-review a study is that the study has been identified for review by the WWC and the original review was conducted under version 2.0 or earlier of the *WWC Standards and Procedures Handbook.*

In both of these cases, a new review is needed. The review process unfolds as described in section V.B, whereby two independent reviewers conduct a new review using the updated *WWC Standards Handbook, Version 4.1,* and protocol. To prevent unintended re-reviews of studies that

do not meet the criteria outlined above, the Online Study Review Guide has a study locking feature to prevent duplicative effort.

D. What Works Clearinghouse reviews and Standards for Excellence in Education Research

The Standards for Excellence in Education Research are a set of IES-wide principles, distinct from WWC design standards, to encourage and acknowledge high-quality education research studies along several dimensions, such as documentation of core components of the intervention and of the counterfactual condition and reporting of cost information. For more information about the Standards for Excellence in Education Research principles and their use across IES, visit https://ies.ed.gov/seer.asp.

Figure V.1. Roadmap of the study review process for group design studies by the What Works Clearinghouse

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, such as corrections for 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, including systematic reviews for 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.

1. Magnitude of a finding

In general, the WWC reports the magnitude of study findings in two ways: effect sizes—that is, standardized mean differences—for continuous outcomes or outcome domains containing both continuous and dichotomous outcomes and percentage-point impacts for strictly dichotomous outcomes. The WWC also sometimes uses the Cox index to translate impacts on dichotomous outcomes into effect sizes and to calculate *t* statistics to assess the statistical significance of impacts on dichotomous outcomes in order to better compare them with impacts on continuous outcomes.

In addition, the WWC may report the magnitude of impacts using other metrics, such as Cohen's U_3—a measure of the percentage of the intervention group with outcomes above the comparison group's average—or Cohen's U_3 minus 50 percentage points, which the *WWC Procedures Handbook, Version 4.0,* described as the WWC "improvement index." More details for all of these calculations are provided next.

Effect sizes for group design studies

For all studies, the WWC records the study findings in the units reported by the study authors. In addition, for continuous outcomes or dichotomous outcomes being synthesized with continuous outcomes, 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 sample bias. For group design studies, this effect size 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, given by

[VI.1.0] $= \frac{(\quad)}{(\quad)}.$

The WWC uses the unadjusted student-level standard deviations to estimate equation VI.1.0. When unadjusted student-level standard deviations are not available (for example, when adjusted standard deviations are reported), the WWC sends an author query requesting the unadjusted standard deviations. Should the unadjusted standard deviations not be available after sending an author query, the WWC uses one of the procedures described in appendix E to estimate the effect.

In addition, we define ω as the small sample size correction the effect size (Hedges, 1981), which is given by

[VI.1.1] $\omega = [1 - 3/(4N - 9)],$

where N is the sum of n_i and n_c defined above.

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 logistic similar to a 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

[VI.1.2] $= \quad\quad /1.65.$

The WWC follows these guidelines when calculating effect sizes from continuous outcomes:

- **The study provides pretest-adjusted means.** The WWC prefers the pretest-adjusted means over the unadjusted means when estimating effect sizes. Therefore, when both are available, the WWC uses the pretest-adjusted means and unadjusted standard deviations to estimate effect sizes.

- **The study provides unadjusted means at pretest and posttest using the same test.** The WWC prefers the difference-in-difference adjustment that subtracts the pretest mean from the posttest mean, within each condition. See appendix E for a full description of the effect estimate. 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 section II.A of the *WWC Standards Handbook, Version 4.1*.

- **The study provides unadjusted means at pretest and posttest using a different, but sufficiently related test.** 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. See appendix E for a full description of the effect estimate. 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 section II.A of the *WWC Standards Handbook, Version 4.1*.

- **The study provides unadjusted means at posttest.** The WWC estimates using equation V1.1.0.

The WWC reports statistical significance levels for the adjusted differences that reflect the adjustment in the effect size. For example, consider a preintervention difference of 0.2 on an achievement test. If the postintervention difference were 0.3, then 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. Standard errors for all effect size estimates can be found in appendix E.

Finally, 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 previously.

Effect sizes for regression discontinuity designs studies

For RDD studies that are rated *Meets RDD Standards With Reservations* or *Meets RDD Standards Without Reservations*, the WWC will calculate the effect size in the same manner as a group design study. For both continuous and dichotomous outcomes, the predicted means or probabilities must be calculated using the same statistical model that is used to estimate the impact on the outcome at the cutoff.

For continuous outcomes, the numerator of the effect size is the difference between the predicted group means, with each mean estimated using data from the corresponding side of the cutoff. The standard deviations and sample sizes used to standardize the impact estimate should be the standard deviations of the treatment and comparison groups from the full sample (as opposed to just those units within an optimal bandwidth that weights observations relative to their distance from the cutoff). If it might be possible to compose more than one treatment and comparison group (such as with a fuzzy RDD), then the treatment and comparison groups should be formed based on treatment assignment.

For dichotomous outcomes, the Cox index should be calculated using the predicted probabilities at the cutoff for the intervention and comparison groups, using the corresponding data above and below the cutoff.

Effect sizes for SCD studies

For SCD studies that are rated *Meets WWC SCD Standards With Reservations* or *Meets WWC SCD Standards Without Reservations,* the WWC will calculate a design-comparable effect size (D-CES) where feasible and appropriate in the judgment of review team leadership. The D-CES is comparable with a standardized mean-difference effect size, that is intended to be interpreted similarly to the Hedges' *g*, the effect size used by WWC for group design studies (Pustejovsky, Hedges, & Shadish, 2014; Shadish, Hedges, & Pustejovsky, 2014).

SCDs involve multiple observations in treatment and comparison conditions for each individual. Despite the name, SCDs typically involve data from several individuals. For each individual, there are multiple observations within each treatment phase.

A D-CES can be computed for a study that has three or more participants in a design that is multiple baseline across individuals, multiple probe across individuals, or a treatment reversal $(AB)^k$ design. In each case, the numerator of the effect size is a mean of the difference between observations in the treated and comparison conditions, averaged across individuals. The denominator of the effect size is an estimate of the between-person-within-condition standard deviation. Because the observations within persons are correlated, the computation of the degrees of freedom of the denominator and the variance of the effect size is more complex than in conventional between-subjects designs. Moreover, the number of degrees of freedom in the denominator is typically close to the number of subjects, which is often rather small so that the bias correction, analogous to that used to compute Hedges' *g*, is quite important.

The statistical details and formulas for computing design-comparable effect sizes are given in appendix E. For a more complete exposition, see Hedges, Pustejovsky, and Shadish (2012); Hedges, Pustejovsky, and Shadish (2013); and Pustejovsky et al. (2014).

Computing the D-CES requires access to raw outcome data by case, by observation occasion, and by treatment phase. The preferred method of obtaining raw data, if not presented in a suitable form in the paper being evaluated, is from the study authors. If study authors do not provide raw data but clear graphs are provided in the paper, then WWC reviewers may also use a graph-digitizing software to extract the individual points from a graph.

When estimating the D-CES, the WWC will begin with the following default specifications:

1. Use restricted maximum likelihood as the default estimator.

2. Specify the intervention effect as a fixed effect.

3. Assume "no trend" at baseline or any later phases for the estimation of the D-CES in multiple baseline designs.

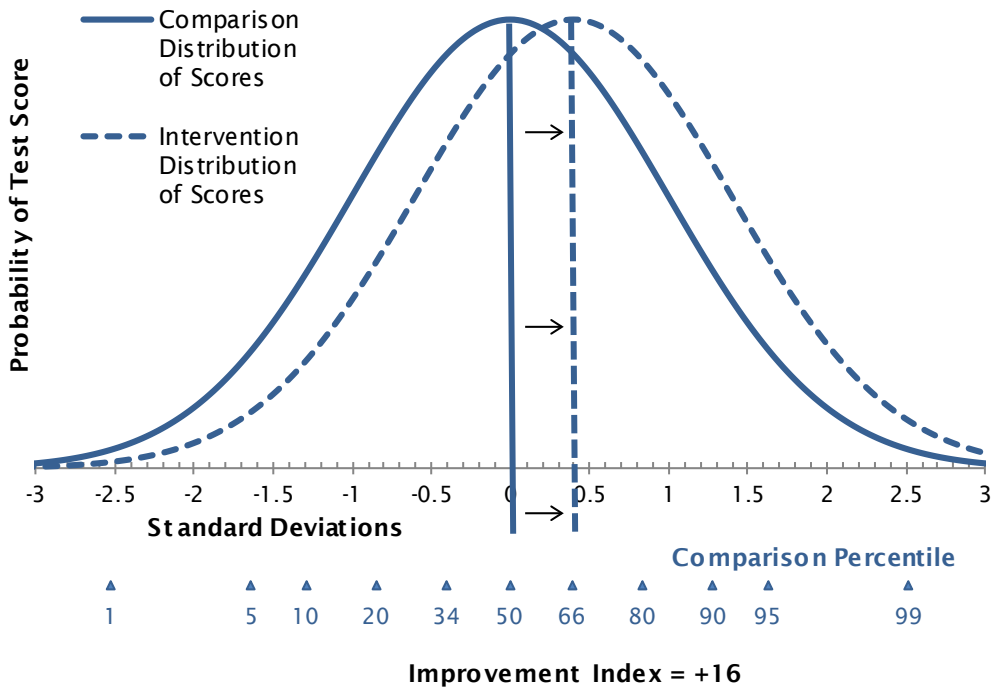
Review team leadership may determine, on the basis of visual analysis or an appropriate algorithm, that the underlying data do not conform to the above specifications. The review team may, after consultation with the content and methodological experts, either change the above specifications or not compute the D-CES, if an appropriate method is not available. The WWC will document in the SRG the rationale for any departures from the default specifications for computing the D-CES.

Improvement index

To help readers judge the practical importance of the magnitude of an intervention's effect, the WWC may translate 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 in the comparison group distribution (the latter being 50 percent by definition). 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 that student had received the intervention.

Figure VI.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 deviation 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 deviation above the mean (or at the 66th percentile of students) if he or she had received the intervention. 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 VI.1. Computation of the What Works Clearinghouse 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 and effect size, as described previously. For the WWC, a statistically significant estimate of an effect is one for which the null hypothesis was evaluated and rejected using a nondirectional test and a type I error rate of $\alpha = .05$.

The WWC generally accepts the statistical significance levels reported by the author(s) of the study. In some cases, the WWC will need to compute statistical significance for an outcome, for example, if statistical significance is unreported by study authors. To compute statistical significance, the WWC will use the effect size and standard error formulas reported in appendix E. For example, the *t* statistic for group mean differences on continuous measures is calculated using:

[VI.2.0] $= \frac{}{\quad \quad}$,

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 (Hedges & Olkin, 1985).

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

[VI.2.1] $= 1.65 \frac{}{\quad (\quad) \quad \quad (\quad)}$,

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). These WWC-calculated or recalculated estimates will be used in WWC products; the WWC's technical documentation will include a discussion of any corrections or modifications of author-reported probability values. A comprehensive list of all WWC-calculated effect sizes and their standard errors may be found in appendix E.

Clustering correction for "mismatched" analyses

In clustered trials (either random or nonrandom), groups of participants—such as classrooms or schools, as opposed to individuals—are assigned to conditions. Participants in preexisting groups tend to be more similar to one another than they are to participants in other preexisting groups. For example, students in one school are more like each other than they are like students in other schools. This similarity of individuals within a cluster means that students in the same cluster cannot be treated as independent, a core assumption underlying most of the statistical tests that are done in education, the social sciences, and in medicine. It is relatively common for analyses to be carried out at the individual level when assignment was done at the cluster level. The failure to account for clustering in the data analysis is sometimes known as a "mismatch" problem. The primary issue is that ignoring the correlation between responses among individuals within the same clusters results in standard errors that are too small, and therefore, the

probability values arising from the statistical tests are also too small. In other words, the null hypothesis is too likely to be rejected when the data analysis does not account for clustering.

To assess the statistical significance of an intervention's effects in cases where study authors have assigned at the cluster level but analyzed at the student level, the WWC computes clustering-corrected statistical significance estimates based on guidance from 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 intraclass correlation coefficient (ICC). As defaults, the WWC uses the ICC values of .20 for achievement outcomes and .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, the WWC reviews all eligible main findings as defined by the applicable study review protocol. The study rating is specified as the highest rating obtained across all eligible main findings. In general, *main findings*

- Answer confirmatory rather than exploratory research questions.
- Correspond with the full sample assigned to the intervention rather than subsets of that sample, unless the full sample is ineligible for review under the study review protocol.
- Rely on aggregate or composite measures and main outcomes as defined in the review protocol, rather than subscales or supplementary outcomes that may also be eligible for review.
- Focus on benchmark analyses rather than sensitivity analyses.
- Focus, for interventions in grade 12 or lower, on the earliest time point after receipt of the intervention, unless a different time point is specified in the study review protocol.

Supplementary findings include additional analyses eligible for review under the protocol. The WWC will review supplementary findings only if specified in advance for the purpose of the review.

Author queries are conducted as needed to evaluate all eligible analyses (main and supplemental) reviewed from the study.

1. Presentation of findings from multiple analyses

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 grade 8 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 supplementary 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 supplementary 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 supplementary finding.
- **The pooled analysis meets standards, and none of the cohort-specific analyses meet standards.** The pooled analysis is presented as a main finding, with no supplementary findings.
- **The pooled analysis does not meet standards, but all 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 in section VI.B.2. The findings from the analyses for separate cohorts are presented as supplementary 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, then 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 VI.B.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 supplementary 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.[2] If the only finding meeting standards in this example were instead for a separate subscale of a composite measure, then the WWC would report the finding for the subscale that meets WWC design standards as the main finding.

[2] 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, but not both, 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.

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 previously 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 supplementary findings from among those findings 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 details about assessing statistical significance in reviews of studies with multiple comparison groups.

When an author reports a set of sensitivity analyses that focus on the same or very similar samples, but applies different analytic methods to obtain each finding, the WWC designates as the main finding the finding that receives the highest WWC rating, accounts for the baseline measure(s) specified in the review protocol, uses the most comprehensive sample, and is most robust to threats to internal validity, based on the judgment of 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 supplementary 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, as described next, the WWC adjusts for multiple comparisons among all main findings, but not supplementary findings.

2. Magnitude of findings

To determine the magnitude of an aggregate effect, 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. For other studies, the aggregate result may not meet WWC design standards. Examples include a middle school math study that presents the effects separately for students in grades 6, 7, and 8; 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 may query authors to learn whether they conducted an analysis on the full sample. 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, then the WWC averages results **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, then the WWC computes an aggregate finding. For continuous outcomes, defining n_{gj}, m_{gj}, and s_{gj} as the sample size, outcome mean, and standard deviation for subsample g in group j, respectively, the combined group mean (M_j) across all subsamples and the combined standard deviation (S_j) are given by

[VI.3.0] $= \frac{}{}$ and $= \frac{}{}$.

The effect size g is then given by

[VI.3.1] $= \frac{(\quad)}{(\quad)}$.

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 combined probabilities across subsamples P_i and P_c using:

[VI.3.2] $= \frac{}{}$ and $= \frac{}{}$.

Then, the effect size is given by the Cox index using P_i and P_c:

[VI.3.3] $= \frac{}{} \frac{}{} /1.65.$

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.

For WWC products 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 fixed-effects meta-analysis of the study average effect sizes.

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 VI.A. For study average effect sizes based on continuous outcome measures, g is the average effect size across findings. For study average effect sizes based on dichotomous outcome measures expressed using the Cox index, 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 effect sizes for the sample outcome measure across *independent*[3] subsamples, the *t* statistic is derived in the same way as described for single effects, using the standard error formulas reported in appendix E. However, the sample sizes for the intervention and comparison groups become cumulative (that is, the total number of intervention and comparison participants across subsamples). For example, the *t* statistic for a mean differences on continuous measures is calculated using:

[VI.4.0a] $= \frac{}{}$, or

where *g* is the effect size based on M_j and S_j 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.

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

[VI.4.0b] $= 1.65 \frac{}{\sqrt{\frac{}{(\quad)}\frac{}{}+\frac{}{(\quad)}\frac{}{}}}$,

where d_{Cox} is the effect size based on the Cox index using P_i and P_c as the average probability of a positive outcome for students in the intervention and comparison groups, respectively, and N_i and N_c are the total sample sizes across the subsamples for the intervention and comparison groups, respectively.

For WWC-aggregated findings from *dependent* samples, the variance of the domain average effect is a function of the correlation among effect sizes, the number of effect sizes, and the effect size variances. For example, the *t* statistic for the standardized mean difference effect size is calculated using the following:

[VI.4.1] $= \frac{}{\sqrt{\frac{}{}}}$,

where *k* is the total number of dependent effect sizes within an outcome domain within a study, is the average correlation among outcome measures,[4] is the *i*th effect size variance, and is the *j*th effect size variance.[5] Any missing study correlations relevant to are assumed to be 1.0. The denominator in VI.4.1 is general and applicable to any of the effect size and variance standard error estimators presented in appendix E.

[3] Independent samples are those that do not share any participants. Dependent samples are those that share any study participants.

[4] The variance estimator for dependent effects within an outcome domain relies on the correlations between effect sizes, which is a function of the correlations between outcome measures. In general, the two correlations are *very* similar, especially when the correlation between measures is positive, which is reasonable in this context. When they differ, the correlation between measures will be slightly larger than the correlation between effect sizes, resulting in a slightly conservative variance estimate (Thompson & Becker, 2014).

[5] The summation notation treats pairs as unordered (for example, $i = 2$ and $j = 4$ is distinct from $i = 4$ and $j = 2$), meaning that $1 = (\quad 1)$.

Benjamini-Hochberg correction for multiple comparisons

Sometimes there is more than one hypothesis test within a domain. In these cases, 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 sign and statistical significance of findings as well as the extent of evidence. The summaries are based on findings that meet WWC design standards, regardless of study design, 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 three categories: statistically significant positive—that is, favorable—effect, indeterminate effect, and 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 three categories apply.

Table IV.1. What Works Clearinghouse characterization of findings of an effect based on a single outcome measure within a domain

Characterization	Criteria
Statistically significant positive effect	The estimated effect is positive and statistically significant, correcting for clustering when not properly aligned.
Indeterminate effect	The estimated effect is not statistically significant.
Statistically significant negative effect	The estimated effect is negative and statistically significant, correcting for clustering when not properly aligned.

Note: For the WWC, a statistically significant estimate of an effect is one for which the null hypothesis was evaluated and rejected using a nondirectional test and a type I error rate of $\alpha = .05$. 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.

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

Because they are not directly comparable with individual-level effect sizes, the results based on the analysis of aggregate data cannot be combined with student-level findings when calculating pooled effect sizes and an intervention effectiveness rating. 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. However, the statistical significance and sign of cluster-level findings is taken into account in determining the characterization of study findings.

Table IV.2. What Works Clearinghouse characterization of findings of an effect based on multiple outcome measures within a domain

Characterization	Criteria
Statistically significant positive effect	When any of the following is true: 1. At least one main 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-aggregated main finding 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.
Indeterminate effect	When any of the following is true: 1. None of the main findings are statistically significant, accounting for multiple comparisons, and correcting for clustering when not properly aligned; and the WWC-aggregated main finding is not statistically significant, correcting for clustering when not properly aligned. 2. At least one main finding is statistically significant and positive and at least one main finding is statistically significant and negative, accounting for multiple comparisons, and correcting for clustering when not properly aligned.
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-aggregated main finding 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: For the WWC, a statistically significant estimate of an effect is one for which the null hypothesis was evaluated and rejected using a nondirectional test and a type I error rate of $\alpha = .05$. 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.

In addition to characterizing study findings as described above, the WWC uses the U.S. Department of Education's definitions for "evidence-based" interventions from the final

regulation under 34 C.F.R. §77.1(c) to characterize the "evidence tier" of study findings meeting WWC standards.[6] These designations are separate from the review of each study using the *WWC Standards Handbook, Version 4.1,* and do not affect the rating of a study as meeting WWC design standards. Necessary criteria for a study being a source of tier 1 ("strong") evidence include requirements that the study be rated *Meets WWC Standards Without Reservations* and report a statistically significant and positive effect confirmed by the WWC. A study rated *Meets WWC Standards With Reservations* can be a source of tier 2 ("moderate") evidence.

Studies reviewed by the WWC may be found at the WWC's review of individual studies and data from individual studies' webpages. Users can filter and download study and finding-level information by topic area, WWC study rating, and evidence tier as well as by other criteria.

2. Summary of evidence for a What Works Clearinghouse intervention report

An intervention report is a publication that characterizes the effectiveness of an intervention on the basis of a systematic review of studies by the WWC. The intervention, which may be a "branded" program or product, is a replicable combination of core components identified by the review team in collaboration with content experts. If findings on the intervention meet WWC standards, the WWC provides a rating of the intervention's effectiveness within each outcome domain and characterizes the extent of evidence for that rating.

Intervention rating

As illustrated in table IV.3, the intervention rating for each outcome domain is composed of three elements. The first element is the number of studies. An outcome domain must be assessed in at least two studies meeting WWC standards in order for the intervention to receive the highest rating of "positive effects" for that outcome domain. The second element is the statistical significance of the outcome. If only one study in an intervention report assesses a particular outcome domain, then the WWC's assessment of the statistical significance of that outcome is used. When multiple studies included in an intervention report assess outcomes in the same domain, the WWC computes an average effect size using a fixed-effects model (Hedges & Vevea, 1998; see appendix H). In this model, the effect sizes observed in the individual studies are weighted by a function of their effective sample size. Larger studies receive proportionally more weight in the analysis. The statistical significance of the average effect size for each outcome domain is then derived using a *z* test. The third element relates to the relative contribution of those studies that receive a study rating of *Meets WWC Standards Without Reservations* versus those studies that receive a study rating of *Meets WWC Standards With Reservations*. To be eligible for the highest rating of "positive effects," more than 50.0 percent of the meta-analytic weight needs to be attributable to studies that are rated *Meets WWC Standards Without Reservations*. These procedures and rules apply regardless of whether the studies in the intervention report are group design studies, RDD studies, SCD studies, or a combination of the three.

[6] The regulatory definitions of the evidence tiers are themselves based on the statutory definitions of "strong," "moderate" and "promising" evidence included in the *Every Student Succeeds Act* (ESSA), P.L. 114-95, and 20 U.S.C. §7801(21).

Table IV.3. What Works Clearinghouse characterization of findings in intervention reports

Characterization	Criteria
Positive effects	• At least two studies are rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*; AND • The mean effect from a fixed-effects meta-analysis of these studies is statistically significant and positive; AND • More than 50.0 percent of the fixed-effects meta-analytic weight comes from studies that are rated *Meets WWC Standards Without Reservations.*
Potentially positive effects	• At least two studies are rated *Meets WWC Standards Without Reservations* or *Meet WWC Standards With Reservations*; AND • The mean effect from a fixed-effects meta-analysis of these studies is statistically significant and positive; AND • 50.0 percent or less of the fixed-effects meta-analytic weight comes from studies that are rated *Meets WWC Standards Without Reservations*. OR • One study is rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations;* AND • The study has a statistically significant and positive effect.
Uncertain effects	• At least two studies are rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*; AND • The mean effect from a fixed-effects meta-analysis of these studies is not statistically significant. OR • One study is rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*; AND • The study does not have a statistically significant effect.
Potentially negative effects	• At least two studies are rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*; AND • The mean effect from a fixed-effects meta-analysis of these studies is statistically significant and negative; AND • 50.0 percent or less of the fixed-effects meta-analytic weight comes from studies that are rated *Meet WWC Standards Without Reservations.* OR • One study is rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*; AND • The study has a statistically significant and negative effect.
Negative effects	• At least two studies are rated *Meet WWC Standards Without Reservations* or *Meet WWC Standards With Reservations*; AND • The mean effect from a fixed-effects meta-analysis of these studies is statistically significant and negative; AND • More than 50.0 percent of the fixed-effects meta-analytic weight comes from studies that are rated *Meets WWC Standards Without Reservations*.

Note: A fixed-effects meta-analytic estimate is the WWC's default method of synthesis across studies. If the WWC reports meta-analytic estimates using other methods of synthesis, then the WWC will document these methods in a supplement to this *Procedures Handbook.*

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 What Works Clearinghouse extent of evidence for an intervention

Extent of Evidence	Criteria
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 individuals.
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 individuals.

- 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 increase the sensitivity of the analysis by reducing the effects of estimation 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, such as one 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 individuals was selected because it is the smallest sample size needed to have an 80 percent probability of detecting an impact of 0.30 standard deviation or larger as statistically significant at the .05 level for a simple RCT with equally sized intervention and comparison groups and no covariates used in the data analysis.

3. Summary of evidence for a What Works Clearinghouse practice guide

A practice guide is a publication that presents recommendations from across the empirical literature to help educators address particular challenges in their classrooms and schools. In contrast with an intervention report, a practice guide does not focus on characterizing evidence of effectiveness for individual "branded" programs or products, but on identifying a set of intervention components that, when implemented appropriately, may improve student outcomes or other outcomes relevant for education. Each guide is based on a systematic review of studies by the WWC, on practitioner experience, and on the opinions of a panel of nationally recognized experts.

When assessing the evidence for each practice recommendation, the expert panel and WWC review staff consider the following:

- The extent of evidence meeting WWC standards, as defined in table IV.4.
- The effectiveness ratings, as defined in table IV.3, for the relevant outcome domain(s) based on the studies that both meet WWC standards and inform that recommendation.
- The relevance of the studies for representing 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.
- Panel confidence in the effectiveness of the recommended practice.

Practice guide panels and WWC staff rely on a set of criteria to determine the level of evidence supporting each practice guide recommendations (table IV.5).

Table IV.5. Levels of evidence for practice guide recommendations

Requirement	Strong evidence base	Moderate evidence base	Minimal evidence base
Extent of evidence	The research includes studies that meet WWC standards and provide a "medium to large" extent of evidence as defined in table IV.4.	The research includes at least one study that meets WWC standards and provides a "small" extent of evidence as defined in table IV.4.	The research does not include evidence from studies that meet the requirements for moderate or strong evidence.
Effects on relevant outcomes	The research shows, for the relevant outcome domain(s), a preponderance of evidence of "positive effects" as defined in table IV.3, without contradictory evidence of "negative effects" or "potentially negative effects."	The research shows, for the relevant outcome domain(s), a preponderance of evidence of "positive effects" or "potentially positive effects" as defined in table IV.3. Contradictory evidence of "negative effects" or "potentially negative effects" 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, uncertain, 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.

Requirement	Strong evidence base	Moderate evidence base	Minimal evidence base
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, but 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 on which the recommendation is focused.	Not applicable.

Note: A recommendation must satisfy *all* applicable requirements in the same column for the WWC to characterize the practice as supported by an evidence base at that level.

Appendix A. Policies for prioritizing studies for review

Because of the large amount of research literature in the field of education, the What Works Clearinghouse (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 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 topic areas determined by the Institute of Education Sciences (IES) to be of highest priority for informing education decisions. 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 education associations; input from state and federal policymakers; patterns of searches for education topics on the WWC website or on the Internet more generally; and scans of the literature or of research funded by the U.S. Department of Education to determine how much evidence on the effectiveness of interventions exists in various topic areas. In consultation with the WWC contractors participating in the corresponding reviews, IES determines the topic areas within which the WWC will conduct intervention reviews. To date, topic areas for intervention reports 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 may conduct 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 content experts, 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 scoring system specified in the study review protocol being used for interventions in the corresponding topic area.

B. Prioritizing topics for practice guides

Practice guide topics are selected based on their potential to improve key 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 individual studies for review

Reviews of individual studies are generally initiated in two ways: IES requests a WWC review of a particular study or a study is prioritized from a list of eligible studies receiving significant media attention, studies submitted to the WWC Help Desk, and other studies not yet reviewed by the WWC.

First, IES may request that one of the WWC contractors complete a review of a specific study for a variety of reasons. For example, IES may request a review of a publicly available study funded by the Department, or of a publicly available study that has been cited as evidence for a discretionary grant competition.

A second method by which studies are selected for WWC review is through "prioritization" lists of eligible studies not currently under review for WWC intervention reports or practice guides or at the specific request of IES. The prioritization lists include studies receiving significant media attention, studies submitted through the WWC Help Desk, studies funded by the Department, studies screened but not reviewed for WWC publications, and studies assessed by WWC-certified reviewers but not yet included in the official WWC database. "Significant media attention" means the study was recently released and reported on in a major national news source or a major education news publication.

Appendix B. Principles for searching for studies to review

Some What Works Clearinghouse (WWC) products, including intervention reports and practice guides, are the result of systematic, comprehensive searches of the literature. Review teams should employ research librarians to help design the search strategy. In addition, review team leadership should provide training to staff so that they reliably identify and screen potentially relevant literature. The review team should write WWC review protocols to include specified search terms, string, limiters, and all necessary and relevant search databases and auxiliary search procedures (for example, specific websites and reference harvesting). Protocols should include relevant terms from the ERIC Thesaurus or complementary databases and include search terms related to the intervention, population, outcomes, and study designs. Example search terms are given in table B.1.

Table B.1. Example search terms for WWC literature searches

Category	Example search term
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 (*) is a Boolean operator and allows the truncation of the term so that the search returns any word that begins with the specified letters. Boolean operators vary across online databases; review teams should consult the specified online database to ensure accurate usage.

Review teams should use the freely available, public version of ERIC as the initial source of studies for WWC reviews. The public version of ERIC is an up-to-date index of education research and gray literature available. Protocols should be written in way that requires review teams to search using best practices of searching for systematic reviews (see this webinar). Protocol authors should consider a wide range of multidisciplinary databases (for example, Academic Search Premier, ProQuest Dissertations and Theses, PsycInfo, Education Research Complete, and/or EconLit) that complement ERIC.

Protocol authors should consider including specific websites or sources of gray literature (for example, research firms, government agencies, or nonprofit organizations) to be searched that are not indexed in ERIC; the entire ERIC index can be found here. Protocol authors should also consider conducting forward and backward reference harvesting of eligible studies. The specific searches conducted may vary across protocols but should include enough detail and scale to ensure that all relevant sources are searched and all identifiable research found.

Appendix C. Staffing, reviewer certification, and quality assurance

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

A. Staffing for What Works Clearinghouse 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 Institute of Education Sciences (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 appendix C, section B, "Reviewer certification," 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 at least six members. Each panel is chaired by a nationally recognized researcher with expertise in the topic. The panel consists of at least 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 review protocol, review studies, and draft the guide. There are four primary roles for WWC research staff on practice guide review teams: an evidence coordinator, who ensures that the research used to support recommendations is rigorous and relevant; a practice coordinator, who ensures that the discussion of how to implement each recommendation is concrete, specific, and appropriate;

WWC-certified reviewers, who assess whether supporting literature meets WWC standards; and 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.

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 are completed using a set of video modules on the WWC website that include an overview of the WWC and its products and in-depth instruction on the WWC review standards, review tools, policies, and practices (https://ies.ed.gov/ncee/wwc/OnlineTraining). Trainings for RDDs and SCDs are each one day and are conducted in person.

At the conclusion of training, participants pursuing certification are expected to take and pass a certification examination including several multiple-choice questions and examples from studies reviewed by the WWC. The exam is graded by the certification team, with feedback provided to the participant. If the participant does not satisfactorily complete the exam, then he or she will have two more opportunities to receive certification.

Upon the release of updated *WWC Procedures Handbook, Version 4.1,* and *WWC Standards Handbook, Version 4.1*, certified reviewers are required to view a recertification video and answer exam items 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. Review of draft What Works Clearinghouse publications

At each stage, reviewers examine the accuracy of WWC study reviews, evaluate draft WWC publications for consistency and clarity, and ensure that a draft publication conforms to WWC processes. It is only after intense review from several perspectives that a WWC publication such as an intervention report or practice guide is released to the public.

After an extensive drafting and revision process with multiple layers of internal review, the completed draft of each WWC publication is submitted to IES, which reviews the document internally and sends it for peer review by researchers who are knowledgeable about WWC standards and are not staff with the WWC contractor that prepared the draft publication. Both

sets of comments are returned to the contractor's drafting team, which responds to each comment and documents all responses in a memo. Each intervention report undergoes a final review by IES staff to ensure that any issues have been addressed appropriately. Practice guides also undergo an external peer review process through IES's Standards and Review Office (https://ies.ed.gov/director/sro/about.asp).

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 be submitted in writing to the WWC Help Desk through the Contact Us page (https://ies.ed.gov/ncee/wwc/ContactUs.aspx), pertain to a specific study or set of studies, and 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 policies and procedures are intended to avoid conflicts of interest and promote transparency in the review process.

Appendix D. Examples of study definition

When two findings share at least three of the following characteristics, the What Works Clearinghouse (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 members 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 grade 3 students after one year (cohort 1) and after two 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 fourth 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 replication studies by the same authors. After releasing a report with findings from a randomized controlled trial (RCT), study authors conduct a replication analysis using the same group formation and analysis procedures on a distinct sample: students in different schools and districts. The background characteristics used in the replication analysis differed from those in the original analysis because of differences in administrative data collection. Additionally, the authors introduced a new data collection procedure designed to limit sample attrition. The WWC would consider the replication analysis to be a separate study from the original analysis because the two sets of findings share neither the same sample members nor the same data collection procedures. If the only difference in the data collection procedures had been the background characteristics, the review team could exercise discretion and determine whether the difference is significant enough to consider these separate

studies. For example, if the characteristics are specified in the review protocol as required for baseline equivalence, then how they are collected and measured may be significant.

Example 9: Findings from related samples, based on different designs. Study authors randomly assigned students to a condition and conducted an RCT analysis. Using a subsample of the randomly assigned students, the same authors also examined a quasi-experimental design (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 10: Findings reported for multiple contrasts. If authors compare an intervention group with 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 (that is, 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 main findings. However, if a contrast is between two versions of the intervention, then the findings should be presented as supplementary.

Appendix E. Magnitude of findings and accompanying standard errors

The results of analyses can be presented in a number of ways, with varying amounts of comparability and utility. To the extent possible, the What Works Clearinghouse (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, and the formulas for computing the standard error of the effect size.

A. Effect sizes

To assist in the interpretation of study findings and facilitate comparisons of findings across studies, the WWC computes the effect size and standard error 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 improves the comparability of effect size estimates across studies. 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 postintervention means adjusted for the preintervention measure. If a study reports both unadjusted and adjusted postintervention means, then the WWC reports the adjusted means and unadjusted standard deviations.

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:

[E.1.0] $= \frac{}{}$, and

[E.1.1] $= \frac{\sqrt{(\quad)\quad(\quad)}}{}$,

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

[E.1.2] $= \frac{}{(\quad)}$.

The effect size index thus computed is referred to as Hedges' *g*; note that it is very similar to the well-known Cohen's *d* effect size. The standardized mean difference effect size, 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. The correction involves multiplying Hedges' *g* by a factor of $= 1 \quad \frac{}{}$, where *N* is 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.

[E.1.3] $= \frac{(\quad)}{(\quad)}$.

The default standard error calculation for the Hedges' *g* effect size measure corrected for small-sample bias is given by (Borenstein & Hedges, 2019):

[E.1.4] $(\ ) = \sqrt{\frac{}{} + \frac{}{(\quad)}}$.

In certain situations, 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, then 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 , 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. Deviations from the default will be clearly described and justified in the WWC's review documentation for that study.

Effect sizes from student-level *t* tests or ANOVA. For 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, then 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,

[E.1.5] = .

For effect sizes based on ANOVA *F* test results,

[E.1.6] = ( ),

where the sign is determined by the sign of the main difference.

Effect sizes from an analysis of covariance. An analysis of covariance (ANCOVA) is a commonly used analytic method for 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 a student-level ANCOVA, the WWC computes Hedges' *g* as the covariate-adjusted mean difference divided by the unadjusted pooled within-group SD:

[E.2.0] = ( ),

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 directly from 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 pretest and posttest *r* is known, and the pretest is the only measure controlled for in the ANCOVA analysis, then we can derive Hedges' *g* from the ANCOVA *F* statistic as follows:

[E.2.1] = ( )( ).

The standard error calculation for the ANCOVA-based effect size is given by

[E.2.2] $$( ) = \sqrt{\frac{}{}(1 \quad) + \frac{}{(\quad)}},$$

where R^2 is the multiple correlation between the covariates and the outcome. To compute equation E.2.2, the WWC will use the value of R^2 provided in the study report or from an author query. If R^2 is not available, the then WWC will take a cautious approach to calculating the standard error and assume a value of zero for R^2. This cautious approach will overestimate the magnitude of the standard error but protects against type I error.

Difference-in-differences adjustment. Study authors will occasionally report unadjusted group means on both pretest and posttest but not adjusted group means and adjusted group mean differences on the posttest. If the pretest and posttest are based on the same test, then the WWC computes the effect size of the difference between the two groups using the gain score effect size formula in Morris (2008) and the pooled posttest SD:

[E.3.0] $$= \frac{\quad, \quad, \quad, \quad,}{\quad, (\quad) \quad,},$$

where $\quad,$ and $\quad,$ are the posttest and pretest means, respectively, for the intervention group and $\quad,$ and $\quad,$ are the posttest and pretest means, respectively, for the comparison group.

The standard error calculation for the gain score effect size formula includes the population correlation between the pretest and posttest measures, :

[E.3.1] $$( ) = \sqrt{\frac{}{}2(1 \quad) + \frac{}{(\quad)}}.$$

If the pretest and posttest are not based on the same test, then 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:

[E.3.2] $$= \quad (\quad).$$

The standard error calculation for this difference-in-differences effect size calculation is given by:

[E.3.3] $$( ) = \sqrt{\frac{}{}2(1 \quad) + \frac{}{(\quad)}}.$$

The standard error calculations for both difference-in-differences approaches (equations E.3.1 and E.3.3) require the correlation between the baseline and outcome measures. For equations E.3.1, E.3.2, and E.3.3, the WWC will use the sample correlation between the baseline and outcome measures if provided in the study report or, if not available from the study, from an author query. For equations E.3.1 and E.3.3, if the correlation is not available, the WWC will

take a cautious approach to calculating the standard error and assume a value of .5 for .[7] For equation E.3.2, if that correlation is not available, then the WWC will take the cautious approach to estimating the effect size and assume a value of 1 for . The lead methodologist may choose to use a different value for if dependable empirical data on the relationship between the baseline and outcome measures are available. 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.[8]

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

[E.4.0] $= \frac{}{(\quad)}$,

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:

[E.4.1] $= \frac{(\quad)}{(\quad)}$,

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 OR 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, such as school-level means, 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:

[7] Using a value of .50 for , the variance becomes proportional to the variance for a posttest only mean difference.

[8] Future updates to the *WWC Procedures Handbook* may include empirical values of the correlation.

[E.4.2a] = ( ).

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

[E.4.2b] = ( ) ( ),

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

The LOR has an important relation to the standardized mean difference. The WWC has adopted the Cox index as the default effect size measure for dichotomous outcomes when these are being synthesized with continuous outcomes (Sanchez-Meca et al., 2003). The computation of the Cox index is straightforward:

[E.4.3] = —.

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) noted, 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.

The standard error calculation for the Cox index effect size is given by (Sanchez-Meca et al., 2003):

[E.4.4] ( ) = — — + ——— + — + ———.

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* (that is, the Cox index in equation E.4.3), with the final effect size given by their difference:

[E.4.5] = ( ).

To date, there is no clear guidance for how to calculate the standard error for the dichotomous outcomes difference-in-differences adjustment. As a result, the WWC will not estimate the standard error of a difference-in-difference effect size with dichotomous outcomes.

Gain scores

Some studies report only the means and SD 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 SD of the outcome measure after adjusting for baseline measures are not comparable with effect sizes calculated using SDs of unadjusted posttest scores. The effect size based on the gain score SDs will generally be larger because the standard deviation of gain scores is typically smaller than the SD of unadjusted posttest scores. The WWC will not report effect sizes based on the gain score SDs, 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, for example when schools or teachers are assigned to conditions, but 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.

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.

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 because the intraclass correlation coefficient (ICC) yields cluster-level SDs that are typically much smaller than student-level SDs,

[E.5.0] $= \overline{\text{ICC}},$

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.

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 in the *cluster-level effect sizes* section on the previous page, 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; Snijders & Bosker, 1999). If the outcome is a

standardized measure that has been administered to a norming sample at the national or state level, then the effect size may be calculated using the SD from the norming sample.

Student-level effect sizes from multilevel modeling

With recent methodological advances, multilevel analysis has gained increased popularity in education and other social science fields. Researchers have begun to employ the hierarchical linear modeling (HLM) method to analyze data of a nested nature, such as 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 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, for example, 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 (Borenstein & Hedges, 2019): adjusted group mean difference divided by unadjusted pooled within-group SD. Specifically,

[E.5.1] $= \frac{}{(\quad)}$,

where 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 of students nested with teachers or classrooms. The idea could easily be extended to a three-level model, for example, students nested with teachers who were, in turn, nested within schools.

The standard error for the effect size of a cluster-level assignment design can be computed with the following calculation for a two-level design (Borenstein & Hedges, 2019):

[E.5.2] $( ) = \sqrt{\frac{}{} (1 + (\quad 1)) + \frac{(\quad)(\quad) \quad (\quad) \quad (\quad) (\quad)}{(\quad)[(\quad) \quad (\quad)]}}$,

where N is the total student-level sample size, N_i and N_c are the total number of students in the intervention and comparison groups, respectively; n is cluster sample size when cluster sample sizes are equal; and is the ICC. While equation E.5.2 can be expanded to account for situations where cluster sample sizes are not equal, the WWC will use equation E.5.2 to calculate standard errors for studies with equal and unequal cluster sample sizes.

3. Design-comparable effect sizes from single-case designs

As outlined in section VI.A, the WWC reports the results from SCDs as a design-comparable effect size (D-CES). A D-CES can be computed for a study that has three or more participants in a design that is multiple baseline across individuals, multiple probe across individuals, or a treatment reversal design. Shadish, Hedges, and Pustejovsky (2014) provided a formula to compute the effect size = – for the treatment reversal design where:

[E.6.0] = — – ( ) – ( ) ( ) ,

where is the observation of case *i* at time *j* in phase pair *a*, *m* is the number of cases, *n* is the number of timepoints per phase, and *k* is the number of AB phase pairs.

[E.6.1] = ( ) ( ) . ,

where . is the mean across individuals at the *t*th time point given by:

[E.6.2] . = – .

The D-CES for the multiple baseline (across individuals) and multiple probe (across individuals) designs is also defined as = – but where:

[E.6.3] = – (. .),

where . and . are the average outcomes for individual *i* within the intervention and baseline conditions, respectively, and:

[E.6.4] = —— , . ,

where *N* is the total number of timepoints, *K* is a degrees-of-freedom correction, and indicates which cases are in condition *p* at time point *j*, for *j* = 1,…, *N* and *p* = *B* for baseline, *T* for treatment. Finally, the WWC applies the small-sample correction and estimates the standard error of the small-sample corrected D-CES following equations 7 and 8, respectively, in Shadish, Hedges, and Pustejovsky (2014).

4. When student-level effect sizes cannot be computed

In some cases, the WWC will be unable to calculate an appropriate effect size from the data reported by the study authors that can be compared with 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 reports or practice guides, 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 and practice guides and potentially inform the discussion in those publications. 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 and the assessment of statistical significance.

B. Improvement index

To help readers judge the practical importance of an intervention's effect, the WWC may translate 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 (that is, 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 percent (10 percent + 50 percent = 60 percent) of the students in the intervention group scored above the comparison group mean. Specifically, the improvement index is computed as described next.

1. Step 1. Convert the effect size (Hedges' *g*) to Cohen's U_3 index

The U_3 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 U_3 of 60 percent, 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 U_3 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, U_3 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.

2. Step 2. Compute improvement index = U_3 – 50 percent

Given that U_3 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 percent, the improvement index, defined as U_3 – 50 percent, 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.

Note that the estimate of U_3 described above is not an unbiased estimate of the tail area (the U_3 parameter) and so the improvement index is also not an unbiased estimate. Nevertheless, the bias is usually small for studies with sample sizes likely to be found in WWC reviews (see Hedges & Olkin, 2016).

Appendix F. Statistical significance for randomized controlled trials and quasi-experimental designs

In order to assess the effects of an intervention adequately, 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, such as the 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.

To present a fair judgment about an intervention's effects, the What Works Clearinghouse (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 intraclass 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.

1. Step 1. Compute the *t* statistic for the effect size, ignoring clustering

[F.1.0] $= \frac{}{\sqrt{\quad + \quad}}$,

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.

2. Step 2. Correct the *t* statistic for clustering

[F.1.1] $= \frac{(\quad) \quad -}{(\quad) \quad -}$,

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 is the intraclass correlation coefficient (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 .20 for achievement outcomes and .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.

3. Step 3. Compute the degrees of freedom associated with the *t* statistic corrected for clustering

[F.1.2] $= \frac{(\quad) \quad -}{(\quad)(\quad) \quad - \quad - \quad - (\quad)}$

4. Step 4. Obtain the statistical significance of the effect corrected for clustering

The clustering-corrected statistical 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). If the cluster-corrected p value from a two-tailed t test is less than .05, then the effect is statistically significant.

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 $1/m$, where m is 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 instead of family-wise error rate. 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 false-discovery-rate-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) pointed 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 two types of situations: studies that tested multiple outcome measures in the same outcome domain with a single comparison group, and studies that tested one or more outcome measures 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

[F.2.0] $\quad < \quad < \quad < \quad < \quad ,$

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:

[F.2.1] $= \frac{x\alpha}{M}$,

where x is the rank for p_x, with $x = 1, 2, \ldots, m$; 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: The authors may not have reported findings from the complete set of comparisons they had made and 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 employs a type I error rate of $\alpha = .05$ when implementing this correction.

Step 3. Identify the cutoff point

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

[F.2.2] .

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 a 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 it 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 F.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 $\alpha = .05$ 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, $y = 5$, $p_5 = .030$, and $p'_5 = .031$. 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 (.027) lower than the highest p value (.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 ($p'_x = .05x/8$)	Finding *p* value ≤ new critical *p* value? ($p_x \leq p'_x$)	Statistical significance after BH correction?
.002	1	.006	Yes	Yes
.009	2	.013	Yes	Yes
.014	3	.019	Yes	Yes
.027	4	.025	No	Yes
.030	5	.031	Yes	Yes
.042	6	.038	No	No
.052	7	.044	No	No
.076	8	.050	No	No

BH is Benjamini-Hochberg.

Thus, we can claim that the five findings associated with a *p* value of .030 or smaller are statistically significant at the .05 level after correction for multiple comparisons. The sixth finding (*p* value = .042), although reported as being statistically significant, is no longer significant after the correction.

2. One or more outcome measures 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 including all comparisons they consider relevant, calculating separate effect sizes for each comparison, and averaging these findings together in a manner similar to multiple outcomes in a domain, as discussed above. 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.

Appendix G. Reporting requirements for studies that present a complier average causal effect

A. Reporting of complier average causal effects estimates in What Works Clearinghouse products

Among randomized controlled trials (RCTs), any complier average causal effects (CACE) estimate that addresses a research topic relevant to a What Works Clearinghouse (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 ITT estimates, as follows.

RCT studies that report both an intent to treat (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 an intervention report or practice guide, then only one of the two types of estimates will contribute to the effectiveness 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 decisionmakers. 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 effectiveness ratings or levels of evidence determinations.

B. Reporting requirements for estimated variances of complier average causal effects 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, described by 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, then the WWC will not use the reported statistical significance of the CACE estimate in characterizing the study's findings.

Appendix H. Estimating the fixed-effects meta-analytic average in intervention reports

A. Estimating the fixed-effects meta-analytic average in intervention reports and practice guides

What Works Clearinghouse (WWC) intervention reports and practice guides are systematic reviews of educational products, policies, practices, or curricula. These reports synthesize studies that meet WWC standards. When more than one study in an intervention report or practice guide estimates an effect size in the same outcome domain, the WWC estimates a fixed-effects meta-analytic average. The WWC chose the fixed-effects model because its goal is to make inferences about the studies in WWC intervention reports and practice guides. Unlike the fixed-effect (singular) model, the fixed-effects (plural) model does not assume that the studies are estimating a common effect. Instead, the fixed-effects model assumes that the observed variation among the effect sizes in the meta-analysis reflects the true variation in population effects. Accordingly, inferences to larger study populations are constrained to those that share the same patterns of important study characteristics that are related to effect size.

Most meta-analyses involve weighting the studies by some value. Although a number of different weighting procedures have been proposed, the most popular weighting scheme involves using weights that correspond approximately to sample size, with larger studies receiving more weight in the analysis. For example, a simple randomized experiment with 300 students will have approximately three times the weight of a simple randomized experiment with 100 students. This is similar to how grade point averages in college are computed: A grade earned in a three-credit-hour course will have three times the weight of a grade earned in a one-credit-hour course. Formally, effect sizes are weighted by the inverse of their variances; hence, this procedure is known as inverse variance weighting.

Appendix E provides the formula for Hedges' *g*, a common effect size for continuous outcomes. Appendix E also includes formulas for each effect size's standard error. Each standard error may be converted to a variance following:

[H.1.0] = ,

where is a standard error estimated using a formula.

Given the , we next estimate the weight associated with each effect size:

[H.1.1] = —,

where is the weight for the effect size in study *s* and is the variance of *g* for study *s*.

The effect size and effect size weight are all that are needed to estimate the fixed-effects meta-analytic average, defined as:

[H.1.2] = ————,

where is the fixed-effects meta-analytic average and *k* is the total number of studies. The numerator sums the product of each study's weight by each study's effect size and the denominator sums each study's weight.

We estimate the standard error of the fixed-effects average by:

[H.1.3] = ——— ,

where is the standard error of the fixed-effects meta-analytic average. A statistically significant estimate of an effect is one for which the null hypothesis was evaluated and rejected using a nondirectional *z* test and a type I error rate of $\alpha = .05$.

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.

Borenstein, M. & Hedges, L. V. (2019). Effect sizes for meta-analysis. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (3rd ed., pp. 207–244). New York, NY: Russell Sage Foundation.

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

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.

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

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. (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.

Hedges, L. V., & Olkin, I. (2016). Overlap between treatment and control group distributions of an experiment as an effect size measure. *Psychological Methods*, *21*, 61–68.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. A. (2012). A standardized mean difference effect size for single case designs. *Journal of Research Synthesis Methods, 3*, 224–239.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. A. (2013). A standardized mean difference effect size for multiple baseline designs. *Journal of Research Synthesis Methods, 4,* 324–341.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods, 3*(4), 486–504.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Morris, S. B. (2008). Estimating effect sizes from pretest-posttest-control group designs. *Organizational Research Methods, 11*, 364–386.

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. L. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics*, *39*, 368–393.

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.

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.

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

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*(2), 123–147.

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

Thompson, C. G., & Becker, B. J. (2014), The impact of multiple endpoint dependency on Q and I^2 in meta-analysis. *Research Synthesis Methods*, 5, 235–253.

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.

January 2020

This report was prepared for the Institute of Education Sciences (IES) under Contract 91990018C0019 by the American Institutes for Research. The mention of trade names, commercial products, or organizations does not imply endorsement by the U.S. Government.

What Works Clearinghouse. (2020). What Works Clearinghouse Procedures Handbook, Version 4.1. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. This report is available on the What Works Clearinghouse website at https://ies.ed.gov/ncee/wwc/handbooks.