What Works Clearinghouse

Procedures and Standards Handbook
(Version 2.1)

The randomized controlled trial and comparison group quasi-experimental design standards published in 2008 are reproduced in this document, along with the pilot standards for regression discontinuity designs and single-case designs that were completed in 2010.

TABLE OF CONTENTS

Chapter **Page**

Chapter **Page**

FOREWORD

The What Works Clearinghouse (WWC) is part of the U.S. Department of Education's Institute of Education Sciences (IES), which was established under the Education Sciences Reform Act of 2002. With its critical assessments of scientific evidence on the effectiveness of education programs, policies, and practices (referred to as "interventions"), and a range of products summarizing this evidence, the WWC 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.

It is critical that educators have access to the best evidence about the effectiveness of education programs, policies and practices in order to make sound decisions. But, without a service like the What Works Clearinghouse, it can be difficult, time-consuming, and costly for educators to access the relevant studies and reach sound conclusions about the effectiveness of particular interventions. The WWC meets this needs for credible, succinct information by reviewing existing research studies, assessing the quality of the research, summarizing the evidence of effectiveness on student-related outcomes, and disseminating its findings broadly. The WWC's mission is to assess the quality and findings of existing research; thus, it does not conduct original research on education programs, policies, or practices.

The WWC only reports impacts for studies for which it has high or moderate confidence that the effect can be attributed solely to the intervention rather than to the many other factors that are at play in schools and in the lives of students. Moreover, when the WWC undertakes a review of a particular intervention or issue, it conducts a thorough search for all relevant literature meeting the WWC evidence standards. For example, educators who want to know whether a particular intervention is effective can read a WWC Intervention Report and know that it represents both a thorough review of the identified research literature on that intervention and a critical assessment and summary of the evidence reported by the study authors.

The What Works Clearinghouse uses objective and transparent standards and procedures to make its assessment of the scientific merit of studies of the effectiveness of education interventions, and then summarizes the results of its systematic reviews in a set of products that currently includes:

- *Intervention reports.* These reports summarize all studies published during a specific time period that examine the effectiveness of an intervention. Using its objective and transparent standards, the WWC rates the quality of the evidence in each study. For studies that meet WWC evidence standards (with or without reservations), the WWC combines the findings to generate overall estimates of the *size of effects* for the intervention. The WWC also rates the *level of evidence* on the intervention's effectiveness, taking into consideration the number of studies, the sample sizes, and the magnitude and statistical significance of the estimates of effectiveness.

- *Practice guides.* These guides contain practical recommendations that educators can use to address specific challenges in their classrooms and schools. The recommendations are based on reviews of research as well as the expertise and professional judgments of a panel of nationally recognized experts that includes both researchers and educators.

- *Quick Reviews.* These reports are designed to provide education practitioners and policymakers with timely and objective assessments of the quality of the research evidence from recently released research papers and reports whose public release is reported in a major national news source. These reviews focus on studies of the effectiveness of education or school-based interventions serving students in pre-kindergarten through post-secondary settings.

The systematic review is the basis of all What Works Clearinghouse products. Systematic reviews use explicit methods to identify, select, and critically appraise relevant research, and to extract and analyze data from studies. There are five basic steps in the WWC review process:

1. **Develop a review protocol.** Protocols define the scope of studies that will be reviewed, the process through which studies will be identified, and the outcomes that will be examined. Protocols also specify the time period during which relevant studies will have been conducted, the outcomes to be examined in the review, and keyword strategies for the literature search.

2. **Identify relevant studies,** often through a systematic search of the literature.

3. **Screen studies** for relevance and the adequacy of study design, implementation, and reporting.

4. **Retrieve and summarize** information on the intervention studied, the study characteristics, and the study findings.

5. **Combine findings** within studies and across studies when relevant.

In this version of the Handbook, pilot standards for judging the conditions under which studies using regression discontinuity or single-case designs meet WWC standards for causal validity have been added. As the WWC continues to refine processes, develop new standards, and create new products, the Handbook will be revised or augmented to reflect these changes. Readers who want to provide feedback on the Handbook, or the WWC more generally, may contact the WWC Help Desk at http://ies.ed.gov/ncee/wwc/help/webmail.

I. CONTRIBUTORS TO TOPIC AREA REVIEW

A large number of people are involved in conducting a review for the WWC. Although the Topic Area Team is directly responsible for the content of the review, team members are aided by many others outside the team. This chapter describes the roles of those who contribute to the topic area reviews, along with details on participating organizations and conflicts of interest.

A. WWC ORGANIZATIONS

The WWC is administered by the U.S. Department of Education's Institute of Education Sciences through a contract with Mathematica Policy Research, Inc. (MPR), a nationally recognized leader in education research and in rigorous reviews of scientific evidence. Experts and staff from a variety of organizations participate in the development of WWC topic areas and reports. Subcontractors that may also be involved include Analytica; Chesapeake Research Associates; Communications Development, Inc.; CommunicationWorks; Empirical Education, Inc.; ICF-Caliber; Optimal Solutions Group; RAND Corporation; RG Research Group; SRI International; Twin Peaks Partners; the University of Arkansas; and the University of Wisconsin. For more information about key staff and principal investigators, visit the About Us page of the website (http://ies.ed.gov/ncee/wwc/aboutus).

B. TOPIC AREA TEAM

Once a topic area is selected, the WWC identifies leaders of the Topic Area Team. Each review team consists of a principal investigator (PI), deputy principal investigator (Deputy PI), content expert, project coordinator (PC), and reviewers. All Topic Area Team leaders (PI, Deputy PI, and content expert) are approved to serve in their positions by the IES.

1. Principal Investigator

The principal investigator is an expert in the research methodology of the topic area. Initially, the PI works with the deputy principal investigator to develop a review protocol for the topic area that defines the scope of the review, specifies the literature search parameters, summarizes the search results, and suggests prioritization of interventions for review. Throughout the topic area review, the PI 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 the IES.

2. Deputy Principal Investigator

The deputy principal investigator is an established researcher with relevant methodological and substantive expertise in the topic area. The Deputy PI oversees the day-to-day work of the review team, assists in the development of the review protocol, and reviews research ratings. The

Deputy PI also reconciles differences between reviewers of a particular study, along with writing and reviewing reports on interventions.

3. Content Expert

The content expert, a well-established researcher with substantive expertise in the topic area, serves as a consultant to a Topic Area Team to help the PI and Deputy PI with content-specific questions that arise in reviews.

4. Project Coordinator

Coordinators are WWC staff with an interest in the topic area whose role is to support PIs, Deputy PIs, reviewers, and other Topic Area Team members. These individuals are responsible for coordinating the literature search process, conducting screens of the literature, organizing and maintaining the topic area's communication and management, tracking the review process, and managing the production process.

5. Reviewers

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. As part of the team, these individuals review, analyze, and summarize relevant literature for evidence of effectiveness, and also draft intervention reports.

Each reviewer must complete an extensive training and certification process before working on WWC reviews and authoring intervention reports. Potential reviewers, who are employees of MPR or WWC subcontractors, submit their resumes to WWC training and certification staff for screening. Those who pass the initial screening are invited to participate in reviewer training, a required two-day interactive session detailing the WWC and its products, review standards, and policies.

Within one week of the conclusion of training, participants must pass a multiple-choice certification examination. Those who pass the certification exam are required to complete a full review of an article. The review is graded by the certification team, with feedback provided to the trainee. If the trainee has not satisfactorily completed the review, he or she will be asked to review a second article, which is again graded and comments given. If the potential reviewer still has not attained a passing grade, he or she may be asked to complete a third review as long as the second review showed improvement. If there is no apparent improvement or the trainee does not adequately complete the third review, he or she will not receive certification.

Those who do complete satisfactory reviews are granted "provisional certification" status and are assigned to a Topic Area Team. Reviewers work closely with the Deputy PI and the topic area coordinator to complete reviews. Once reviewers have satisfactorily completed several WWC reviews, they are granted "final certification" status as a WWC reviewer.

C. STATISTICAL, TECHNICAL, AND ANALYSIS TEAM

The Statistical, Technical, and Analysis Team (STAT) is a group of highly-experienced researchers who are employees of MPR or WWC subcontractors. This team considers 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.

D. QUALITY REVIEW TEAM

The Quality Review Team addresses concerns about WWC reports and reviews raised by external inquiries through a quality review process. Inquiries must be submitted in writing to the WWC through the Contact Us page (http://ies.ed.gov/ncee/wwc/help/webmail), pertain to a specific study or set of studies, identify the specific issue(s) in the review that the inquirer thinks are incorrect, and provide an explanation as to why the review may be incorrect.[1] The Quality Review Team addresses the following issues regarding the application of standards:

- 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 procedures for computing effect sizes were implemented correctly.

After an inquiry is forwarded to the Quality Review Team, a team member verifies that the inquiry meets criteria for a quality review and, if so, notifies the inquirer that a review will be conducted. A reviewer is assigned to conduct an independent review of the study, examine the original review and relevant author and developer communications, notify the topic area PI of the inquiry, and interview the original reviewers. Throughout the process, all actions and conversations are documented and logged. When the process is complete, the reviewer makes a determination on the inquiry.

If the original assessment is validated, the reviewer drafts a response to the inquirer explaining the steps taken and the disposition of the review. If the inquirer's concerns are validated, the reviewer notifies the WWC project director, who subsequently notifies the IES. A revised review may be conducted at the request of the IES.

E. CONFLICTS OF INTEREST

Given the central importance of the WWC, the Department of Education's National Center for Education Evaluation and Regional Assistance (NCEERA) has established guidelines

[1] Additionally, the Contact Us web page allows users to ask questions about publications, topic areas, and evidence standards, as well as to suggest topics, interventions, or studies to be reviewed; however, these issues are not addressed by the Quality Review Team.

regarding actual or perceived conflicts of interest specific to the WWC. MPR administers this conflict of interest policy on behalf of the 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 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 project team from situations in which reports and products could be reasonably questioned, discredited, or dismissed due to apparent or actual conflicts of interest and to maintain standards for high-quality, unbiased policy research and analysis. All WWC Topic Area Team members, including the principal investigator, deputy principal investigator, content expert, 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.

For its reviews, the WWC does not exclude studies conducted or outcomes created by the developer of the product being reviewed; the WWC clearly lists authors of studies and indicates when outcomes were created by the developer. Additionally, as part of the review process, the WWC will occasionally uncover studies that have been conducted by organizations or researchers associated with the WWC. In these cases, review and reconciliation of the study are conducted by reviewers from organizations not directly connected to the research. Furthermore, the detailed processes undertaken to avoid any potential conflict are described in the intervention report. These procedures, along with explicit review guidelines, IES review, and external peer review, protect the review process from bias.

II. IDENTIFYING TOPIC AREAS, RESEARCH, AND INTERVENTIONS TO DEVELOP INTERVENTION REPORTS

Since research on education covers a wide range of topics, interventions, and outcomes, a clear protocol is used to set the parameters for locating, screening, and reviewing literature in a topic area according to WWC evidence standards. Senior WWC staff, along with the PI and the Deputy PI, develop the formal review area protocol to define the parameters for the interventions within the scope of the review, the literature search, and any area-specific applications of the evidence standards. Protocols are subject to IES approval.

A. IDENTIFYING REVIEW AREAS

The WWC seeks to review the effectiveness of interventions for a wide range of educational outcomes. Topics to be reviewed are prioritized based on their potential to improve important student outcomes; applicability to a broad range of students or to particularly important subpopulations; policy relevance and perceived demand within the education community; and likely availability of scientific studies about the effectiveness of specific, identifiable interventions.

The IES selects topics based on nominations received from the public, meetings and presentations sponsored by the WWC, suggestions presented by senior members of education associations, policymakers, and the U.S. Department of Education, and reviews of existing research. A list of current topics is available on the Topic Areas page.

B. SCOPE OF THE REVIEW

The protocol includes guidance regarding the following issues:

- **Topic area focus.** A very brief overview of the topic area, including the outcomes of interest and key questions to be addressed by the review.
- **Key definitions.** Definitions of terms and concepts that will be used frequently within a topic area, particularly the key outcomes on which the review will focus, along with the domains in which they will be classified.
- **General inclusion criteria.** Specification of the population, types of interventions, and types of research to be included in the review, including detail on timeframe, sample, study design, and outcomes.
- **Specific topic parameters.** Specification of which studies are to be considered for review and which aspects of those studies are to be examined. Considerations include characteristics of interventions, elements of intervention replicability, issues for outcome relevance and reliability, characteristics relevant to equating groups, effectiveness of the intervention across different groups and settings, preferences for

measuring post-intervention effects, identification of differential and severe overall attrition, and statistical properties important for computing effect sizes.

- **Literature search methodology.** List of the requirements for searching literature, including databases to search, parameters and keywords for the searches, and any specific instructions regarding hand searches and exploration of the gray literature. Databases typically included in the literature search are ERIC, PsychINFO, Dissertation Abstracts, Sociological Collection, Professional Development Collection, Wilson Educational Abstracts PlusText, Academic Search Premier, WorldCat, and Google Scholar. Searching gray literature typically includes public submissions, materials sent directly to the WWC website or staff, requests for research made to developers of specific interventions, prior reviews and syntheses, requests for research made via listservs, and searches of organizational websites.

The PI is responsible for assuring that the topic area protocol accurately reflects the work of the review team, as well as a comprehensive review of the topic area. The protocol may be revised and updated as needed, although all revisions must be approved by the IES.

C. LITERATURE SEARCH

Identifying and reviewing literature begins after the topic area, review protocol, and Topic Area Team leadership are approved by the IES. Studies are gathered through an extensive search of published and unpublished research literature, including submissions from intervention developers, researchers, and the public. The WWC staff use the search parameters set by the protocol to search relevant databases and store all references in the reference-tracking software for the topic area.

Trained WWC staff members use the following strategies in collecting studies:

- **Electronic databases.** Identify keywords for each topic and search a variety of electronic databases for relevant studies.
- **Website searches.** Search the websites of core and topic-relevant organizations and collect potentially relevant studies.
- **Extensive outreach.** Contact topic experts and relevant organizations to request studies as well as to request recommendations of other people and organizations that are able to provide studies.
- **Submissions.** Incorporate studies submitted by the public.

D. ELIGIBILITY SCREENING

In each area, the WWC collects published and unpublished studies that are potentially relevant to the topic. Gathered studies that meet broad relevancy and methodology criteria are then screened regarding the relevance of the intervention to the topic area, the relevance of the sample to the population of interest, the timeliness of the study, the relevance and validity of the

outcome measure, and other criteria specified in the topic area protocol. Across topic areas, three general criteria apply:

- *Was the study published in the relevant time range?* Studies need to have been published within 20 years of the beginning of the topic area review. This time frame encompasses research that adequately represents the current status of the field and of analytical methods and avoids inclusion of research conducted with populations and in contexts that may be very different from those existing today.
- *Is the study a primary analysis of the effect of an intervention?* Some research studies identified in the literature search will not be primary studies of an intervention's impacts or effectiveness, and cannot provide evidence of the effects of the intervention for the WWC review. For example, studies of how well the intervention was implemented, literature reviews, or meta-analyses are not eligible to be included in the review of an intervention.
- *Does the study have an eligible design?* The focus of the WWC is on scientifically-based evidence. Therefore, to be included in the WWC review, a study must use one of the following designs (described in the later section on evidence standards): randomized controlled trial, quasi-experimental, regression discontinuity, or single subject.

Across topic areas, specifics of studies to be included may vary. The screening for a topic area includes four criteria.

- *Is the intervention a program, product, policy, or practice with the primary focus aligned with the topic area?*
- *Does the study examine students in the age or grade range specified for the topic area?*
- *Does the study examine students in a location specified for the topic area?*
- *Does the study address at least one student outcome in a relevant domain?*

Studies that do not meet one or more of these criteria are categorized as "Does Not Meet Eligibility Screens," indicating that they are out of the scope of the review as defined by the topic area protocol. At this stage, a study is screened out if it

- Does not examine the effectiveness of an intervention.
- Is not a primary analysis of the effectiveness of an intervention.
- Does not provide enough information about its design to assess whether it meets standards.
- Does not use a comparison group.

- Does not include a student outcome.
- Does not include an outcome within a domain specified in the protocol.
- Does not occur within the time frame specified in the protocol.
- Does not examine an intervention conducted in English.
- Does not take place in the geographic area specified in the protocol.
- Does not use a sample within the age or grade range specified in the protocol.
- Does not disaggregate findings for the age or grade range specified in the protocol.
- Does not examine an intervention implemented in a way that falls within the scope of the review.

E. PRIORITIZING INTERVENTIONS FOR REVIEW

After the initial literature screen is completed, studies are screened and ranked to prioritize interventions to review for the upcoming review year. Only studies that relate to the protocol of the topic area (those that include the correct age range, achievement outcome measured, and so on) are included in the ranking process. Using information in the title and the abstract or introduction, the coordinator ranks the study based on internal validity, objectivity, size, and differential contrast. Once all studies are screened, the coordinator organizes the information by intervention, and interventions are ranked by their scores. After a prioritization of interventions for review has been approved, the WWC Library staff work to identify additional studies by conducting targeted searches on the named interventions.

Upon approval of the intervention ranking by the IES, the Topic Area Team can begin contacting intervention developers—the person or company that researched and created the intervention. At this point, the PI sends a letter notifying the developer of the WWC review. The letter provides a list of all WWC-identified citations related to the intervention, inquires if the list is complete, invites comment on the intervention description slated for use in the report, and requests that the developer sign an agreement not to release any information about the review. If developers have questions about the report or review process, they are encouraged to contact the WWC in writing.

III. THE REVIEW PROCESS AND EVIDENCE STANDARDS

The purpose of the WWC review of a study is to assess its quality using the evidence standards. The process is designed to ensure that the standards are applied correctly and that the study is represented accurately. Evidence standards for randomized controlled trials and comparison group quasi-experimental designs are described in detail in section B. Newly created pilot standards for regression discontinuity and single-case research designs are also described in this chapter (sections C and D) and in Appendices E and F. As of the publication of this Handbook, the WWC standards for regression discontinuity and single-case designs are being applied only in judging evidence based on individual studies. The WWC has not determined whether or how findings from studies using these two designs should be incorporated into syntheses of evidence.

A. THE REVIEW PROCESS

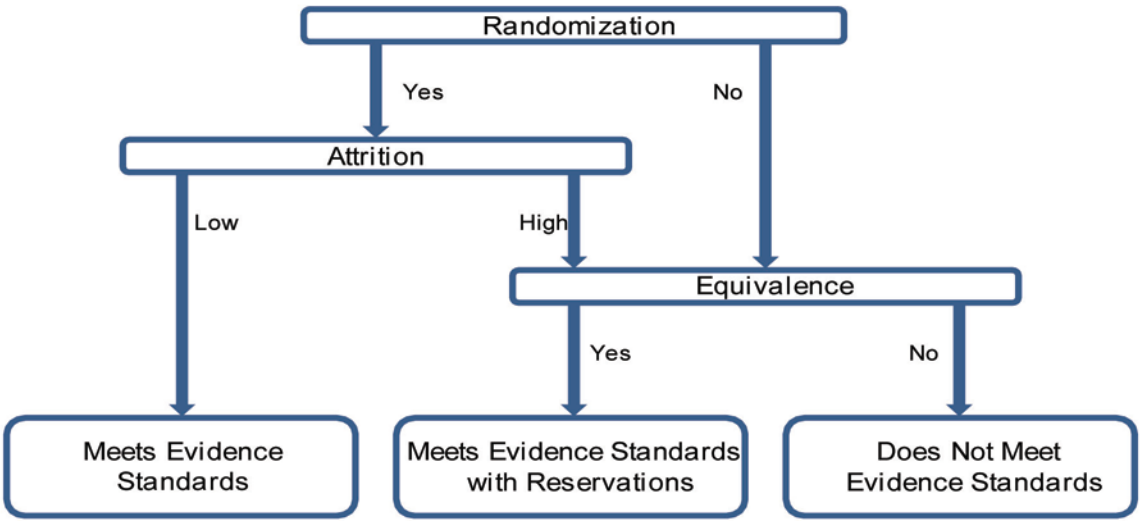
Initially, two reviewers are assigned to independently examine each study that has not been screened out as ineligible. Each reviewer completes a study review guide, which documents the study design, outcomes, samples and attrition, and analysis methods. After they complete their review, they hold a reconciliation meeting with a senior WWC reviewer to discuss any differences between their reviews and any remaining issues about the study. Following the reconciliation meeting, a master study review guide is developed to reflect the decisions of the reviewers and reconciler pertaining to the study. The review and reconciliation process typically occurs over a two-week period.

The reviews and reconciliation may result in some unresolved issues. Some of these may be technical issues regarding the application of standards, which are brought to the PI or STAT for guidance, or content issues, which may require assistance from the content expert. Others may be questions about the study itself, for which the WWC submits a query to the author. Author queries communicate a specific set of questions from the study reviewers to the study author(s), and answers to these queries clarify the questions that arose in the review. As with developer correspondence, all author queries are sent by the PI. Author responses to the query direct future review of the study, and any information provided by the author(s) is documented in the intervention report.

B. EVIDENCE STANDARDS FOR RANDOMIZED CONTROLLED TRIALS AND COMPARISON GROUP QUASI-EXPERIMENTAL DESIGNS

The WWC reviews each study that passes eligibility screens to determine whether the study provides strong evidence (*Meets Evidence Standards*), weaker evidence (*Meets Evidence Standards with Reservations*), or insufficient evidence (*Does Not Meet Evidence Standards*) for an intervention's effectiveness. Currently, only well-designed and well-implemented randomized controlled trials (RCTs) are considered strong evidence, while quasi-experimental designs (QEDs) with equating may only meet standards with reservations; evidence standards for regression discontinuity and single-case designs are under development.

A study's rating is an indication of the level of evidence provided by the study and can be affected by attrition and equivalence, in addition to study design. The following figure illustrates the contributions of these three factors in determining the rating of a study:



1. Study Design

In an RCT, researchers use random assignment to form two groups of study participants. Carried out correctly, random assignment results in groups that are similar on average in both observable and unobservable characteristics and any differences in outcomes between the two groups are due to the intervention alone, within a known degree of statistical precision. Therefore, such an RCT can receive the highest rating of *Meets Evidence Standards*.

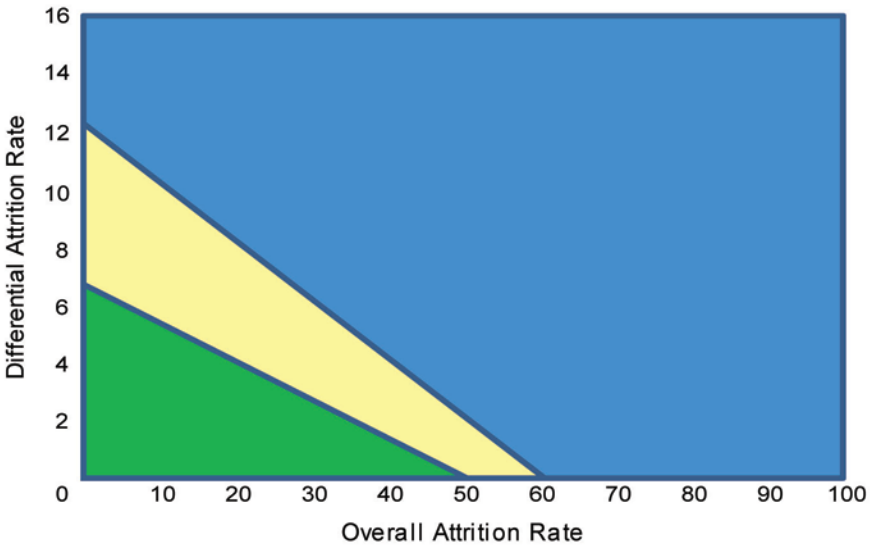
Randomization is acceptable if the study participants (students, teachers, classrooms, or schools) have been placed into each study condition through random assignment or a process that was functionally random (such as alternating by date of birth or the last digit of an identification code). Any movement or nonrandom placement of students, teachers, classrooms, or schools after random assignment jeopardizes the random assignment design of the study.

In a QED, the intervention group includes participants who were either self-selected (for example, volunteers for the intervention program) or were selected through another process, along with a comparison group of nonparticipants. Because the groups may differ, a QED must demonstrate that the intervention and comparison groups are equivalent on observable characteristics. However, even with equivalence on observable characteristics, there may be differences in unobservable characteristics; thus, the highest rating a well-implemented QED can receive is *Meets Evidence Standards with Reservations*.

2. Attrition

Randomization, in principle, should result in similar groups, but attrition from these groups may create dissimilarities. Attrition occurs when an outcome variable is not available for all participants initially assigned to the intervention and comparison groups. The WWC is concerned about overall attrition as well as differences in the rates of attrition for the intervention and comparison groups. If there are high levels of attrition, the initial equivalence of the intervention and comparison groups may be compromised and the effect size estimates may be biased.

Both overall and differential attrition contribute to the potential bias of the estimated effect. The WWC has developed a model of attrition bias to calculate the potential bias under assumptions about the relationship between response and the outcome of interest.[2] The following figure illustrates the combination of overall and differential attrition rates that generates acceptable, potentially acceptable, and unacceptable levels of expected bias under certain circumstances that characterize many studies in education. In this figure, an acceptable level of bias is defined as an effect size of 0.05 of a standard deviation or less on the outcome.



The blue/top-right region shows combinations of overall and differential attrition that result in high levels of potential bias, and the green/bottom-right region shows combinations that result in low levels of potential bias. However, within the yellow/middle region of the figure, the potential bias depends on the assumptions of the model.

In developing the topic area review protocol, the PI considers the types of samples and likely relationship between attrition and student outcomes for studies in the topic area. In cases where a PI has reason to believe that much of the attrition is exogenous—such as parent mobility

[2] For details on the model of attrition bias and the development of the standard, please see Appendix A.

with young children—more optimistic assumptions regarding the relationship between attrition and the outcome might be appropriate. On the other hand, in cases where a PI has reason to believe that much of the attrition is endogenous—such as high school students choosing whether to participate in an intervention—more conservative assumptions may be appropriate. This results in a specific set of combinations of overall and differential attrition that separates high and low levels of attrition to be applied consistently for all studies in a topic area:

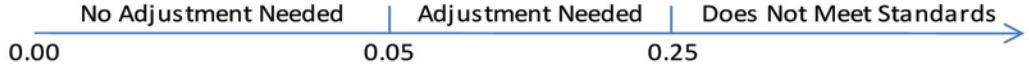
- For a study in the green/bottom-right region, attrition is expected to result in an acceptable level of bias even under conservative assumptions, which yields a rating of *Meets Evidence Standards*.
- For a study in the blue/top-right region, attrition is expected to result in an unacceptable level of bias even under optimistic assumptions, and the study can receive a rating no higher than *Meets Evidence Standards with Reservations*, provided that it establishes baseline equivalence of the analysis sample.
- For a study in the yellow/middle region, the PI's judgment about the sources of attrition for the topic area determines whether a study *Meets Evidence Standards*. If a PI believes that optimistic assumptions are appropriate for the topic area, then a study that falls in this range is treated as if it were in the green/bottom-right region. If a PI believes that conservative assumptions are appropriate, then a study that falls in this range is treated as if it were in the blue/top-right region. The choice of the boundary establishing acceptable levels of attrition is articulated in the protocol for each topic area.

3. Establishing Equivalence in RCTs with Attrition and QEDs

The WWC requires that RCTs with high levels of attrition and all QEDs present evidence that the intervention and comparison groups are alike. Demonstrating equivalence minimizes potential bias from attrition (RCTs) or selection (QEDs) that can alter effect size estimates.

Baseline equivalence of the analytical sample must be demonstrated on observed characteristics defined in the topic area protocol, using these criteria:

- The reported difference of the characteristics must be less than 0.25 of a standard deviation (based on the variation of that characteristic in the pooled sample).[3]
- In addition, the effects must be statistically adjusted for baseline differences in the characteristics if the difference is greater than 0.05 of a standard deviation.



[3] The standard limiting pre-intervention differences between groups to 0.25 standard deviations is based on Ho, Imai, King, and Stuart (2007).

Statistical adjustments include, but are not necessarily limited to, techniques such as ordinary least squares regression adjustment for the baseline covariates, fixed effects (difference-in-differences) models, and ANCOVA analysis.

4. Confounding Factor

In some studies, a component of the design lines up exactly with the intervention or comparison group (for example, studies in which there is one "unit"—teacher, classroom, school, or district—in one of the conditions). In these studies, the confounding factor may have a separate effect on the outcome that cannot be eliminated by the study design. Because it is impossible to separate how much of the observed effect was due to the intervention and how much was due to the confounding factor, the study cannot meet standards, as the findings cannot be used as evidence of the program's effectiveness.

5. Reasons for Not Meeting Standards

A study may fail to meet WWC evidence standards if

- It does not include a valid or reliable outcome measure, or does not provide adequate information to determine whether it uses an outcome that is valid or reliable.
- It includes only outcomes that are overaligned with the intervention or measured in a way that is inconsistent with the protocol.
- The intervention and comparison groups are not shown to be equivalent at baseline.
- The overall attrition rate exceeds WWC standards for an area.
- The differential attrition rate exceeds WWC standards for an area.
- The estimates of effects did not account for differences in pre-intervention characteristics while using a quasi-experimental design.
- The measures of effect cannot be attributed solely to the intervention—there was only one unit of analysis in one or both conditions.
- The measures of effect cannot be attributed solely to the intervention—the intervention was combined with another intervention.
- The measures of effect cannot be attributed solely to the intervention—the intervention was not implemented as designed.

6. Corrections and Adjustments

Different types of effect size indices have been developed for different types of outcome measures, given their distinct statistical properties. For continuous outcomes, the WWC has adopted the most commonly-used effect size index—the standardized mean difference, which is defined as the difference between the mean outcome of the intervention group and the mean

outcome of the comparison group divided by the pooled within-group standard deviation on that outcome measure. (See Appendix B for the rationale for the specific computations conducted by the WWC and their underlying assumptions.)

When the unit of assignment differs from the unit of analysis, the resulting analysis yields statistical tests with greater apparent precision than they actually have. Although the point estimates of the intervention's effects are unbiased, the standard errors of the estimates are likely to be underestimated, which would lead to overestimated statistical significance. In particular, a difference found to be statistically significant without correcting for this issue might actually not be statistically significant.

When a statistically significant finding is reported from a misaligned analysis, and the author is not able to provide a corrected analysis, the effect sizes computed by the WWC incorporate a statistical adjustment for clustering. The default (based on Hedges' summary of a wide range of studies) intraclass correlation used for these corrections is 0.20 for achievement outcomes and 0.10 for behavioral and attitudinal outcomes. (See Appendix C.)

When a study examines many outcomes or findings simultaneously (for example, a study examines multiple outcomes in a domain or has more than one treatment or comparison condition), the statistical significance of findings may be overstated. Without accounting for these multiple comparisons, the likelihood of finding a statistically significant finding increases with the number of comparisons. The WWC uses the Benjamini-Hochberg method to correct for multiple comparisons. (See Appendix D.)

The WWC makes no adjustments or corrections for variations in implementation of the intervention; however, if a study meets standards and is included in an intervention report, descriptions of implementation are provided in the report appendices to provide context for the findings. Similarly, the WWC also makes no adjustments for non-participation (intervention group members given the opportunity to participate in a program who chose not to) and contamination (control group members who receive the treatment). The PI for a topic area has the discretion to determine whether these issues are substantive enough to warrant reducing the rating of a study.

C. PILOT STANDARDS FOR REGRESSION DISCONTINUITY DESIGNS

Regression discontinuity (RD) designs are increasingly used by researchers to obtain unbiased estimates of the effects of education-related interventions. These designs are applicable when a continuous "scoring" rule is used to assign the intervention to study units (e.g., school districts, schools, or students). Units with scores below a preset cutoff value are assigned to the treatment group and units with scores above the cutoff value are assigned to the comparison group, or vice versa. For example, students may be assigned to a summer school program if they score below a preset point on a standardized test, or schools may be awarded a grant based on their score on an application.

Under an RD design, the effect of an intervention can be estimated as the difference in mean outcomes between treatment and comparison group units, adjusting statistically for the relationship between the outcomes and the variable used to assign units to the intervention,

typically referred to as the "forcing" or "assignment" variable. A regression line (or curve) is estimated for the treatment group and similarly for the comparison group, and the difference in average outcomes between these regression lines at the cutoff value of the forcing variable is the estimate of the effect of the intervention. Stated differently, an effect occurs if there is a "discontinuity" in the two regression lines at the cutoff. This estimate pertains to average treatment effects for units right at the cutoff. RD designs generate unbiased estimates of the effect of an intervention if (a) the relationship between the outcome and forcing variable can be modeled correctly and (b) the forcing variable was not manipulated to influence treatment assignments.

A study qualifies as an RD study if it meets *all* of the following criteria:

- **Treatment assignments are based on a forcing variable; units with scores at or above (or below) a cutoff value are assigned to the treatment group while units with scores on the other side of the cutoff are assigned to the comparison group.** For example, an evaluation of a tutoring program could be classified as an RD study if students with a reading test score at or below 30 are admitted to the program and students with a reading test score above 30 are not. As another example, a study examining the impacts of grants to improve teacher training in local areas could be considered an RD study if grants are awarded to only those sites with grant application scores that are at least 70. In some instances, RD studies may use multiple criteria to assign the treatment to study units. For example, a student may be assigned to an after-school program if the student's reading score is below 30 *or* math score is below 40. As with RCTs, noncompliance with treatment assignment is permitted, but the study must still meet the criteria below to meet evidence standards.

- **The forcing variable must be ordinal with a sufficient number of unique values.** This condition is required to model the relationship between the outcomes and forcing variable. The forcing variable should never be based on cardinal (non-ordinal) categories (such as gender or race). The analyzed data also must include at least four unique values of the forcing variable below the cutoff and four unique values above the cutoff.

- **There must be no factor confounded with the forcing variable.** The cutoff value for the forcing variable must not be used to assign students to interventions other than the one being tested. For example, free/reduced-price lunch (FRPL) status cannot be the basis of an RD design, because FRPL is used as the eligibility criteria for a wide variety of services. This criterion is necessary to ensure that the study can isolate the causal effects of the tested intervention from the effects of other interventions.

If a study claims to be based on an RD design but does not have these properties, the study *Does Not Meet Evidence Standards* as an RD design. Once a study is determined to be an RD design, the study can receive one of three designations based on the following criteria: (1) integrity of the forcing variable; (2) attrition; (3) continuity of the outcome—forcing variable relationship; and (4) functional form and bandwidth.[4]

[4] Details on these criteria and the rating of regression discontinuity studies are presented in Appendix E.

D. PILOT STANDARDS FOR SINGLE-CASE DESIGNS

Single-case designs can provide causal evidence of intervention effects. Although the basic SCD has many variations, these designs often involve repeated, systematic measurement of a dependent variable before, during, and after the active manipulation of an independent variable (e.g., applying an intervention). For example, a behavior may be measured for a student over four phases (periods of time), ABAB, where A is a phase without the intervention and B is a phase with the intervention. Thus, an individual "case" is the unit of intervention administration and data analysis, and the case provides its own control for purposes of comparison. The outcome variable must be measured repeatedly within and across different conditions or levels of the independent variable. This differs from the pre-post design, which simply examines data once before and once after participation in a program.

These standards are intended to guide WWC reviewers in identifying and evaluating single-case designs (SCDs). The first section assists in identifying an SCD. If a study is an eligible SCD, it is then reviewed using the study rating criteria to determine whether it receives a rating of *Meets Evidence Standards, Meets Evidence Standards with Reservations,* or *Does Not Meet Evidence Standards*. A study that meets standards (with or without reservations) is then reviewed using visual analysis to determine whether it provides *Strong Evidence of a Causal Relation, Moderate Evidence of a Causal Relation,* or *No Evidence of a Causal Relation* for each outcome. SCDs are identified by the following features:

- An individual "case" is the unit of intervention administration and data analysis. A case may be a single participant or a cluster of participants (e.g., a classroom or community).
- Within the design, the case provides its own control for purposes of comparison. For example, the case's series of outcome variables prior to the intervention is compared with the series of outcome variables during (and after) the intervention.
- The outcome variable is measured *repeatedly* within and across *different* conditions or levels of the independent variable. These different conditions are referred to as "phases" (e.g., baseline phase, intervention phase).

The standards for SCDs apply to a wide range of designs, including ABAB and other designs.[5] Even though SCDs can be augmented by including one or more independent comparison cases (i.e., a comparison group), in this document the standards address only the core SCDs and are not applicable to the augmented independent comparison SCDs. If the study appears to be an SCD, a set of rules are used to determine whether the study's design *Meets Evidence Standards, Meets Evidence Standards with Reservations,* or *Does Not Meet Evidence Standards.*

[5] Details on these designs, along with the rules and rating criteria, are presented in Appendix F.

IV. SUMMARIZING THE REVIEW TO DEVELOP AN INTERVENTION REPORT

After reviewing all studies of an intervention within a topic area, the WWC will write an intervention report summarizing the findings of the review. This chapter describes the types of intervention reports, the process of preparing the report, components of the intervention report, the rating system used to determine the evidence rating, and the metrics and computations used to aggregate and present the evidence.

A. TYPES OF INTERVENTION REPORTS

If an intervention has at least one study meeting standards or meeting standards with reservations, an intervention report is prepared that presents the empirical findings, the rating of the evidence, and the improvement index for the magnitude of the effect synthesized from the evidence. As described earlier, the information for preparing these reports is generated from the study review guides developed by the reviewers.

If an intervention is determined not to have studies that meet standards or meet standards with reservations, an intervention report is prepared indicating that no evidence was found that met standards. The report provides additional details on the studies, categorized by the reason that each did not meet standards. As with the intervention report based on studies meeting standards, it includes a full list of all studies that were reviewed, along with the specific reason that each did not meet standards. These reports are careful to note that because there are no studies that meet standards, they cannot provide any statement about the effectiveness of the intervention.

Because educational research is ongoing during the review process, the WWC periodically revisits interventions, examining all new research that has been produced since the release of the intervention report. After the review of additional studies is complete, the WWC will release an updated intervention report. If some of the new research meets standards, the summary measures (effect size, improvement index, and rating) may change.

B. PREPARING THE REPORT

Based on reviews of the literature for a particular intervention, an intervention report examines all studies of the intervention within a topic area.[6] An intervention report provides a description of the intervention and references all relevant research. Intervention reports undergo a rigorous peer review process.

[6] An intervention may be reviewed in more than one topic area. For example, one intervention may affect outcomes in both beginning reading and early childhood, and therefore result in a separate intervention report for each area.

1. Draft Report

After a review of research on an intervention is complete, a topic area PI will assign drafting a report on the intervention to a certified reviewer. The WWC produces intervention reports even for those interventions for which no studies fall into the scope of the review or meet standards, as well as reports for interventions for which one or more studies meet standards or meet standards with reservations. The report writer completes the report by filling in the appropriate report template based on information from reviews of the studies.

Draft revisions occur at numerous points of the writing and production processes. After the report writer has developed the draft, the PI or Deputy PI reviews the report draft and provides feedback and suggestions. Based on PI feedback, the writer edits the draft and provides another draft to the PI or Deputy PI for additional comments. After approval is received from the PI or Deputy PI, the draft is reviewed by WWC staff to verify, among other things, that the correct template was used, study counts match the number of studies listed in the references, current study disposition codes were used, and all parts of the template have been completed.

2. Quality Assurance Review

At this point, the draft is submitted to a quality assurance (QA) reviewer who is a senior member of the WWC staff. The QA reviews the document and returns comments or changes to the report writer. When QA comments have been addressed, the PI sends the report to IES for external peer review.

3. IES and External Peer Review

Upon receiving the report from the PI, the IES reviews the report, sends it for external peer review, collects peer reviewer comments, and returns them to the Topic Area Team. The external peer reviewers are researchers who are not affiliated with the WWC but are knowledgeable about WWC standards. The report writer and the PI address the comments, resubmitting a revised draft to the IES for final approval. Intervention reports for which no studies meet evidence standards are subject only to IES review, not external peer review.

4. Production and Release

The production process begins when final approval for the intervention report is received from the IES. In addition to developing a PDF version of the report, production includes developing an HTML version for the website; creating a rotating banner image to advertise the release of the report on the WWC website home page; and writing text for the "What's New" announcement and e-mail blasts, which are sent to all WWC and IES NewsFlash subscribers.

Additionally, the PI sends a letter to the developer indicating that the WWC is posting an intervention report on its website. Developers receive an embargoed copy of the intervention report 24 hours prior to its release on the WWC website. This is not a review stage, and the report will not be immediately revised based on developer comments. If developers have

questions about the report, they are encouraged to contact the WWC in writing, and the issues will be examined by the quality review team described in Chapter I.

C. COMPONENTS OF THE REPORT

The intervention report is a summary of all the research reviewed for an intervention within a topic area. It contains three types of information—program description, research, and effectiveness—presented in a number of ways. This section describes the contents of the intervention report.

1. Front Page

The front page of the intervention report provides a quick summary of all three types of the information just noted. The *Program description* section describes the intervention in a few sentences and is drafted using information from publicly available sources, including studies of the intervention and the developer's website. The description is sent to the developer to solicit comments on accuracy and to ask for any additional information, if appropriate.

The *Research* section summarizes the studies on which the findings of effectiveness were based, delineating how many studies met standards with and without reservations. The section also provides a broad picture of the scope of the research, including the number of students and locations, along with domains for which the studies examined outcomes.

Finally, the *Effectiveness* section reports the rating of effectiveness (detailed in the later section on report appendices) taken from Appendix A5 of the report, along with the improvement index average and range taken from Appendix A3 of the report, by domain. These ratings and indices are the "bottom line" of the review and appear in the summary of evidence tables in both the topic report and the user-generated summary tables available for each topic area on the website.

2. Body of the Report

The text of the report covers all three types of information again, but with more detail. The *Additional program information* section provides a more in-depth description of the intervention, including contact information for the developer, information on where and how broadly the intervention is used, a more detailed description of the intervention, and an estimate of the cost of the program. Again, these are obtained from publicly-available sources and reviewed by the developer for accuracy and completeness.

The *Research* section in this part of the report gives a more complete picture of the research base, detailing all the studies that were reviewed for the report and the disposition for each study. For those that meet WWC evidence standards, with or without reservations, a paragraph describes the study design and samples, along with any issues related to the rating, using information from Appendix A1 of the intervention report.

For each domain with outcomes examined in the studies, the *Effectiveness* section includes a paragraph describing the findings. Taken from Appendix A3, these include the specific sample examined, the outcome(s) studied, the size(s) of the effect, and whether the findings are statistically significant or substantively important. This section also describes the rating of effectiveness and improvement index generally, as well as the specific ratings and indices found for the intervention, followed by a paragraph summarizing all the research and effectiveness findings.

The body of the report concludes with a list of *References*, broken down by study disposition. Additional sources that provide supplementary information about a particular study are listed with the main study. Finally, for each study that was not used in the measures of effectiveness, because it either was outside the scope of the review or did not meet WWC evidence standards, an explanation of the exact reason for its exclusion is provided.

3. Appendices

Following the body of the report are technical appendices that provide the details of studies underlying the presented ratings. Appendix A1 provides much more detail and context for each study that meets standards, including a table containing the full study citation, details of the study design, a description of study participants, the setting in which the study was conducted, descriptions of the intervention and comparison conditions as implemented in the study, the outcomes examined, and any training received by staff to implement the intervention. Appendix A2 provides more detail on the outcomes examined in the studies that meet standards, grouped by domain.

Appendix A3 consists of tables that summarize the study findings by domain. For each outcome, a row includes the study sample, sample size, the means and standard deviations of the outcome for the treatment and comparison groups, the difference in means, the effect size, an indicator for statistical significance, and the improvement index. An average is presented for all outcomes (within a domain) for a study, along with an average for all studies in a domain. Footnotes describe the table components, as well as any issues particular to the studies, such as whether corrections needed to be made for clustering or multiple comparisons.

Appendix A4 consists of tables similar to those in Appendix A3, summarizing findings by domain, with rows for each outcome. However, these tables contain supplemental findings that are not used in the determination of the rating for an intervention. Findings in these tables may include those for subgroups of interest, subscales of a test, or a different follow-up period.

The information in Appendices A1 through A4 comes from the studies and the reviewer summaries. Appendix A5 uses information and findings from all the studies to create aggregate measures of effectiveness. For each domain, the intervention rating scheme is applied to determine the rating for the intervention in that domain, based on the number of studies, study designs, and findings. The criteria for each rating are evaluated, with the intervention receiving the highest rating for which it meets the associated criteria, and the criteria for unattained higher ratings are described.

Appendix A6 aggregates the setting information of the passing studies, including the number of studies, schools, classrooms, and students, to create a measure of the extent of evidence for the intervention in each domain. The summaries from Appendices A5 and A6 are the source of the bottom-line rating information presented in the table at the foot of the front page of the intervention report.

D. INTERVENTION RATING SCHEME

As it does in rating studies, the WWC uses a set of guidelines to determine the rating for an intervention. To obtain this rating, the intervention rating scheme provides rules for combining the findings from multiple studies. An additional complexity, relative to rating a single study, is that different studies can yield different findings. Similarly, interventions may receive different ratings in different domains, since the evidence varies across types of outcomes.

The WWC's intervention rating scheme has six mutually exclusive categories that span the spectrum from positive effects to negative effects, with two categories for potentially positive and potentially negative effects, and two other categories of mixed evidence (when positive and negative effects are found in studies meeting standards) and no discernible effects (when all of studies meeting standards show statistically insignificant and substantively small effects).

Both statistical significance and the size of the effect play a role in rating interventions. Statistically significant effects are noted as "positive" (defined as favoring the intervention group) or "negative" in the ratings. Effects that are not statistically significant but have an effect size of at least 0.25 are considered "substantively important" and are also considered in the ratings. A third factor contributing to the rating is whether the quality of the research design generating the effect estimate is strong (RCT) or weak (QED).

The rating scheme based on these factors is presented next; the detailed descriptions for making the judgments on these factors for each study and outcome are presented in Appendix G of this handbook.

Positive Effects: Strong evidence of a positive effect with no overriding contrary evidence.

- Two or more studies showing *statistically significant* positive effects, at least one of which met WWC evidence standards for a *strong* design.
- No studies showing *statistically significant* or *substantively important* negative effects.

Potentially Positive Effects: Evidence of a positive effect with no overriding contrary evidence.

- At least one study showing a statistically significant or substantively important positive effect.

- No studies showing a statistically significant or substantively important negative effect AND fewer or the same number of studies showing indeterminate effects than showing statistically significant or substantively important positive effects.

Mixed Effects: Evidence of inconsistent effects, demonstrated through either of the following:

- At least one study showing a *statistically significant* or *substantively important* positive effect AND at least one study showing a *statistically significant* or *substantively important* negative effect, but no more such studies than the number showing a *statistically significant* or *substantively important* positive effect.
- At least one study showing a *statistically significant* or *substantively important* effect AND more studies showing an *indeterminate* effect than showing a *statistically significant* or *substantively important* effect.

No Discernible Effects: No affirmative evidence of effects.

- None of the studies shows a *statistically significant* or *substantively important* effect, either positive or negative.

Potentially Negative Effects: Evidence of a negative effect with no overriding contrary evidence.

- At least one study showing a *statistically significant* or *substantively important* negative effect.
- No studies showing a *statistically significant* or *substantively important* positive effect OR more studies showing *statistically significant* or *substantively important* negative effects than showing *statistically significant* or *substantively important* positive effects.

Negative Effects: Strong evidence of a negative effect with no overriding contrary evidence.

- Two or more studies showing *statistically significant* negative effects, at least one of which met WWC evidence standards for a *strong* design.
- No studies showing statistically significant or substantively important positive effects.

E. AGGREGATING AND PRESENTING FINDINGS

Several additional WWC standards are used in preparing intervention reports. To compare results across studies, effect sizes are averaged for studies meeting standards or meeting them with reservations. Based on the average effect size, an improvement index is calculated, and the intervention report also indicates the maximum and minimum effect size for studies meeting standards that have outcomes within a domain. Additionally, the extent of evidence is another

consideration in rating interventions. This section describes these concepts, with technical details presented in Appendices B, F, and G.

1. Effect Size

To assist in the interpretation of study findings and to facilitate comparisons of findings across studies, the WWC computes the effect sizes associated with study findings on outcome measures relevant to the topic area review. In general, the WWC focuses on student-level findings, regardless of the unit of assignment or the unit of intervention. Focusing on student-level findings not only improves the comparability of effect size estimates across studies, but also allows us to draw upon existing conventions among the research community to establish the criterion for substantively important effects for intervention rating purposes.

Different types of effect size indices have been developed for different types of outcome measures, given their distinct statistical properties. For continuous outcomes, the WWC has adopted the most commonly-used effect size index—the standardized mean difference, which is defined as the difference between the mean outcome of the intervention group and the mean outcome of the comparison group, divided by the pooled within-group standard deviation on that outcome measure. Given the focus on student-level findings, the default standard deviation used in the effect size computation is the student-level standard deviation. This effect size index is referred to as Hedges's *g*. For binary outcomes, the effect size measure of choice is the odds ratio. In certain situations, however, the WWC may present study findings using alternative measures. For details on these calculation and others, see Appendix B on effect size computations.

The WWC potentially performs two levels of aggregation to arrive at the average effect size for a domain in an intervention report. First, if a study has more than one outcome in a domain, the effect sizes for all of that study's outcomes are averaged into a study average. Second, if more than one study has outcomes in a domain, the study average for all of those studies is averaged into a domain average.

2. Improvement Index

In order to help readers judge the practical importance of an intervention's effect, the WWC translates effect sizes into an improvement index. The improvement index represents the difference between the percentile rank corresponding to the intervention group mean and the percentile rank corresponding to the comparison group mean (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.

In addition to the improvement index for each individual finding, the WWC also computes a study average improvement index for each study, as well as a domain average improvement index across studies for each outcome domain. The study average improvement index is computed based on the study 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 average effect size for individual studies. The computation of the improvement index is detailed in Appendix H.

3. Extent of Evidence

The extent of evidence categorization was developed to tell readers how much evidence was used to determine the intervention rating, focusing on the number and sizes of studies. Currently, this scheme has two categories: small and medium to large. The extent of evidence categorization described here is not a rating on external validity; instead, it serves as an indicator that cautions readers when findings are drawn from studies with small samples, a small number of school settings, or a single study. Details of the computation, along with the rationale, are described in Appendix I.

REFERENCES

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

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

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

Bloom, H. S., Bos, J.M., & Lee, S.W. (1999). Using cluster random assignment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review, 234,* 445–469.

Bonferroni, C. E. (1935). Il calcolo delle assicurazioni su gruppi di teste. In *Studi in onore del Professore Salvatore Ortu Carboni* (pp. 13–16). Rome.

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

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

Donner, A., & Klar, N. (2000). *Design and analysis of cluster randomized trials in health research.* London: Arnold Publishing.

Dunnett, C. (1955). A multiple comparisons procedure for comparing several treatments with a control. *Journal of American Statistical Association, 50*, 1096–1121.

Flay, B. R., & Collins, L. M. (2005). Historical review of school-based randomized trials for evaluating problem behavior prevention programs. *The Annals of the American Academy of Political and Social Science, 599*, 147–175.

Fleiss, J. L. (1994). Measures of effect size for categorical data. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 245–260). New York: Russell Sage Foundation.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*, 107–128.

Hedges, L. V. (2005). *Correcting a significance test for clustering.* Unpublished manuscript.

Ho, D., Imai, K., King, G., & Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, *15*, 199–236.

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

Murray, D. M. (1998). *Design and analysis of group-randomized trials* (Vol. 27). New York: Oxford University Press.

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

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

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (pp. 231–244). New York: Russell Sage Foundation.

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

Sanchez-Meca, J., Marin-Martinez, F., & Chacon-Moscoso, S. (2003). Effect-size indices for dichotomous outcomes in meta-analysis. *Psychological Methods, 8*(4), 448–467.

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

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

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

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

APPENDIX A. ASSESSING ATTRITION BIAS

A. INTRODUCTION

In a randomized controlled trial (RCT), researchers use random assignment to form two groups of study participants that are the basis for estimating intervention effects. Carried out correctly, the groups formed by random assignment have similar observable and unobservable characteristics, allowing any differences in outcomes between the two groups to be attributed to the intervention alone, within a known degree of statistical precision.

Though randomization (done correctly) results in statistically similar groups at baseline, the two groups also need to be equivalent at follow-up, which introduces the issue of attrition. Attrition occurs when an outcome is not measured for all participants initially assigned to the two groups. Attrition can occur for the overall sample, and it can differ between the two groups; both aspects can affect the equivalence of the groups. Both overall and differential attrition create potential for bias when the characteristics of sample members who respond in one group differ systematically from those of the members who respond in the other.

To support its efforts to assess design validity, the What Works Clearinghouse (WWC) needs a standard by which it can assess the likelihood that findings of RCTs may be biased due to attrition. This appendix develops the basis for the RCT attrition standard. It uses a statistical model to assess the extent of bias for different rates of overall and differential attrition under different assumptions regarding the extent to which respondent outcomes are correlated with the propensity to respond. The validity of these assumptions is explored using data from a past experimental evaluation.

A key finding is that there is a trade-off between overall attrition rates and differential attrition rates such that a higher rate of overall attrition can be offset by a lower rate of differential attrition (and vice versa). For example, the bias associated with an overall attrition rate of 10% and a differential attrition rate of 5% can be equal to the bias associated with an overall attrition rate of 30% and a differential attrition rate of 2%.

Assessing design validity requires considering both overall and differential attrition within a framework in which both contribute to possible bias. An approach for doing so is developed in the next section. Under various assumptions about tolerances for potential bias, the approach yields a set of attrition rates that falls within the tolerance and a set that falls outside it. Because different topic areas may have factors generating attrition that lead to more or less potential for bias, the approach allows for refinement within a review protocol that expands or contracts the set of rates that yield tolerable bias. This approach is the basis on which WWC attrition standards can be set.

B. ATTRITION AND BIAS

Both overall and differential attrition may bias the estimated effect of an intervention.[7] However, the sources of attrition and their relation to outcomes rarely can be observed or known with confidence (an important exception being clearly exogenous "acts of nature," such as hurricanes or earthquakes, which can cause entire school districts to drop out of a study), which limits the extent to which attrition bias can be quantified. The approach here is to develop a model of attrition bias that yields potential bias under assumptions about the correlation between response and outcome. This section describes the model and its key parameters. It goes on to identify values of the parameters that are consistent with the WWC's current standards, and it assesses the plausibility of the parameters using data from a recent randomized trial.

1. Model of Attrition Bias

Attrition that arises completely at random reduces sample sizes but does not create bias. However, researchers rarely know whether attrition is random and not related to outcomes. When attrition *is* related to outcomes, different rates of attrition between the treatment and control groups can lead to biased impacts. Furthermore, if the relationship between attrition and outcomes differs between the treatment and control groups, then attrition can lead to bias even if the attrition rate is the same in both groups. The focus here is to model the relationship between outcomes and attrition in a way that allows it to be manipulated and allows bias to be assessed under different combinations of overall and differential attrition.

To set up the model, consider a variable representing an individual's latent (unobserved) propensity to respond, z. Assume z has an N(0,1) distribution. If the proportion of individuals who respond is ρ, an individual is a respondent if his or her value of z exceeds a threshold:

$$(1) \quad z > Q(z, 1-\rho)$$

where the quantile function, Q, is the inverse of the cumulative distribution function. That is, if z is greater than the value that corresponds to a particular percentile of the z distribution (given ρ), then an individual responds at follow-up.

The outcome at follow-up, y, is the key quantity of interest. It can be viewed as the sum of two unobserved quantities, the first a factor that is unrelated to attrition (u) and the second the propensity to respond (z). The outcome can be modeled as

$$(2) \quad y = \alpha * z + \beta * u$$
$$\alpha = \delta * \theta$$
$$\beta = 1 - \theta$$

[7] Throughout this appendix, the word *bias* refers to a deviation from the true impact *for the analysis sample*. An alternative definition of bias could also include deviation from the true impact for a larger population. We focus on the narrower goal of achieving causal validity for the analysis sample because nearly all studies reviewed by the WWC involve purposeful samples of students and schools.

where u is a random variable that is assumed to be normally distributed N(0,1), θ is the proportion of the variation in y that is explained by z, and δ takes a value of +1 or –1 to allow y to be positively or negatively correlated with z.[8] Note that there are no covariates and the model assumes no effect of the treatment on the outcome. If θ is one, the entire outcome is explained by the propensity to respond. If θ is zero, none of the outcome is explained by the propensity to respond, which is the case when attrition is completely random.

The proportion of individuals responding at follow-up may differ by treatment status. Therefore, for treatment and control group members:

$$y_t = \alpha_t * z_t + \beta_t * u_t$$
$$y_c = \alpha_c * z_c + \beta_c * u_c$$

If α is the same for both treatment and control group members, then equal rates of attrition in the treatment and control groups do not compromise the causal validity of the impact because the same kind of individuals attrite from both groups.[9] However, if the rates of attrition differ between the treatment and control groups, then the causal validity of the impact is compromised even when $\alpha_t = \alpha_c$. If $\alpha_t \neq \alpha_c$, then impacts will be biased even if the attrition rate is the same in both groups because the types of students who attrite differ between the treatment and control groups.[10]

In this model, *bias* is the difference between y_t and y_c among respondents. It is generated by differences in the response rates (ρ_t and ρ_c) or in the proportion of the variation in y explained by z (θ_t and θ_c) for the two groups.

2. Using the Model to Assess Current Standards

The inputs to the model are the parameters θ_t, θ_c, δ_t, δ_c, ρ_t, and ρ_c. With values chosen for the parameters, the model yields outcomes and estimates of bias once the two random variables z and u are given values.

Using a program written in R, 5,000 draws of z_t, z_c, u_t, and u_c were created and inserted into the model. For each individual, follow-up response (0 or 1) was then determined using equation (1), and the outcome was determined using equation (2).

[8] In a regression of y on z, θ would be the regression R^2.

[9] Those who attrite, nonetheless, will differ systematically from those who do not attrite, which possibly creates issues for external validity.

[10] It is possible that a difference in the rate of attrition between groups could offset a difference between α_t and α_c. However, throughout this appendix, we conservatively assume the opposite—that these differences are reinforcing, not offsetting.

Bias is the difference in mean outcomes between treatment and control respondents. Table A1 reports bias in effect size units for various assumptions about the parameters. The key finding in this table is that given a set of assumptions regarding the correlation between outcomes and the propensity to respond (these assumptions vary by column), bias can be reduced by either increasing the overall response rate or reducing the differential response rate. For example, column 4 shows that an overall response rate of 60% yields a bias of 0.05 only if the differential rate is 2% or less, but that if the overall rate is 90%, the differential rate can be as high as 5%.

TABLE A1

BIAS BY RESPONSE RATE AND PROPORTION OF OUTCOME EXPLAINED BY RESPONSE (EFFECT SIZE UNITS)

P_T	P_C	$\alpha_t = 0.075$ $\alpha_c = 0.050$	$\alpha_t = 0.01$ $\alpha_c = 0.05$	$\alpha_t = 0.15$ $\alpha_c = 0.05$	$\alpha_t = 0.20$ $\alpha_c = 0.15$	$\alpha_t = 0.30$ $\alpha_c = 0.20$	$\alpha_t = 0.50$ $\alpha_c = 0.20$	$\alpha_t = 1.00$ $\alpha_c = 1.00$	$\alpha_t = 1.00$ $\alpha_c = -1.00$
0.900	0.900	0.01	0.02	0.03	0.01	0.02	0.05	0.00	0.39
0.890	0.910	0.02	0.03	0.04	0.03	0.04	0.07	0.03	0.39
0.875	0.925	0.03	0.04	0.06	0.05	0.06	0.10	0.08	0.39
0.865	0.935	0.04	0.05	0.07	0.06	0.08	0.12	0.12	0.39
0.850	0.950	0.05	0.06	0.08	0.08	0.10	0.15	0.17	0.38
0.800	0.800	0.02	0.03	0.06	0.02	0.03	0.09	0.00	0.70
0.790	0.810	0.02	0.04	0.07	0.03	0.05	0.11	0.03	0.70
0.775	0.825	0.04	0.05	0.08	0.05	0.07	0.13	0.07	0.70
0.765	0.835	0.04	0.06	0.09	0.06	0.09	0.15	0.10	0.70
0.750	0.850	0.05	0.07	0.10	0.08	0.11	0.18	0.15	0.70
0.700	0.700	0.02	0.05	0.08	0.03	0.05	0.13	0.00	0.99
0.690	0.710	0.03	0.05	0.09	0.04	0.06	0.15	0.03	0.99
0.675	0.725	0.04	0.07	0.10	0.06	0.09	0.17	0.07	0.99
0.665	0.735	0.05	0.07	0.11	0.07	0.10	0.19	0.10	0.99
0.650	0.750	0.06	0.09	0.13	0.09	0.12	0.21	0.15	0.99
0.600	0.600	0.03	0.06	0.11	0.04	0.06	0.17	0.00	1.29
0.590	0.610	0.04	0.07	0.12	0.05	0.08	0.18	0.03	1.29
0.575	0.625	0.05	0.08	0.13	0.07	0.10	0.21	0.07	1.29
0.565	0.635	0.06	0.09	0.14	0.08	0.12	0.23	0.10	1.29
0.550	0.650	0.07	0.10	0.15	0.10	0.14	0.25	0.15	1.29

But what assumptions are appropriate regarding the extent to which response is related to outcome (the magnitudes of α coefficients that vary across the columns of Table A1)? We could infer possible appropriate assumptions from existing studies if we could somehow measure the extent of differences in outcomes between respondents and nonrespondents, and whether those differences are themselves different between the treatment and control groups. We could then compare those observed differences to what those differences would be for different values

of α_t and α_c using our model of attrition. Of course, we cannot do this directly, because we do not observe outcomes for nonrespondents. However, in studies that have both follow-up and baseline test scores, we can use the baseline test scores as proxies for the follow-up test scores.

The example used here is Mathematica's evaluation of education technology interventions. The evaluation had overall response rates above 90% for its sample and almost no differential response, which means that it is close to the first line of Table A1 (equal response rates of 90% in the groups). The study's data allow calculations of differences in *baseline* test scores for follow-up respondents and nonrespondents. Baseline test scores are highly correlated with follow-up test scores, which means the baseline scores can proxy for follow-up scores.

The education technology study had four interventions that were implemented in four grade levels (first, fourth, sixth, and ninth) that essentially operated as distinct studies. Overall effect size differences between respondents and nonrespondents for the four study components were 0.41, 0.44, 0.51, and 0.23, an average of 0.40. The differences between the treatment and control groups in these respondent-nonrespondent differences were 0.10, 0.11, 0.10, and 0.10.

Table A2 shows the difference in effect size units between respondents and nonrespondents, and the difference in that difference between the treatment and control groups for the same α assumptions as in Table A1, but restricting attention to the case of 90% response and no differential response (the same rates observed in the education technology data). In Table A2, the closest match for the respondent-nonrespondent difference of 0.40 is found in the first column, in which the difference is 0.49. The closest match for the treatment-control difference in the respondent-nonrespondent difference is also in the first column, in which the difference-in-difference is 0.10. In other words, in the education technology study, response had little correlation with the baseline test score (our proxy for the study's outcome measure), and this correlation did not differ significantly between the treatment and control groups.

TABLE A2

OVERALL DIFFERENCES BETWEEN RESPONDENTS AND NONRESPONDENTS AND THE DIFFERENCE IN THAT DIFFERENCE BETWEEN THE TREATMENT AND CONTROL GROUPS IN THE CASE OF 90% RESPONSE AND NO DIFFERENTIAL ATTRITION

	(1) $\alpha_t = 0.075$ $\alpha_c = 0.05$	(2) $\alpha_t = 0.10$ $\alpha_c = 0.05$	(3) $\alpha_t = 0.15$ $\alpha_c = 0.05$	(4) $\alpha_t = 0.20$ $\alpha_c = 0.15$	(5) $\alpha_t = 0.30$ $\alpha_c = 0.20$	(6) $\alpha_t = 0.50$ $\alpha_c = 0.20$	(7) $\alpha_t = 1.00$ $\alpha_c = 1.00$	(8) $\alpha_t = 1.00$ $\alpha_c = -1.00$
Difference between all respondents and all nonrespondents	0.49	0.52	0.60	0.81	0.97	1.12	1.95	0.00
Difference between the treatment and control groups in the difference between respondents and nonrespondents	0.10	0.18	0.32	0.12	0.20	0.50	0.00	3.90

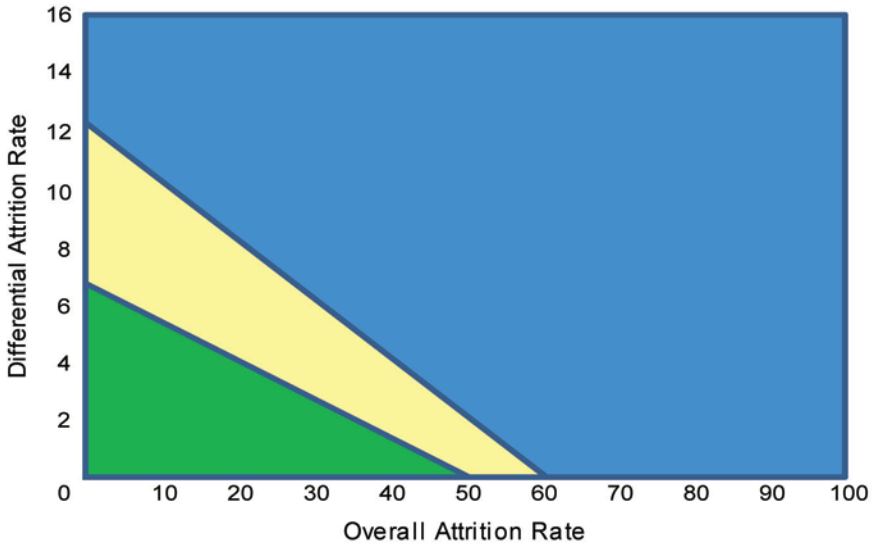
Intuitively, this conclusion is reasonable because students were not likely to attrite from the study because of their treatment or control status. The classroom was randomly assigned to use or not use a technology product and students had no discretion. Attrition in the education technology evaluation is more likely related to family mobility because of both the students' age and the nature of the intervention. However, for other populations of students, such as older students who volunteer to participate in a dropout prevention program, attrition may be more correlated with the outcome.

3. Attrition Trade-offs Assuming a Constant Relative Bias

The trade-off between response rates can be illustrated graphically by assuming a threshold degree of tolerable bias and examining values of overall and differential response that exceed or fall below the threshold. Figure A1 uses a bias threshold of 0.05 standard deviations of the outcome measure. The green/bottom-right region shows combinations of overall and differential attrition that yield attrition bias less than 0.05 under pessimistic (but still reasonable) assumptions (column 4 in Tables A1 and A2), the yellow/middle region shows additional combinations that yield attrition bias less than 0.05 under the most optimistic assumptions (column 1 in the tables), and the blue/top-right region shows combinations that yield bias greater than 0.05 even under the most optimistic assumptions.

FIGURE A1

TRADE-OFFS BETWEEN OVERALL AND DIFFERENTIAL ATTRITION



The model shows that both the overall attrition rate and the differential attrition rate can be viewed as contributing to bias, and it illuminates a relationship between the two rates. Operationalizing a standard requires choosing an appropriate degree of bias. There is no right or wrong answer to the amount of bias that can be tolerated. Empirically, the WWC would accept as evidence of effectiveness a study that reported an effect size of 0.25 that was statistically insignificant even though the true effect of the intervention might be as low as 0.20 (the WWC deems an effect size of 0.25 to be substantively important and factors this into its ratings for studies that meet standards).

To get some indication of how large the relative bias is, note that for a nationally normed test, a difference of 0.05 represents about 2 percentile points for a student at the 50th percentile. For example, if the reported effect suggests the intervention will move the student from the 50th percentile to the 60th percentile (a 0.25 effect size), the true effect may be to move the student from the 50th percentile to the 58th percentile (a 0.20 effect size). Doubling the tolerable bias to 0.10 means that an intervention that reportedly moves a student from the 50th percentile to the 60th percentile may move the student only to the 56th percentile. A relative bias of 67% (with a true effect of an increase of 6 percentile points and a reported effect of an increase of 10 percentile points, the bias would be 4 percentile points) seems large.

4. Using the Attrition Bias Model to Create a Standard

In developing the topic area review protocol, the principal investigator (PI) considers the types of samples and likely relationship between attrition and student outcomes for studies in the topic area. When a PI has reason to believe that much of the attrition is exogenous—for example, parent mobility with young children—more optimistic assumptions regarding the relationship between attrition and outcome might be appropriate. On the other hand, when a PI has reason to believe that much of the attrition is endogenous—for example, high school students choosing whether to participate in an intervention—more conservative assumptions may be appropriate. The combinations of overall and differential attrition that are acceptable given either optimistic or conservative assumptions are illustrated in Figure A1, and translate into evidence standards ratings:

- For a study in the green/bottom-right region, attrition is expected to result in an acceptable level of bias even under conservative assumptions, which yields a rating of *Meets Evidence Standards*.
- For a study in the blue/top-right region, attrition is expected to result in an unacceptable level of bias even under optimistic assumptions, and the study can receive a rating no higher than *Meets Evidence Standards with Reservations*, provided it establishes baseline equivalence of the analysis sample.
- For a study in the yellow/middle region, the PI's judgment about the sources of attrition for the topic area determines whether a study *Meets Evidence Standards*. If a PI believes that optimistic assumptions are appropriate for the topic area, then a study that falls in this range is treated as if it were in the green/bottom-right region. If a PI believes that conservative assumptions are appropriate, then a study that falls in this range is treated as if it were in the red area.

To help reviewers implement this standard, the WWC needs to develop a simple formula to determine whether a study falls in the blue/top-right, yellow/middle, or green/bottom-right region for a topic area. The inputs to this formula will be the overall and differential attrition rates, which are already collected by WWC reviewers. When entire school districts are lost from a study due to clearly exogenous "acts of nature," the attrition standard will be applied to the remaining districts (that is, the districts lost due to the act of nature will not count against the attrition rate). Future considerations may include attrition in multilevel models.

APPENDIX B. EFFECT SIZE COMPUTATIONS

Different types of effect size (ES) indices have been developed for different types of outcome measures, given their distinct statistical properties. The purpose of this appendix is to provide the rationale for the specific computations conducted by the WWC, as well as their underlying assumptions.

A. STUDENT-LEVEL ANALYSES

1. Continuous Outcomes—ES as Standardized Mean Difference (Hedges's *g*)

For continuous outcomes, the WWC has adopted the most commonly used ES index—the standardized mean difference, which is defined as the difference between the mean outcome of the intervention group and the mean outcome of the comparison group divided by the pooled within-group standard deviation (SD) on that outcome measure. Given that the WWC generally focuses on student-level findings, the default SD used in ES computation is the student-level SD.

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

$$g = (X_1 - X_2) / S_{pooled}$$

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

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

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

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

The ES index thus computed is referred to as Hedges's *g*.[11] This index, however, has been shown to be upwardly biased when the sample size is small. Therefore, we have applied a simple

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

correction for this bias developed by Hedges (1981), which produces an unbiased ES estimate by multiplying the Hedges's *g* by a factor of (1 - 3/[4N - 9]), with N being the total sample size. Unless otherwise noted, Hedges's *g* corrected for small-sample bias is the default ES measure for continuous outcomes used in the WWC's review.

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

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

2. Continuous—ES Based on Results from Student-Level *t*-tests or ANOVA

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

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

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

For ESs based on ANOVA F-test results,

$$g = \sqrt{\frac{F(n_1 + n_2)}{n_1 n_2}}$$

3. Continuous—ES Based on Results from Student-Level ANCOVA

Analysis of covariance (ANCOVA) is a commonly used analytic method for quasi-experimental designs. It assesses the effects of an intervention while controlling for important covariates, particularly pretests, which might confound the effects of the intervention. ANCOVA is also used to analyze data from randomized controlled trials so that greater statistical precision of parameter estimates can be achieved through covariate adjustment.

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

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

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

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

A final note about ANCOVA-based ES computation is that Hedges's *g* cannot be computed based on the F-statistic from an ANCOVA. Unlike the F-statistic from an ANOVA, which is based on unadjusted within-group variance, the F-statistic from an ANCOVA is based on covariate-adjusted within-group variance. Hedges's *g*, however, requires the use of unadjusted within-group SD. Therefore, we cannot compute Hedges's *g* with the F-statistic from an ANCOVA in the same way as we can compute it with the F-statistic from an ANOVA. If the pretest-posttest correlation is known, however, we can derive Hedges's *g* from the ANCOVA F-statistic as follows:

$$g = \sqrt{\frac{F(n_1 + n_2)(1 - r^2)}{n_1 n_2}}$$

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

4. Continuous—Difference-in-Differences Approach

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

$$g = \frac{(X_1 - X_{1-pre}) - (X_2 - X_{2-pre})}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}}$$

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

This "difference-in-differences" approach to estimating an intervention's effects while taking into account group difference in pretest is not necessarily optimal, as it is likely to either overestimate or underestimate the adjusted group mean difference, depending on which group performed better on the pretest.[12] Moreover, this approach does not provide a means for adjusting the statistic significance of the adjusted mean difference to reflect the covariance between the pretest and the posttest. Nevertheless, it yields a reasonable estimate of the adjusted group mean difference, which is equivalent to what would have been obtained from a commonly used alternative to the covariate adjustment-based approach to testing an intervention's effect—the analysis of gain scores.

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

The difference-in-differences approach presented earlier also assumes that the pretest-posttest correlation is unknown. In some areas of educational research, however, empirical data on the relationships between pretest and posttest may be available. If such data are dependable, the WWC PIs and the review team in a given topic area may choose to use the empirical relationship to estimate the adjusted group mean difference that is unavailable from the study report or study authors, rather than using the default difference-in-differences approach. The

[12] If the intervention group had a higher average pretest score than the comparison group, the difference-in-difference approach is likely to underestimate the adjusted group mean difference. If the opposite occurs, it is likely to overestimate the adjusted group mean difference.

advantage of doing so is that if, indeed, the empirical relationship between pretest and posttest is dependable, the covariate-adjusted estimates of the intervention's effects will be less biased than those based on the difference-in-differences (gain score) approach. If the PIs and review teams choose to compute ESs using an empirical pretest-posttest relationship, they will need to provide an explicit justification for their choice as well as evidence on the credibility of the empirical relationship. Computationally, if the pretest and posttest have a correlation of *r*, then

$$g = \frac{(X_1 - X_2) - r(X_{1-pre} - X_{2-pre})}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}}$$

5. Dichotomous Outcomes

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

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

$$Odds = \frac{p}{1 - p}$$

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

$$OR = \frac{p_1(1 - p_2)}{p_2(1 - p_1)}$$

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

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

Following conventional practice, the WWC transforms the odds ratio to a logged odds ratio (LOR; that is, the natural log of the odds ratio) to simplify statistical analyses:

$$LOR = \ln(OR)$$

The logged odds ratio has a convenient distribution form, which is approximately normal with a mean of 0 and a SD of π/sqrt(3), or 1.81.

The logged odds ratio can also be expressed as the difference between the logged odds, or logits, for the two groups compared:

$$LOR = \ln(Odds_1) - \ln(Odds_2)$$

which shows more clearly the connection between the logged odds ratio index and the standardized mean difference index (Hedges's *g*) for ESs. To make the logged odds ratio comparable to the standardized mean difference and thus facilitate the synthesis of research findings based on different types of outcomes, researchers have proposed a variety of methods for "standardizing" logged odds ratio. Based on a Monte Carlo simulation study of seven different types of ES indices for dichotomous outcomes, Sanchez-Meca, Marin-Martinez, and Chacon-Moscoso (2003) concluded that the ES index proposed by Cox (1970) is the least biased estimator of the population standardized mean difference, assuming an underlying normal distribution of the outcome. The WWC, therefore, has adopted the Cox index as the default ES measure for dichotomous outcomes. The computation of the Cox index is straightforward:

$$LOR_{Cox} = LOR / 1.65$$

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

B. CLUSTER-LEVEL ANALYSES

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

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

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

1. Computing Student-Level ESs for Studies with Cluster-Level Analyses

For studies that reported findings from only cluster-level analyses, it might be tempting to compute ESs using cluster-level means and SDs. This, however, is not appropriate for the purpose of the WWC reviews for at least two reasons. First, because cluster-level SDs are typically much smaller than student-level SDs,[13] ESs based on cluster-level SDs will be much larger than and, therefore, incomparable with student-level ESs that are the focus of WWC reviews. Second, the criterion for "substantively important" effects in the WWC Intervention Rating Scheme (ES of at least 0.25) was established specifically for student-level ESs and does not apply to cluster-level ESs. Moreover, there is not enough knowledge in the field as yet for judging the magnitude of cluster-level effects. A criterion of "substantively important" effects for cluster-level ESs, therefore, cannot be developed for intervention rating purposes. An intervention rating of potentially positive effects based on a cluster-level ES of 0.25 or greater (that is, the criterion for student-level ESs) would be misleading.

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

[13] Cluster-level SD = (student-level SD)*sqrt(ICC).

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

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

It is clear from this discussion that in most cases, requesting student-level data, particularly student-level SDs, from the study authors will be the only way that allows us to compute the student-level ESs for studies reporting only cluster-level findings. If the study authors cannot provide the student-level data needed, then we will not be able to compute the student-level ESs. Nevertheless, such studies will not be automatically excluded from the WWC reviews; they could still potentially contribute to intervention ratings as explained in the next section.

2. Handling Studies with Cluster-Level Analyses if Student-Level ESs Cannot Be Computed

A study's contribution to the effectiveness rating of an intervention depends mainly on three factors: (1) the quality of the study design, (2) the statistical significance of the findings, and (3) the effect size(s). For studies that report only cluster-level findings, the quality of their designs is not affected by whether student-level ESs could be computed. Such studies could still meet WWC evidence standards with or without reservations and be included in intervention reports even if student-level ESs were not available.

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

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

As explained in the previous section, multilevel analysis is generally considered the preferred method for analyzing data from studies with cluster-level assignment. With recent methodological advances, multilevel analysis has gained increased popularity in education and other social science fields. More and more researchers have begun to employ the hierarchical linear modeling (HLM) method to analyze data of a nested nature (for example, students nested within classes and classes nested within schools) (Raudenbush & Bryk, 2002).[14] Similar to student-level ANCOVA, HLM can also adjust for important covariates such as pretest when estimating an intervention's effect. Unlike student-level ANCOVA that assumes independence of observations, however, HLM explicitly takes into account the dependence among members within the same higher-level unit (for example, the dependence among students within the same class). Therefore, the parameter estimates, particularly the standard errors, generated from HLM are less biased than those generated from ANCOVA when the data have a multilevel structure.

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

$$g = \frac{\gamma}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{(n_1 + n_2 - 2)}}}$$

where γ is the HLM coefficient for the intervention's effect, which represents the group mean difference adjusted for both level-1 and level-2 covariates, if any; n_1 and n_2 are the student sample sizes; and S_1 and S_2 are the student-level unadjusted posttest SD for the intervention group and the comparison group, respectively.[15]

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

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

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

tends to be smaller than the conventional unadjusted pooled within-group variance, and it would thus lead to an overestimate of the ES if used in the denominator of the ES.

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

APPENDIX C. CLUSTERING CORRECTION OF THE STATISTICAL SIGNIFICANCE OF EFFECTS ESTIMATED WITH MISMATCHED ANALYSES

In order to assess an intervention's effects adequately, it is important to know not only the magnitude of the effects as indicated by the ES, but also the statistical significance of the effects. The correct statistical significance of findings, however, 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 (for example, classroom or school level), but the analysis was conducted at the student level, ignoring the dependence among students within the same clusters. Although the point estimates of the intervention's effects based on such mismatched analyses are unbiased, the standard errors of the effect estimates are likely to be underestimated, which would lead to inflated Type I error and overestimated statistical significance.

In order to present a fair judgment about an intervention's effects, the WWC computes clustering-corrected statistical significance for effects estimated from mismatched analyses and the corresponding domain average effects based on Hedges's (2005) most recent work. As 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 ES that ignores clustering and then to correct both the t-statistic and the associated degrees of freedom for clustering based on sample sizes, number of clusters, and the intra-class correlation (ICC). The statistic significance corrected for clustering could then be obtained from the t-distribution with the corrected t-statistic and degrees of freedom. In the remainder of this section, we detail each step of the process.

Compute the t-statistic for the ES ignoring clustering:

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

where g is the ES that ignores clustering, and n_1 and n_2 are the sample sizes for the intervention group and the comparison group, respectively, for a given outcome. For domain average ESs, n_1 and n_2 are the average sample sizes for the intervention and comparison groups, respectively, across all outcomes within the domain.

Correct the t-statistic for clustering:

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

where N is the total sample size at the student level ($N = n_1 + n_2$), m is the total number of clusters in the intervention and comparison groups ($m = m1 + m2$, $m1$ and $m2$ are the number of clusters in each of the two groups), and ρ is the ICC for a given outcome.

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

For domain average ESs, the ICC used earlier 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 (that is, the largest $m1$ and $m2$ across outcomes). This gives the study the benefit of the doubt by crediting the measure with the most statistical power, so that the WWC's rating of interventions will not be unduly conservative.

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

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

Obtain the statistical significance of the effect corrected for clustering:

The clustering-corrected statistical significance (p-value) is determined based on the t-distribution with the corrected t-statistic (t_A) and the corrected degrees of freedom (h). This p-value can be either 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, h, 2).

APPENDIX D. BENJAMINI-HOCHBERG CORRECTION OF THE STATISTICAL SIGNIFICANCE OF EFFECTS ESTIMATED WITH MULTIPLE COMPARISONS

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

As is the case with clustering correction, the WWC applies the BH correction only to statistically significant findings, because nonsignificant findings will remain nonsignificant after correction. For findings based on analyses in which the unit of analysis was properly aligned with the unit of assignment, we use the p-values reported in the study for the BH correction. If the exact p-values were not available, but the ESs could be computed, we would convert the ESs to t-statistics and then obtain the corresponding p-values.[16] For findings based on mismatched analyses, we first correct the author-reported p-values for clustering and then use the clustering-corrected p-values for the BH correction.

Although the BH correction procedure just described was originally developed under the assumption of independent test statistics (Benjamini & Hochberg, 1995), Benjamini and Yekutieli (2001) point out that it also applies to situations in which the test statistics have positive dependency, and that the condition for positive dependency is general enough to cover many problems of practical interest. For other forms of dependency, a modification of the original BH procedure could be made, which, however, is "very often not needed, and yields too conservative a procedure" (p. 1183).[17] Therefore, the WWC has chosen to use the original BH procedure rather than its more conservative modified version as the default approach to correcting for multiple comparisons.

In the remainder of this section, we describe the specific procedures for applying the BH correction in three types of situations: studies that tested multiple outcome measures in the same outcome domain with a single comparison group, studies that tested a given outcome measure

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

[17] The modified version of the BH procedure uses α over the sum of the inverse of the p-value ranks across the m comparisons instead of α .

with multiple comparison groups, and studies that tested multiple outcome measures in the same outcome domain with multiple comparison groups.

A. BENJAMINI-HOCHBERG CORRECTION OF THE STATISTICAL SIGNIFICANCE OF EFFECTS ON MULTIPLE OUTCOME MEASURES WITHIN THE SAME OUTCOME DOMAIN TESTED WITH A SINGLE COMPARISON GROUPS

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

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

For each p-value (pi), compute:

$$p_i' = \frac{i\alpha}{M}$$

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

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

Identify the largest i—denoted by k—that satisfies the condition: $pi \leq pi'$*. This establishes the cutoff point and allows us to conclude that all findings with p-values smaller than or equal to pk are statistically significant, and findings with p-values greater than pk are not significant at*

the prespecified level of significance (α = 0.05 by default) after correction for multiple comparisons.

One thing to note is that unlike clustering correction, which produces a new p-value for each corrected finding, the BH correction does not generate a new p-value for each finding but rather 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 and reported six statistically significant effects and two nonsignificant effects based on properly aligned analyses. To correct the significance of the findings for multiple comparisons, we would first rank order the p-values of the six author-reported significant findings in the first column of Table D1 and list the p-value ranks in the second column. We then compute $p' = i * \alpha / M$ with $M = 8$ and $\alpha = 0.05$ and record the values in the third column. Next, we identify *k*, the largest *i*, that meets the condition: $pi \leq pi'$. In this example, $k = 4$, and $pk = 0.014$. Thus, we can claim that the four findings associated with a p-value of 0.014 or smaller are statistically significant at the 0.05 level after correction for multiple comparisons. The other two findings, although reported as being statistically significant, are no longer significant after the correction.

TABLE D1

AN ILLUSTRATION OF APPLYING THE BENJAMINI-HOCHBERG CORRECTION FOR MULTIPLE COMPARISONS

Author-reported or clustering-corrected p-value (*pi*)	P-value rank (*i*)	$pi' = i * 0.05/8$	$pi \leq pi'$?	Statistical significance after BH correction (α = .05)
0.002	1	0.006	Yes	significant
0.009	2	0.013	Yes	significant
0.011	3	0.019	Yes	significant
0.014	**4**	**0.025**	**Yes**	**significant**
0.034	5	0.031	No	n.s.
0.041	6	0.038	No	n.s.

Note: n.s. = not statistically significant.

B. BENJAMINI-HOCHBERG CORRECTION OF THE STATISTICAL SIGNIFICANCE OF EFFECTS ON A GIVEN OUTCOME TESTED WITH MULTIPLE COMPARISON GROUPS

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

1. In consultation with the PI and the study authors if needed, the review team selects a single comparison group that best represented the "business as usual" condition or that is considered most relevant to the WWC's review. Only findings based on comparisons between the intervention group and this particular comparison group would be included in the WWC's review. Findings involving the other comparison groups would be ignored, and the multiplicity due to one intervention group being compared with multiple comparison groups would also be ignored.

2. If the PI and the review team believe that it is appropriate to combine the multiple comparison groups, and if adequate data are available for deriving the means and SDs of the combined group, the team may present the findings based on comparisons of the intervention group and the combined comparison group instead of findings based on comparisons of the intervention group and each individual comparison group. The kind of multiplicity due to one intervention group being compared with multiple comparison groups would no longer be an issue in this approach.

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

3. If the PI and the review team believe that neither of these two options is appropriate for a particular study, and that findings from comparisons of the intervention group and each individual comparison group should be presented, they need to make sure that the findings presented in the WWC's intervention report are corrected for multiplicity due to multiple comparison groups if necessary. The review team needs to check the study report or check with the study authors to determine whether the comparisons of the multiple groups were based on a proper statistical test that already took multiplicity into account (for example, Dunnett's test [Dunnett, 1955], the Bonferroni method [Bonferroni, 1935], Scheffe's test [1953], and Tukey's HSD test [1949]). If so, then there would be no need for further corrections. It is also advisable for the team to check with the study authors regarding the appropriateness of correcting its findings for multiplicity due to multiple comparison groups, as the authors might have theoretical or empirical concerns about considering the findings from comparisons of the intervention group and a given comparison group without consideration of other comparisons made within the same study. If the team decides that multiplicity correction is necessary, it will apply such correction using the BH method in the same way as it would apply the method to findings on multiple outcomes within the same domain tested with a single comparison group as described in the previous section.

C. BENJAMINI-HOCHBERG CORRECTION OF THE STATISTICAL SIGNIFICANCE OF EFFECTS ON MULTIPLE OUTCOME MEASURES IN THE SAME OUTCOME DOMAIN TESTED WITH MULTIPLE COMPARISON GROUPS

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

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

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

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

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

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

Although this scenario may be relatively rare, it is possible that the study authors performed a proper multivariate test (for example, MANOVA or MANCOVA) to compare the intervention group with a given comparison group that took into account the multiplicity due to multiple outcomes and performed separate multivariate tests for different comparison groups. For such studies, the review team will need to correct only the findings for multiplicity due to multiple comparison groups. Specifically, separate BH corrections will be made to the findings on different outcomes. With five outcomes and two comparison groups, for instance, the review team will apply the BH correction to the five sets of findings separately—one set of findings (one finding for each comparison group) for each outcome measure.

The decision rules for these three scenarios described are summarized in Table D2.

TABLE D2

DECISION RULES FOR CORRECTING THE SIGNIFICANCE LEVELS OF FINDINGS FROM STUDIES THAT HAD A MULTIPLE COMPARISON PROBLEM DUE TO MULTIPLE OUTCOMES IN A GIVEN DOMAIN AND/OR MULTIPLE COMPARISON GROUPS, BY SCENARIO

Authors' Analyses	Benjamini-Hochberg Correction
1. Did not correct for multiplicity from any source	• BH correction to all 10 individual findings
2. Corrected for multiplicity due to multiple comparison groups only	• BH correction to the 5 findings based on T vs. C1 comparisons • BH correction to the 5 findings based on T vs. C2 comparisons
3. Corrected for multiplicity due to multiple outcomes only	• BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O1 • BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O2 • BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O3 • BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O4 • BH correction to the 2 findings based on T vs. C1 and T vs. C2 comparisons on O5

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

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

APPENDIX E. PILOT STANDARDS FOR REGRESSION DISCONTINUITY DESIGNS

Regression discontinuity (RD) designs are increasingly used by researchers to obtain unbiased estimates of the effects of education-related interventions. These designs are applicable when a continuous "scoring" rule is used to assign the intervention to study units (for example, school districts, schools, or students). Units with scores below a pre-set cutoff value are assigned to the treatment group and units with scores above the cutoff value are assigned to the comparison group, or vice versa. For example, students may be assigned to a summer school program if they score below a preset point on a standardized test, or schools may be awarded a grant based on their score on an application.

Under an RD design, the effect on an intervention can be estimated as the difference in mean outcomes between treatment and comparison group units, adjusting statistically for the relationship between the outcomes and the variable used to assign units to the intervention, typically referred to as the "forcing" or "assignment" variable. A regression line (or curve) is estimated for the treatment group and similarly for the comparison group, and the difference in average outcomes between these regression lines at the cutoff value of the forcing variable is the estimate of the effect of the intervention. Stated differently, an effect occurs if there is a "discontinuity" in the two regression lines at the cutoff. This estimate pertains to average treatment effects for units right at the cutoff. RD designs generate unbiased estimates of the effect of an intervention if (1) the relationship between the outcome and forcing variable can be modeled correctly and (2) the forcing variable was not manipulated to influence treatment assignments.

This document presents criteria under which RD designs *Meet WWC Evidence Standards* and *Meet WWC Evidence Standards with Reservations*.

Assessing Whether a Study Qualifies as an RD Study

A study qualifies as an RD study if it meets *all* of the following criteria:

- ***Treatment assignments are based on a forcing variable; units with scores at or above (or below) a cutoff value are assigned to the treatment group while units with scores on the other side of the cutoff are assigned to the comparison group.*** For example, an evaluation of a tutoring program could be classified as an RD study if students with a reading test score at or below 30 are admitted to the program and students with a reading test score above 30 are not. As another example, a study examining the impacts of grants to improve teacher training in local areas could be considered an RD study if grants are awarded to only those sites with grant application scores that are at least 70. In some instances, RD studies may use multiple criteria to assign the treatment to study units. For example, a student may be assigned to an after-school program if the student's reading score is below 30 or math score is below 40.[18] As with RCTs, noncompliance with treatment assignment is permitted,

[18] For ease of exposition, the remainder of this document will refer to one cutoff.

but the study must still meet the criteria below to meet evidence standards. Two additional criteria for the forcing variable are:

- ***The forcing variable must be ordinal with a sufficient number of unique values.*** This condition is required to model the relationship between the outcomes and forcing variable. The forcing variable should never be based on cardinal (non-ordinal) categories (like gender or race). The analyzed data must also include at least four unique values of the forcing variable below the cutoff and four unique values above the cutoff.
- ***There must be no factor confounded with the forcing variable.*** The cutoff value for the forcing variable must not be used to assign students to interventions other than the one being tested. For example, free/reduced-price lunch (FRPL) status cannot be the basis of an RD design, because FRPL is used as the eligibility criteria for a wide variety of services. This criterion is necessary to ensure that the study can isolate the causal effects of the tested intervention from the effects of other interventions.

If a study claims to be based on an RD design, but does not have these properties, the study does not meet standards as an RD design.

Possible Designations for Studies Using RD Designs

Once a study is determined to be an RD design, the study can receive one of three designations based on the set of criteria described below:

4. ***Meets Evidence Standards.*** To qualify, a study must meet each of the four individual standards listed below without reservations.
5. ***Meets Evidence Standards with Reservations.*** To qualify, a study must meet standards 1, 2, and 4, with or without reservations.
6. ***Does Not Meet Evidence Standards.*** If a study fails to meet standard 1, 2, or 4.

Standard 1: Integrity of the Forcing Variable

A key condition for an RD design to produce unbiased estimates of effects of an intervention is that there was no systematic manipulation of the forcing variable. This situation is analogous to the non-random manipulation of treatment and control group assignments under an RCT. In an RD design, manipulation means that scores for some units were systematically changed from their true values to influence treatment assignments. With nonrandom manipulation, the true relationship between the outcome and forcing variable can no longer be identified, which could lead to biased impact estimates.

Manipulation is possible if "scorers" have knowledge of the cutoff value and have incentives to change unit-level scores to ensure that some units are assigned to a specific research condition. Stated differently, manipulation could occur if the scoring and treatment assignment processes are not independent. It is important to note that manipulation of the forcing variable is *different*

than treatment status noncompliance (which occurs if some treatment group members do not receive intervention services or some comparison group members receive embargoed services).

The likelihood of manipulation will depend on the nature of the forcing variable, the intervention, and the study design. For example, manipulation is likely to be less plausible if the forcing variable is a standardized test score than if it is a student assessment conducted by teachers who also have input into treatment assignment decisions. As another example, manipulation is unlikely if the researchers themselves determined the cutoff value using an existing forcing variable (for example, a score from a test that was administered prior to the implementation of the study).

In all RD studies, the integrity of the forcing variable should be established both institutionally and statistically.

> **Criterion A.** The institutional integrity of the forcing variable should be established by an adequate description of the scoring and treatment assignment process. This description should indicate the forcing variable used, the cutoff value that was selected, who selected the cutoff (for example: researchers, school personnel, curriculum developers), who determined values of the forcing variable (for example, who scored a test), and when the cutoff was selected relative to determining the values of the forcing variable. This description must show that manipulation was unlikely because scorers had little opportunity or little incentive to change "true" scores in order to allow or deny specific individuals access to the intervention. If there is both a clear opportunity to manipulate scores and a clear incentive (for example, in an evaluation of a math curriculum if a placement test is scored by the curriculum developer after the cutoff is known) then the study does not satisfy this standard.
>
> **Criterion B.** The statistical integrity of the forcing variable should be demonstrated by using statistical tests found in the literature or a graphical analysis to establish the smoothness of the density of the forcing variable right around the cutoff. This is important to establish because there may be incentives for scorers to manipulate scores to make units just eligible for the treatment group (in which case, there may be an unusual mass of units near the cutoff). If a statistical test is provided, it should fail to reject the null hypothesis of continuity in the density of the forcing variable. If a graphical analysis is provided (such as a histogram or other type of density plot), there should not be strong evidence of a discontinuity at the cutoff that is obviously larger than discontinuities in the density at other points (some small discontinuities may arise when the forcing variable is discrete). If both are provided then the statistical test will take precedence, unless the statistical test indicates no discontinuity but the graphical analysis provides very strong evidence to the contrary.

To meet this standard without reservations, both criteria must be satisfied.

To meet this standard with reservations, one of the two criteria must be satisfied.

A study fails this standard if neither criterion is satisfied.

Standard 2: Attrition

An RD study must report the number of students (teachers, schools, etc.) who were assigned to the treatment and comparison group samples, and the proportion of students (teachers, schools, etc.) with outcome data who were included in the impact analysis (that is, response rates). Both overall attrition and attrition by treatment status must be reported.

To meet this standard without reservations, an RD study must meet the WWC randomized control trial (RCT) standards for attrition. The study authors can calculate overall and differential attrition either for the entire research sample or for only students near the cutoff value of the forcing variable.

A study fails this standard if attrition information is not available or if the above conditions are not met. A study that fails this standard could potentially be reviewed as a QED if equivalence is established on key baseline covariates (in this case, the forcing variable is not exempt from the equivalence requirement, described below).

Standard 3: Continuity of the Outcome-Forcing Variable Relationship

To obtain a rigorous impact estimate of a key outcome under an RD design, there must be strong evidence that in the absence of the intervention, there would be a smooth relationship between the outcome and the forcing variable at the cutoff score. This condition is needed to ensure that any observed discontinuity in the outcomes of treatment and comparison group units at the cutoff can be attributable to the intervention.

This smoothness condition cannot be checked directly, although there are two indirect approaches that should be used. The first approach is to test whether, conditional on the forcing variable, key *baseline* covariates that are correlated with the outcome variable (as identified in the review protocol for the purpose of establishing equivalence) are continuous at the cutoff. This means that the intervention should have no "impact" on baseline covariates at the cutoff. Particularly important baseline covariates for this analysis are pre-intervention measures of the key outcome variables (for example, pretests). This requirement is waived for any key covariate that is used as the RD forcing variable.

The second approach for assessing the smoothness condition is to use statistical tests or graphical analyses to examine whether there are discontinuities in the outcome-forcing variable relationship at values away from the cutoff. This involves testing for "impacts" at values of the forcing variable where there should be no impacts, such as the medians of points above or below the cutoff value (Imbens and Lemieux 2008). The presence of such discontinuities (impacts) would imply that the relationship between the outcome and the forcing variable at the cutoff may not be truly continuous, suggesting that observed impacts at the cutoff may not be due to the intervention.

Two criteria determine whether a study meets this standard.

> **Criterion A.** Baseline (or pre-baseline) equivalence on key covariates (as identified in the review protocol) should be demonstrated at the cutoff value of the forcing variable. This involves calculating an impact at the cutoff on the covariate of interest. This

requirement is waived if the variable on which equivalence must be established is the forcing variable (for example, a baseline test score).

Criterion B. There should be no evidence (using statistical tests or graphical analyses) of an unexplainable discontinuity in the outcome-score relationship at score values other than at the cutoff value. An example of an "explainable" discontinuity is one that corresponds to some other known intervention that was also administered using the same forcing variable but with a different cutoff value.

To meet this standard without reservations, both criteria must be satisfied. If criterion A is waived (see above), it can be regarded as satisfied.

A study fails this standard if either criterion is not satisfied. If criterion A is waived (see above), it can be regarded as satisfied.

Standard 4: Functional Form and Bandwidth

Unlike with RCTs, statistical modeling plays a central role in estimating impacts in an RD study. The most critical aspects of the statistical modeling are (1) the functional form specification of the relationship between the outcome and the forcing variable, and (2) the appropriate range of forcing variable values for selecting the sample (that is, the *bandwidth* around the cutoff value). Five criteria determine whether a study meets this standard.

Criterion A. The average treatment effect for an outcome must be estimated using a statistical model that controls for the forcing variable. Other baseline covariates may also be included in the statistical models, though they are not required. For both bias and variance considerations, it is never acceptable to estimate an impact by comparing the mean outcomes of treatment and comparison group members without adjusting for the forcing variable (even if there is a weak relationship between the outcome and forcing variable).

Criterion B. A graphical analysis displaying the relationship between the outcome and forcing variable—including a scatter plot and a fitted curve—must be included in the report. The display must be consistent with the choice of bandwidth and the functional form specification for the analysis. For example, if the graphical analysis shows a nonlinear relationship between the outcome and the forcing variable, then the functional form of the impact regression should also be nonlinear, or the bandwidth should be restricted to the range of data that is approximately linear on either side of the cutoff. One way to assess whether the bandwidth or functional form was appropriately chosen is to measure the sensitivity of impacts to the inclusion of observations in the tails of the forcing variable distribution.

Criterion C. Evidence must be provided that an appropriate parametric, semi-parametric, or nonparametric model was fit to the data. For a parametric approach, the adopted functional form (for example, a polynomial specification) must be shown to be the best fit to the data using statistical significance of higher order terms or a recognized "best fit" criterion (for example, the polynomial degree could be chosen to minimize the Akaike Information Criteria). Alternatively, a local regression or related nonparametric

approach can be used, where the chosen bandwidth is justified using an approach such as cross-validation (or other similar approaches found in the literature). In the event that competing models are plausible, evidence of the robustness of impact findings to alternative model specifications should be provided.

Criterion D. If the estimate of the relationship between the outcome and the forcing variable is constrained to be the same on both sides of the cutoff (for example, a line that is constrained to have the same slope on both sides of the cutoff), then empirical support (either a statistical test or graphical evidence) for that constraint must be provided.

Criterion E. If the reported impact is an average of impacts across multiple sites (where, for example, a different cutoff or forcing variable is used in each site), each site impact should be estimated separately. The model used in each site should be justified using the criteria discussed above.

To meet this standard without reservations, all five of the criteria must be satisfied.

To meet this standard with reservations, Criteria A and D must be satisfied. In addition either B or C must also be satisfied.

A study fails this standard if Criterion A is not satisfied, or criterion D is not satisfied, or if both criteria B and C are not satisfied.

Reporting Requirement

Truly continuous forcing variables are likely to be rare in education studies. For example, test scores are not truly continuous—they often have a finite number of unique values because every test has a finite number of questions. If a forcing variable has a very small number of unique values (for example, a letter grade on an A-F scale) then it is not possible to estimate the relationship between the outcome and the forcing variable. Thus, we require at least 4 categories above and below the cutoff for a study to be eligible for review as an RD design. However, even in cases with a larger (but still discrete) number of unique values of the forcing variable standard errors must be estimated appropriately to account for the clustering of students at unique values of the forcing variable (see Lee and Card 2008).

As is the case in RCT designs, clustering of students should not cause biased estimates of the impact of the intervention, so if study authors do not appropriately account for the clustering of students, a study can still meet WWC standards if it meets the standards described above. However, since the statistical significance of findings is used for the rating of the effectiveness of an intervention, study authors must account for clustering using an appropriate method (for example, the method proposed in Lee and Card 2008) in order for findings reported by the author to be included in the rating of effectiveness. If the authors do not account for clustering, then the WWC will not rely on the statistical significance of the findings from the study. However, the findings can still be included as "substantively important" if the effect size is 0.25 standard deviation or greater.

Study authors may also demonstrate that clustering of students into unique test score values does not require adjustments in the calculation of standards errors. This can be done by showing that the forcing variable is continuous around the cutoff and there is no clustering of observation around specific scores.

References

Imbens, G. and T. Lemieux (2008). Regression Discontinuity Designs: A Guide to Practice. *Journal of Econometrics* 142 (2), 615-635.

Lee, David and David Card (2008). Regression Discontinuity Inference With Specification Error. *Journal of Econometrics* 142 (2), 655-674.

APPENDIX F. PILOT STANDARDS FOR SINGLE-CASE DESIGNS

In an effort to expand the pool of scientific evidence available for review, the What Works Clearinghouse (WWC) assembled a panel of national experts in single-case design (SCD) and analysis to draft SCD Standards. In this paper, the panel provides an overview of SCDs, specifies the types of questions that SCDs are designed to answer, and discusses the internal validity of SCDs. The panel then proposes SCD Standards to be implemented by the WWC. The Standards are bifurcated into Design and Evidence Standards (see Figure F1). The Design Standards evaluate the internal validity of the design. Reviewers assign the categories of *Meets Standards, Meets Standards with Reservations* and *Does not Meet Standards* to each study based on the Design Standards. Reviewers trained in visual analysis will then apply the Evidence Standards to studies that meet standards (with or without reservations), resulting in the categorization of each outcome variable as demonstrating *Strong Evidence, Moderate Evidence, or No Evidence.*

A. OVERVIEW OF SINGLE-CASE DESIGNS

SCDs are adaptations of interrupted time-series designs and can provide a rigorous experimental evaluation of intervention effects (Horner & Spaulding, in press; Kazdin, 1982, in press; Kratochwill, 1978; Kratochwill & Levin, 1992; Shadish, Cook, & Campbell, 2002). Although the basic SCD has many variations, these designs often involve repeated, systematic measurement of a dependent variable before, during, and after the active manipulation of an independent variable (e.g., applying an intervention). SCDs can provide a strong basis for establishing causal inference, and these designs are widely used in applied and clinical disciplines in psychology and education, such as school psychology and the field of special education.

SCDs are identified by the following features:

- An individual "case" is the unit of intervention and unit of data analysis (Kratochwill & Levin, in press). A case may be a single participant or a cluster of participants (e.g., a classroom or a community).
- Within the design, the case provides its own control for purposes of comparison. For example, the case's series of outcome variables are measured prior to the intervention and compared with measurements taken during (and after) the intervention.
- The outcome variable is measured repeatedly within and across different conditions or levels of the independent variable. These different conditions are referred to as phases (e.g., baseline phase, intervention phase).

As experimental designs, a central goal of SCDs is to determine whether a causal relation (i.e., functional relation) exists between the introduction of a researcher-manipulated independent variable (i.e., an intervention) and change in a dependent (i.e., outcome) variable (Horner & Spaulding, in press; Levin, O'Donnell, & Kratochwill, 2003). Experimental control

involves replication of the intervention in the experiment and this replication is addressed with one of the following methods (Horner, et al., 2005):

- Introduction and withdrawal (i.e., reversal) of the independent variable (e.g., ABAB design)
- Iterative manipulation of the independent variable across different observational phases (e.g., alternating treatments design)
- Staggered introduction of the independent variable across different points in time (e.g., multiple baseline design)

SCDs have many variants. Although flexible and adaptive, a SCD is shaped by its research question(s) and objective(s) which must be defined with precision, taking into consideration the specifics of the independent variable tailored to the case(s), setting(s), and the desired outcome(s) (i.e., a primary dependent variable). For example, if the dependent variable is unlikely to be reversed after responding to the initial intervention, then an ABAB reversal design would not be appropriate, whereas a multiple baseline design across cases would be appropriate. Therefore, the research question generally drives the selection of an appropriate SCD.

B. CAUSAL QUESTIONS THAT SCDS ARE DESIGNED TO ANSWER

The goal of a SCD is usually to answer "Is this intervention more effective than the current "baseline" or "business-as-usual" condition?" SCDs are particularly appropriate for understanding the responses of one or more cases to an intervention under specific conditions (Horner & Spaulding, in press). SCDs are implemented when pursuing the following research objectives (Horner et al., 2005):

- Determining whether a causal relation exists between the introduction of an independent variable and a change in the dependent variable. For example, a research question might be "Does Intervention B reduce a problem behavior for this case (or these cases)?"
- Evaluating the effect of altering a component of a multi-component independent variable on a dependent variable. For example, a research question might be "Does adding Intervention C to Intervention B further reduce a problem behavior for this case (or these cases)?"
- Evaluating the relative effects of two or more independent variables (e.g., alternating treatments) on a dependent variable. For example, a research question might be "Is Intervention B or Intervention C more effective in reducing a problem behavior for this case (or these cases)?"

SCDs are especially appropriate for pursuing research questions in applied and clinical fields. This application is largely because disorders with low prevalence may be difficult to study with traditional group designs that require a large number of participants for adequate statistical power (Odom, et al., 2005). Further, in group designs, the particulars of who responded to an intervention under which conditions might be obscured when reporting only group means and associated effect sizes (Horner et al. 2005). SCDs afford the researcher an opportunity to provide detailed documentation of the characteristics of those cases that *did* respond to an intervention and those that *did not* (i.e., nonresponders). For this reason, the panel recommends that What Works Clearinghouse (WWC) reviewers systematically specify the conditions under which an intervention is and is not effective for cases being considered, if this information is available in the research report.

Because the underlying goal of SCDs is most often to determine "Which intervention is effective for this case (or these cases)?" the designs are intentionally flexible and adaptive. For example, if a participant is not responding to an intervention, then the independent variables can be manipulated while continuing to assess the dependent variable (Horner et al., 2005). Because of the adaptive nature of SCD designs, nonresponders might ultimately be considered "responders" under particular conditions.[19] In this regard, SCDs provide a window into the process of participant change. SCDs can also be flexible in terms of lengthening the number of data points collected during a phase to promote a stable set of observations, and this feature may provide additional insight into participant change.

C. THREATS TO INTERNAL VALIDITY IN SINGLE-CASE DESIGN[20]

Similar to group randomized controlled trial designs, SCDs are structured to address major threats to internal validity in the experiment. Internal validity in SCDs can be improved through replication and/or randomization (Kratochwill & Levin, in press). Although it is possible to use randomization in structuring experimental SCDs, these applications are still rare. Unlike most randomized controlled trial group intervention designs, most single-case researchers have addressed internal validity concerns through the structure of the design and systematic *replication of the effect* within the course of the experiment (e.g., Hersen & Barlow, 1976; Horner et al., 2005; Kazdin, 1982; Kratochwill, 1978; Kratochwill & Levin, 1992). The former (design structure, discussed in the *Standards* as "Criteria for Designs…") can be referred to as "methodological soundness" and the latter (effect replication, discussed in the *Standards* as "Criteria for Demonstrating Evidence…") is a part of what can be called "evidence credibility" (see, for example, Kratochwill & Levin, in press).

In SCD research, effect replication is an important mechanism for controlling threats to internal validity and its role is central for each of the various threats discussed below. In fact, the replication criterion discussed by Horner et al. (2005, p. 168) represents a fundamental characteristic of SCDs: "In most [instances] experimental control is demonstrated when the

[19] WWC Principal Investigators (PIs) will need to consider whether variants of interventions constitute distinct interventions. Distinct interventions will be evaluated individually with the SCD Standards. For example, if the independent variable is changed during the course of the study, then the researcher must begin the replication series again to meet the design standards.

[20] Prepared by Thomas Kratochwill with input from Joel Levin, Robert Horner, and William Shadish.

design documents *three* demonstrations of the experimental effect at *three* different points in time with a single case (within-case replication), or across different cases (inter-case replication) (emphasis added)." As these authors note, an experimental effect is demonstrated when the predicted changes in the dependent measures covary with manipulation of the independent variable. This criterion of three replications has been included in the *Standards* for designs to "meet evidence" standards. Currently, there is no formal basis for the "three demonstrations" recommendation; rather, it represents a conceptual norm in published articles, research, and textbooks that recommend methodological standards for single-case experimental designs (Kratochwill & Levin, in press).

Important to note are the terms level, trend and variability. "Level" refers to the mean score for the data within a phase. "Trend" refers to the slope of the best-fitting straight line for the data within a phase, and "variability" refers to the fluctuation of the data (as reflected by the data's range or standard deviation) around the mean. See pages 17-20 for greater detail.

Table F1, adapted from Hayes (1981) but without including the original "design type" designations, presents the three major types of SCDs and their variations. In AB designs, a case's performance is measured within each condition of the investigation and compared between or among conditions. In the most basic two-phase AB design, the A condition is a baseline or preintervention series/phase and the B condition is an intervention series/phase. It is difficult to draw valid causal inferences from traditional two-phase AB designs because the lack of replication in such designs makes it more difficult to rule out alternative explanations for the observed effect (Kratochwill & Levin, in press). Furthermore, repeating an AB design across several cases in separate or independent studies would typically not allow for drawing valid inferences from the data (Note: this differs from multiple baseline designs, described below, which introduce the intervention at different points in time). The *Standards* require a minimum of four A and B phases, such as the ABAB design.

There are three major classes of SCD that incorporate phase repetition, each of which can accommodate some form of randomization to strengthen the researcher's ability to draw valid causal inferences (see Kratochwill & Levin, in press, for discussion of such randomization applications). These design types include the ABAB design (as well as the changing criterion design, which is considered a variant of the ABAB design), the multiple baseline design, and the alternating treatments design. Valid inferences associated with the ABAB design are tied to the design's structured repetition. The phase repetition occurs initially during the first B phase, again in the second A phase, and finally in the return to the second B phase (Horner et al., 2005). This design and its effect replication standard can be extended to multiple repetitions of the treatment (e.g., ABABABAB) and might include multiple treatments in combination that are introduced in a repetition sequence as, for example, A/(B+C)/A/(B+C)/A (see Table F1). In the case of the changing criterion design, the researcher begins with a baseline phase and then schedules a series of criterion changes or shifts that set a standard for participant performance over time. The criteria are typically pre-selected and change is documented by outcome measures changing with the criterion shifts over the course of the experiment.

TABLE F1

EXAMPLE SINGLE-CASE DESIGNS AND ASSOCIATED CHARACTERISTICS

Representative Example Designs	Characteristics
Simple phase change designs [e.g., ABAB; BCBC and the changing criterion design].* (In the literature, ABAB designs are sometimes referred to as withdrawal designs, intrasubject replication designs, or reversal designs)	In these designs, estimates of level, trend, and variability within a data series are assessed under similar conditions; the manipulated variable is introduced and concomitant changes in the outcome measure(s) are assessed in the level, trend, and variability between phases of the series, with special attention to the degree of overlap, immediacy of effect, and similarity of data patterns in similar phases (e.g., all baseline phases).
Complex phase change [e.g., interaction element: B(B+C)B; C(B+C)C]	In these designs, estimates of level, trend, and variability in a data series are assessed on measures within specific conditions and across time.
Changing criterion design	In this design the researcher examines the outcome measure to determine if it covaries with changing criteria that are scheduled in a series of predetermined steps within the experiment. An A phase is followed by a series of B phases (e.g., B1, B2, B3…BT), with the Bs implemented with criterion levels set for specified changes. Changes/ differences in the outcome measure(s) are assessed by comparing the series associated with the changing criteria.
Alternating treatments (In the literature, alternating treatment designs are sometimes referred to as part of a class of multi-element designs)	In these designs, estimates of level, trend, and variability in a data series are assessed on measures within specific conditions and across time. Changes/differences in the outcome measure(s) are assessed by comparing the series associated with different conditions.
Simultaneous treatments (in the literature simultaneous treatment designs are sometimes referred to as concurrent schedule designs).	In these designs, estimates of level, trend, and variability in a data series are assessed on measures within specific conditions and across time. Changes/differences in the outcome measure(s) are assessed by comparing the series across conditions.
Multiple baseline (e.g., across cases, across behaviors, across situations)	In these designs, multiple AB data series are compared and introduction of the intervention is staggered across time. Comparisons are made both between and within a data series. Repetitions of a single simple phase change are scheduled, each with a new series and in which both the length and timing of the phase change differ across replications.

Source: Adapted from Hayes (1981) and Kratochwill & Levin (in press). To be reproduced with permission.

* A represents a baseline series; "B" and "C" represent two different intervention series.

Another variation of SCD methodology is the alternating treatments design, which relative to the ABAB and multiple baseline designs potentially allows for more rapid comparison of two or more conditions (Barlow & Hayes, 1979; Hayes, Barlow, & Nelson-Gray, 1999). In the typical application of the design, two separate interventions are alternated following the baseline phase. The alternating feature of the design occurs when, subsequent to a baseline phase, the interventions are alternated in rapid succession for some specified number of sessions or trials. As an example, Intervention B could be implemented on one day and Intervention C on the next, with alternating interventions implemented over multiple days. In addition to a direct comparison of two interventions, the baseline (A) condition could be continued and compared with each intervention condition in the alternating phases. The order of this alternation of interventions across days may be based on either counterbalancing or a random schedule. Another variation, called the simultaneous treatment design (sometimes called the concurrent schedule design), involves exposing individual participants to the interventions simultaneously, with the participant's differential preference for the two interventions being the focus of the investigation. This latter design is used relatively infrequently in educational and psychological research, however.

The multiple baseline design involves an effect replication option across participants, settings, or behaviors. Multiple AB data series are compared and introduction of the intervention is staggered across time. In this design, more valid causal inferences are possible by staggering the intervention across one of the aforementioned units (i.e., sequential introduction of the intervention across time). The minimum number of phase repetitions needed to meet the standard advanced by Horner et al. (2005) is three, but four or more is recognized as more desirable (and statistically advantageous in cases in which, for example, the researcher is applying a randomization statistical test). Adding phase repetitions increases the power of the statistical test, similar to adding participants in a traditional group design (Kratochwill & Levin, in press). The number and timing of the repetitions can vary, depending on the outcomes of the intervention. For example, if change in the dependent variable is slow to occur, more time might be needed to demonstrate experimental control. Such a circumstance might also reduce the number of phase repetitions that can be scheduled due to cost and logistical factors. Among the characteristics of this design, effect replication across series is regarded as the characteristic with the greatest potential for enhancing internal and statistical-conclusion validity (see, for example, Levin, 1992).

Well-structured SCD research that embraces phase repetition and effect replication can rule out major threats to internal validity. The possible threats to internal validity in single-case research include the following (see also Shadish et al., 2002, p. 55):

1. ***Ambiguous Temporal Precedence:*** Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.

Embedded in the SCD *Standards* is a criterion that the independent variable is actively manipulated by the researcher, with measurement of the dependent variable occurring after that

manipulation. This sequencing ensures the presumed cause precedes the presumed effect. A SCD cannot meet *Standards* unless there is active manipulation of the independent variable.[21]

Replication of this manipulation-measurement sequence in the experiment further contributes to an argument of unidirectional causation (Shadish et al., 2002). Effect replication, as specified in the *Standards*, can occur either through within-case replication or multiple-case replication in a single experiment, or by conducting two or more experiments with the same or highly similar intervention conditions included. The *Standards* specify that the study must show a minimum of three demonstrations of the effect through the use of the same design and procedures. Overall, studies that can meet standards are designed to mitigate the threat of ambiguous temporal precedence.

2. ***Selection:*** Systematic differences between/among conditions in participant characteristics could cause the observed effect.

In most single-case research, selection is generally not a concern because one participant is exposed to both (or all) of the conditions of the experiment (i.e., each case serves as its own control, as noted in features for identifying a SCD in the *Standards*). However, there are some conditions under which selection might affect the design's internal validity. First, in SCDs that involve two or more between-case intervention conditions comprised of intact "units" (e.g., pairs, small groups, and classrooms), differential selection might occur. The problem is that the selected units might differ in various respects before the study begins. Because in most single-case research the units are not randomly assigned to the experiment's different intervention conditions, selection might then be a problem. This threat can further interact with other invalidating influences so as to confound variables (a methodological soundness problem) and compromise the results (an evidence credibility problem). Second, the composition of intact units (i.e., groups) can change (generally decrease in size, as a result of participant attrition) over time in a way that could compromise interpretations of a treatment effect. This is a particular concern when within-group individual participants drop out of a research study in a treatment-related (nonrandom) fashion (see also No. 6 below). The SCD *Standards* address traditional SCDs and do not address between-case group design features (for *Standards* for group designs, see the WWC Handbook). Third, in the multiple baseline design across cases, selection might be an issue when different cases sequentially begin the intervention based on "need" rather than on a randomly determined basis (e.g., a child with the most serious behavior problem among several candidate participants might be selected to receive the treatment first, thereby weakening the study's *external* validity).

3. ***History:*** Events occurring concurrently with the intervention could cause the observed effect.

[21] Manipulation of the independent variable is usually either described explicitly in the Method section of the text of the study or inferred from the discussion of the results. Reviewers will be trained to identify cases in which the independent variable is not actively manipulated and in that case, a study *Does Not Meet Standards*.

History is typically the most important threat to any time series, including SCDs. This is especially the case in *ex post facto* single-case research because the researcher has so little ability to investigate what other events might have occurred in the past and affected the outcome, and in simple (e.g., ABA) designs, because one need find only a single plausible alternative event about the same time as treatment. The most problematic studies, for example, typically involve examination of existing databases or archived measures in some system or institution (such as a school, prison, or hospital). Nevertheless, the study might not always be historically confounded in such circumstances; the researcher can investigate the conditions surrounding the treatment and build a case implicating the intervention as being more plausibly responsible for the observed outcomes relative to competing factors. Even in prospective studies, however, the researcher might not be the only person trying to improve the outcome. For instance, the patient might make other outcome-related changes in his or her own life, or a teacher or parent might make extra-treatment changes to improve the behavior of a child. SCD researchers should be diligent in exploring such possibilities. However, history threats are lessened in single-case research that involves one of the types of phase repetition necessary to meet standards (e.g., the ABAB design discussed above). Such designs reduce the plausibility that extraneous events account for changes in the dependent variable(s) because they require that the extraneous events occur at about the same time as the multiple introductions of the intervention over time, which is less likely to be true than is the case when only a single intervention is done.

4. ***Maturation:*** Naturally occurring changes over time could be confused with an intervention effect.

In single-case experiments, because data are gathered across time periods (for example, sessions, days, weeks, months, or years), participants in the experiment might change in some way due to the passage of time (e.g., participants get older, learn new skills). It is possible that the observed change in a dependent variable is due to these natural sources of maturation rather than to the independent variable. This threat to internal validity is accounted for in the *Standards* by requiring not only that the design document three replications/demonstrations of the effect, but that these effects must be demonstrated at a minimum of three different points in time. As required in the *Standards*, selection of an appropriate design with repeated assessment over time can reduce the probability that maturation is a confounding factor. In addition, adding a control series (i.e., an A phase or control unit such as a comparison group) to the experiment can help diagnose or reduce the plausibility of maturation and related threats (e.g., history, statistical regression). For example, see Shadish and Cook (2009).

5. ***Statistical Regression (Regression toward the Mean):*** When cases (e.g., single participants, classrooms, schools) are selected on the basis of their extreme scores, their scores on other measured variables (including re-measured initial variables) typically will be less extreme, a psychometric occurrence that can be confused with an intervention effect.

In single-case research, cases are often selected because their pre-experimental or baseline measures suggest high need or priority for intervention (e.g., immediate treatment for some

problem is necessary). If only pretest and posttest scores were used to evaluate outcomes, statistical regression would be a major concern. However, the repeated assessment identified as a distinguishing feature of SCDs in the *Standards* (wherein performance is monitored to evaluate level, trend, and variability, coupled with phase repetition in the design) makes regression easy to diagnose as an internal validity threat. As noted in the *Standards*, data are repeatedly collected during baseline and intervention phases and this repeated measurement enables the researcher to examine characteristics of the data for the possibility of regression effects under various conditions.

6. ***Attrition:*** Loss of respondents during a single-case time-series intervention study can produce artifactual effects if that loss is systematically related to the experimental conditions.

Attrition (participant dropout) can occur in single-case research and is especially a concern under at least three conditions. First, premature departure of participants from the experiment could render the data series too short to examine level, trend, variability, and related statistical properties of the data, which thereby may threaten data interpretation. Hence, the *Standards* require a minimum of five data points in a phase to meet evidence standards without reservations. Second, attrition of one or more participants at a critical time might compromise the study's internal validity and render any causal inferences invalid; hence, the *Standards* require a minimum of three phase repetitions to meet evidence standards. Third, in some single-case experiments, intact groups comprise the experimental units (e.g., group-focused treatments, teams of participants, and classrooms). In such cases, differential attrition of participants from one or more of these groups might influence the outcome of the experiment, especially when the unit composition change occurs at the point of introduction of the intervention. Although the *Standards* do not automatically exclude studies with attrition, reviewers are asked to attend to attrition when it is reported. Reviewers are encouraged to note that attrition can occur when (1) an individual fails to complete all required phases of a study, (2) the case is a group and individuals attrite from the group, or (3) the individual does not have adequate data points within a phase. Reviewers should also note when the researcher reports that cases were dropped and record the reason for that (for example, being dropped for nonresponsiveness to treatment). To monitor attrition through the various phases of single-case research, reviewers are asked to apply a template embedded in the coding guide similar to the flow diagram illustrated in the CONSORT Statement (Moher, Schulz, & Altman, 2001) and adopted by the American Psychological Association for randomized controlled trials research (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008). See Appendix F.1 for the WWC SCD attrition diagram. Attrition noted by reviewers should be brought to the attention of principal investigators (PIs) to assess whether the attrition may impact the integrity of the study design or evidence that is presented.

7. ***Testing:*** Exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confused with an intervention effect.

In SCDs, there are several different possibilities for testing effects—in particular, many measurements are likely to be "reactive" when administered repeatedly over time. For example, continuous exposure of participants to some curriculum measures might improve their performance over time. Sometimes the assessment process itself influences the outcomes of the study, such as when direct classroom observation causes change in student and teacher behaviors. Strategies to reduce or eliminate these influences have been proposed (Cone, 2001). In single-case research, the repeated assessment of the dependent variable(s) across phases of the design can help identify this potential threat. The effect replication standard can enable the researcher to reduce the plausibility of a claim that testing *per se* accounted for the intervention effect (see *Standards*).

8. ***Instrumentation:*** The conditions or nature of a measure might change over time in a way that could be confused with an intervention effect.

Confounding due to instrumentation can occur in single-case research when changes in a data series occur as a function of changes in the method of assessing the dependent variable over time. One of the most common examples occurs when data are collected by assessors who change their method of assessment over phases of the experiment. Such factors as reactivity, drift, bias, and complexity in recording might influence the data and implicate instrumentation as a potential confounding influence. Reactivity refers to the possibility that observational scores are higher as a result of the researcher monitoring the observers or observational process. Observer drift refers to the possibility that observers may change their observational definitions of the construct being measured over time, thereby not making scores comparable across phases of the experiment. Observational bias refers to the possibility that observers may be influenced by a variety of factors associated with expected or desired experimental outcomes, thereby changing the construct under assessment. Complexity may influence observational assessment in that more complex observational codes present more challenges than less complex codes with respect to obtaining acceptable levels of observer agreement. Numerous recommendations to control these factors have been advanced and can be taken into account (Hartmann, Barrios, & Wood, 2004; Kazdin, 1982).

9. ***Additive and Interactive Effects of Threats to Internal Validity:*** The impact of a threat can be added to that of another threat or may be moderated by levels of another threat.

In SCDs the aforementioned threats to validity may be additive or interactive. Nevertheless, the "Criteria for Designs that Meet Evidence Standards" and the "Criteria for Demonstrating Evidence of a Relation between an Independent and an Outcome Variable" have been crafted largely to address the internal validity threats noted above. Further, reviewers are encouraged to follow the approach taken with group designs, namely, to consider other confounding factors that might have a separate effect on the outcome variable (i.e., an effect that is not controlled for by the study design). Such confounding factors should be discussed with PIs to determine whether the study *Meets Standards*.

D. THE SINGLE-CASE DESIGN STANDARDS

The PI within each topic area will: (1) define the independent and outcome variables under investigation,[22] (2) establish parameters for considering fidelity of intervention implementation,[23] and (3) consider the reasonable application of the *Standards* to the topic area and specify any deviations from the *Standards* in that area protocol. For example, when measuring self-injurious behavior, a baseline phase of fewer than five data points may be appropriate. PIs might need to make decisions about whether the design is appropriate for evaluating an intervention. For example, an intervention associated with a permanent change in participant behavior should be evaluated with a multiple baseline design rather than an ABAB design. PIs will also consider the various threats to validity and how the researcher was able to address these concerns, especially in cases in which the *Standards* do not necessarily mitigate the validity threat in question (e.g., testing, instrumentation). Note that the SCD *Standards* apply to both observational measures and standard academic assessments. Similar to the approach with group designs, PIs are encouraged to define the parameters associated with "acceptable" assessments in their protocols. For example, repeated measures with alternate forms of an assessment may be acceptable and WWC psychometric criteria would apply. PIs might also need to make decisions about particular studies. Several questions will need to be considered, such as: (a) Will generalization variables be reported? (b) Will follow-up phases be assessed? (c) If more than one consecutive baseline phase is present, are these treated as one phase or two distinct phases? and (d) Are multiple treatments conceptually distinct or multiple components of the same intervention?

SINGLE-CASE DESIGN STANDARDS

These Standards are intended to guide WWC reviewers in identifying and evaluating SCDs. The first section of the *Standards* assists with identifying whether a study is a SCD. As depicted in Figure F1, a SCD should be reviewed using the 'Criteria for Designs that Meet Evidence Standards', to determine those that *Meet Evidence Standards,* those that *Meet Evidence Standards with Reservations,* and those that *Do Not Meet Evidence Standards.*

Studies that meet evidence standards (with or without reservations) should then be reviewed using the 'Criteria for Demonstrating Evidence of a Relation between an Independent Variable and a Dependent Variable' (see Figure F1).[24] This review will result in a sorting of SCD studies into three groups: those that have *Strong Evidence of a Causal Relation,* those that have *Moderate Evidence of a Causal Relation,* and those that have *No Evidence of a Causal Relation.*

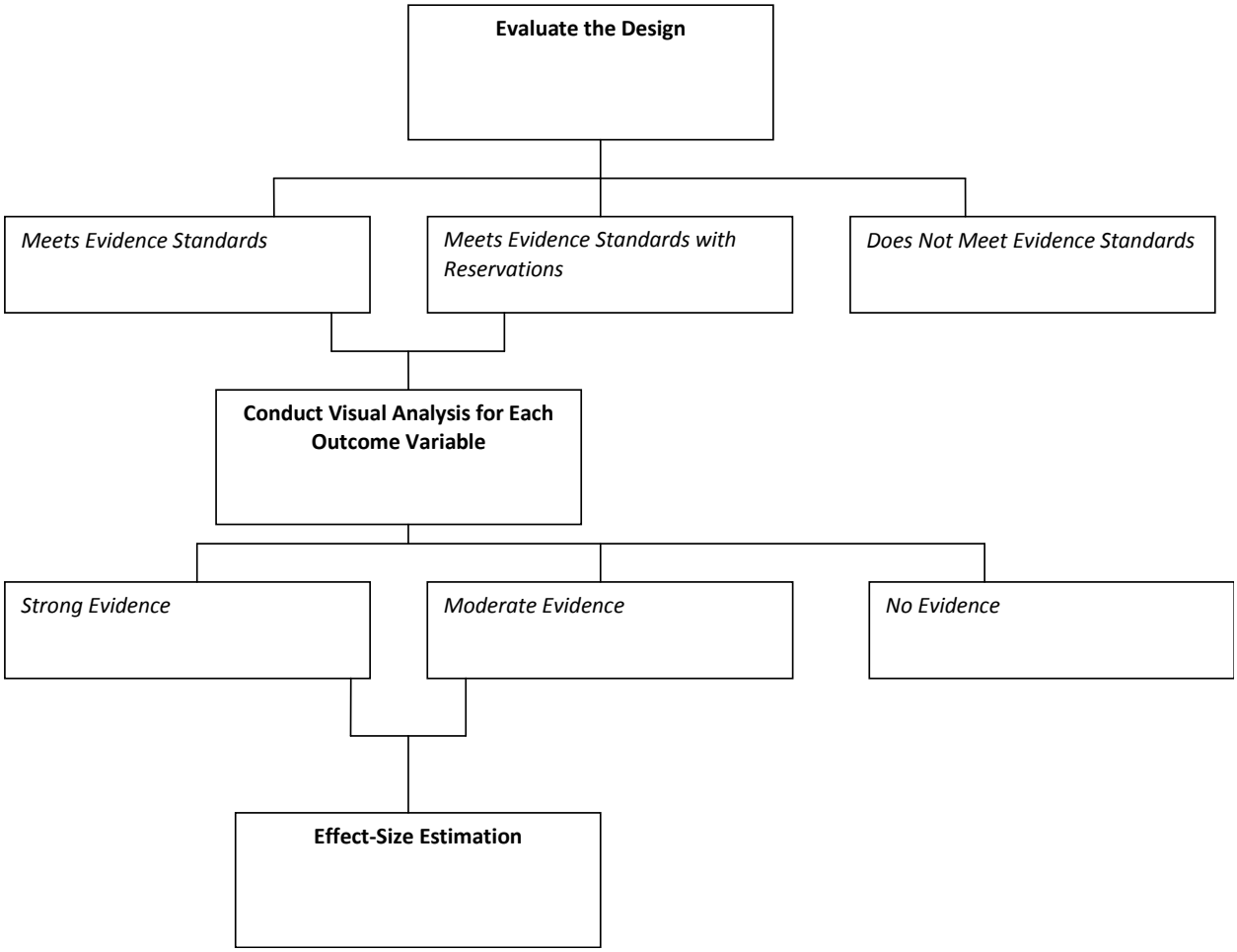
[22] Because SCDs are reliant on phase repetition and effect replication across participants, settings, and researchers to establish external validity, specification of the intervention materials, procedures, and context of the research is particularly important within these studies (Horner et al., 2005).

[23] Because interventions are applied over time, continuous measurement of implementation is a relevant consideration.

[24] This process results in a categorization scheme that is similar to that used for evaluating evidence credibility by inferential statistical techniques (hypothesis testing, effect-size estimation, and confidence-interval construction) in traditional group designs.

FIGURE F1

PROCEDURE FOR APPLYING SCD STANDARDS: FIRST EVALUATE DESIGN, THEN IF APPLICABLE, EVALUATE EVIDENCE



A. SINGLE-CASE DESIGN CHARACTERISTICS

SCDs are identified by the following features:

- An individual "case" is the unit of intervention and the unit of data analysis. A case may be a single participant or a cluster of participants (e.g., a classroom or community).
- Within the design, the case provides its own control for purposes of comparison. For example, the case's series of outcome variables prior to the intervention is compared with the series of outcome variables during (and after) the intervention.
- The outcome variable is measured *repeatedly* within and across *different* conditions or levels of the independent variable. These different conditions are referred to as "phases" (e.g., baseline phase, intervention phase).[25]

The *Standards* for SCDs apply to a wide range of designs, including ABAB designs, multiple baseline designs, alternating and simultaneous treatment designs, changing criterion designs, and variations of these core designs. Even though SCDs can be augmented by including one or more independent comparison cases (i.e., a comparison group), in this document the *Standards* address only the core SCDs and are not applicable to the augmented independent comparison SCDs.

B. CRITERIA FOR DESIGNS THAT MEET EVIDENCE STANDARDS

If the study appears to be a SCD, the following rules are used to determine whether the study's design *Meets Evidence Standards, Meets Evidence Standards with Reservations* or *Does Not Meet Evidence Standards.*

In order to *Meet Evidence Standards,* the following design criteria must be present:

- **The independent variable (i.e., the intervention) must be systematically manipulated, with the researcher determining when and how the independent variable conditions change.** If this standard is not met, the study *Does Not Meet Evidence Standards.*

[25] In SCDs, the ratio of data points (measures) to the number of cases usually is large so as to distinguish SCDs from other longitudinal designs (e.g., traditional pretest-posttest and general repeated-measures designs). Although specific prescriptive and proscriptive statements would be difficult to provide here, what can be stated is: (1) parametric univariate repeated-measures analysis cannot be performed when there is only one experimental case; (2) parametric multivariate repeated-measures analysis cannot be performed when the number of cases is less than or equal to the number of measures; and (3) for both parametric univariate and multivariate repeated-measures analysis, standard large-sample (represented here by large numbers of cases) statistical theory assumptions must be satisfied for the analyses to be credible (see also Kratochwill & Levin, in press, Footnote 1).

- **Each outcome variable must be measured systematically over time by more than one assessor, and the study needs to collect inter-assessor agreement in each phase and on at least twenty percent of the data points in each condition (e.g., baseline, intervention) and the inter-assessor agreement must meet minimal thresholds**. Inter-assessor agreement (commonly called interobserver agreement) must be documented on the basis of a statistical measure of assessor consistency. Although there are more than 20 statistical measures to represent inter-assessor agreement (see Berk, 1979; Suen & Ary, 1989), commonly used measures include percentage agreement (or proportional agreement) and Cohen's kappa coefficient (Hartmann, Barrios, & Wood, 2004). According to Hartmann et al. (2004), minimum acceptable values of inter-assessor agreement range from 0.80 to 0.90 (on average) if measured by percentage agreement and at least 0.60 if measured by Cohen's kappa. Regardless of the statistic, inter-assessor agreement must be assessed for each case on each outcome variable. A study needs to collect inter-assessor agreement in all phases. It must also collect inter-assessor agreement on at least 20% of all sessions (total across phases) for a condition (e.g., Baseline, Intervention.).[26] If this standard is not met, the study *Does Not Meet Evidence Standards.*

- **The study must include at least three attempts to demonstrate an intervention effect at three different points in time or with three different phase repetitions**. If this standard is not met, the study *Does Not Meet Evidence Standards*.[27] Examples of designs meeting this standard include ABAB designs, multiple baseline designs with at least three baseline conditions, alternating/simultaneous treatment designs with either at least three alternating treatments compared with a baseline condition or two alternating treatments compared with each other, changing criterion designs with at least three different criteria, and more complex variants of these designs. Examples of designs not meeting this standard include AB, ABA, and BAB designs.[28]

- **For a phase to qualify as an attempt to demonstrate an effect, the phase must have a minimum of three data points**.[29]

    - To *Meet Standards* a reversal/withdrawal (e.g., ABAB) design must have a minimum of four phases per case with at least 5 data points per phase. To *Meet Standards with Reservations* a reversal /withdrawal (e.g., ABAB) design must have a minimum of four phases per case with at least 3 data

[26] If the PI determines that there are exceptions to this *Standard*, they will be specified in the topic area or practice guide protocol. These determinations are based on the PIs content knowledge of the outcome variable.

[27] The three demonstrations criterion is based on professional convention (Horner, Swaminathan, Sugai, & Smolkowski, under review). More demonstrations further increase confidence in experimental control (Kratochwill & Levin, 2009).

[28] Although atypical, there might be circumstances in which designs without three replications meet the standards. A case must be made by the WWC PI researcher (based on content expertise) and at least two WWC reviewers must agree with this decision.

[29] If the PI determines that there are exceptions to this standard, these will be specified in the topic area or practice guide protocol. (For example, extreme self-injurious behavior might warrant a lower threshold of only one or two data points).

points per phase. Any phases based on fewer than three data *points cannot* b*e used to demonstrat*e existence or lack of an effect.

- To *Meet Standards* a multiple baseline design must have a minimum of six phases with at least 5 data points per phase. To *Meet Standards with Reservations* a multiple baseline design must have a minimum of six phases with at least 3 data points per phase. Any phases based on fewer than three data points *cannot* b*e used to demonstrat*e existence or lack of an effect.

- An alternating treatment design needs *five repetitions* of the alternating sequence to *Meet Standards.* Designs such as ABABBABAABBA, BCBCBCBCBC, and AABBAABBAABB would qualify, even though randomization or brief functional assessment may lead to one or two data points in a phase. A design with four repetitions would *Meet Standards with Reservations*, and a design with fewer than four repetitions *Does Not Meet Standards.*

C. CRITERIA FOR DEMONSTRATING EVIDENCE OF A RELATION BETWEEN AN INDEPENDENT VARIABLE AND AN OUTCOME VARIABLE

For studies that meet standards (with and without reservations), the following rules are used to determine whether the study provides *Strong Evidence*, *Moderate Evidence*, or *No Evidence* of a causal relation. In order to provide *Strong Evidence*, at least two WWC reviewers certified in visual (or graphical) analysis must verify that a causal relation was documented. Specifically this is operationalized as at least three demonstrations of the intervention effect along with no non-effects by[30]

- Documenting the consistency of level, trend, and variability within each phase
- Documenting the immediacy of the effect, the proportion of overlap, the consistency of the data across phases in order to demonstrate an intervention effect, and comparing the observed and projected patterns of the outcome variable
- Examining external factors and anomalies (e.g., a sudden change of level within a phase)

If a SCD does not provide three demonstrations of an effect, then the study is rated as *No Evidence.* If a study provides three demonstrations of an effect and also includes at least one demonstration of a non-effect, the study is rated as *Moderate Evidence.* The following characteristics must be considered when identifying a non-effect:

- Data within the baseline phase do not provide sufficient demonstration of a clearly defined pattern of responding that can be used to extrapolate the

[30] This section assumes that the demonstration of an effect will be established through "visual analysis," as described later. As the field reaches greater consensus about appropriate statistical analyses and quantitative effect-size measures, new standards for effect demonstration will need to be developed.

expected performance forward in time assuming no changes to the independent variable

- Failure to establish a consistent pattern within any phase (e.g., high variability within a phase)
- Either long latency between introduction of the independent variable and change in the outcome variable or overlap between observed and projected patterns of the outcome variable between baseline and intervention phases makes it difficult to determine whether the intervention is responsible for a claimed effect
- Inconsistent patterns across similar phases (e.g., an ABAB design in which the first time an intervention is introduced the outcome variable data points are high, the second time an intervention is introduced the outcome variable data points are low, and so on)
- Comparing the observed and projected patterns of the outcome variable between phases does not demonstrate evidence of a causal relation

When examining a multiple baseline design also consider the extent to which the time in which a basic effect is initially demonstrated with one series (e.g. first five days following introduction of the intervention for participant #1) is associated with change in the data pattern over the same time frame in the other series of the design (e.g. same five days for participants #2, #3, #4). If a basic effect is demonstrated within one series and there is a change in the data patterns in other series, the highest possible design rating is *Moderate Evidence.*

If a study has either *Strong Evidence or Moderate Evidence,* then effect-size estimation follows.

D. VISUAL ANALYSIS OF SINGLE-CASE RESEARCH RESULTS[31]

Single-case researchers traditionally have relied on visual analysis of the data to determine (a) whether evidence of a relation between an independent variable and an outcome variable exists; and (b) the strength or magnitude of that relation (Hersen & Barlow, 1976; Kazdin, 1982; Kennedy, 2005; Kratochwill, 1978; Kratochwill & Levin, 1992; McReynolds & Kearns, 1983; Richards, Taylor, Ramasamy, & Richards, 1999; Tawney & Gast, 1984; White & Haring, 1980). An inferred causal relation requires that changes in the outcome measure resulted from manipulation of the independent variable. A causal relation is demonstrated if the data across all phases of the study document at least three demonstrations of an effect at a minimum of three different points in time (as specified in the *Standards*). An effect is documented when the data pattern in one phase (e.g., an intervention phase) differs more than would be expected from the data pattern observed or extrapolated from the previous phase (e.g., a baseline phase) (Horner et al., 2005).

[31] Prepared by Robert Horner, Thomas Kratochwill, and Samuel Odom.

Our rules for conducting visual analysis involve four steps and six variables (Parsonson & Baer, 1978). The **first step** is documentation of a predictable baseline pattern of data (e.g., student is reading with many errors; student is engaging in high rates of screaming). If a convincing baseline pattern is documented, then the **second step** consists of examining the data within each phase of the study to assess the within-phase pattern(s). The key question is to assess whether there are sufficient data with sufficient consistency to demonstrate a predictable pattern of responding (see below). The **third step** in the visual analysis process is to compare the data from each phase with the data in the adjacent (or similar) phase to assess whether manipulation of the independent variable was associated with an "effect." An effect is demonstrated if manipulation of the independent variable is associated with predicted change in the pattern of the dependent variable. The **fourth step** in visual analysis is to integrate all the information from all phases of the study to determine whether there are at least three demonstrations of an effect at different points in time (i.e., documentation of a causal or functional relation) (Horner et al., in press).

To assess the effects within SCDs, six features are used to examine within- and between-phase data patterns: **(1) level, (2) trend, (3) variability, (4) immediacy of the effect, (5) overlap, and (6) consistency of data patterns across similar phases** (Fisher, Kelley, & Lomas, 2003; Hersen & Barlow, 1976; Kazdin, 1982; Kennedy, 2005; Morgan & Morgan, 2009; Parsonson & Baer, 1978). These six features are assessed individually and collectively to determine whether the results from a single-case study demonstrate a causal relation and are represented in the "Criteria for Demonstrating Evidence of a Relation between an Independent Variable and Outcome Variable" in the *Standards*. "Level" refers to the mean score for the data within a phase. "Trend" refers to the slope of the best-fitting straight line for the data within a phase and "variability" refers to the range or standard deviation of data about the best-fitting straight line. Examination of the data within a phase is used (a) to describe both the observed pattern of a unit's performance, and (b) to extrapolate the expected performance forward in time assuming no changes in the independent variable were to occur (Furlong & Wampold, 1981). The six visual analysis features are used collectively to compare the observed and projected patterns for each phase with the actual pattern observed after manipulation of the independent variable. This comparison of observed and projected patterns is conducted across all phases of the design (e.g., baseline to treatment, treatment to baseline, treatment to treatment, etc.).

In addition to comparing the level, trend, and variability of data within each phase, the researcher also examines data patterns across phases by considering the immediacy of the effect, overlap, and consistency of data in similar phases. "Immediacy of the effect" refers to the change in level between the last three data points in one phase and the first three data points of the next. The more rapid (or immediate) the effect, the more convincing the inference that change in the outcome measure was due to manipulation of the independent variable. Delayed effects might actually compromise the internal validity of the design. However, predicted delayed effects or gradual effects of the intervention may be built into the design of the experiment that would then influence decisions about phase length in a particular study. "Overlap" refers to the proportion of data from one phase that overlaps with data from the previous phase. The smaller the proportion of overlapping data points (or conversely, the larger the separation), the more compelling the demonstration of an effect. "Consistency of data in similar phases" involves looking at data from all phases within the same condition (e.g., all "baseline" phases; all "peer-tutoring" phases) and examining the extent to which there is consistency in the data patterns from phases with the same conditions. The greater the consistency, the more likely the data represent a causal relation.

These six features are assessed both individually and collectively to determine whether the results from a single-case study demonstrate a causal relation.

Regardless of the type of SCD used in a study, visual analysis of: (1) level, (2) trend, (3) variability, (4) overlap, (5) immediacy of the effect, and (6) consistency of data patterns across similar phases are used to assess whether the data demonstrate at least three indications of an effect at different points in time. If this criterion is met, the data are deemed to document a causal relation, and an inference may be made that change in the outcome variable is causally related to manipulation of the independent variable (see *Standards*).

Figures 1–8 provide examples of the visual analysis process for one common SCD, the ABAB design, using proportion of 10-second observation intervals with child tantrums as the dependent variable and a tantrum intervention as the independent variable. The design is appropriate for interpretation because the ABAB design format allows the opportunity to assess a causal relation (e.g., to assess if there are three demonstrations of an effect at three different points in time, namely the B, A, and B phases following the initial A phase).

Step 1: The first step in the analysis is to determine whether the data in the Baseline 1 (first A) phase document that: (a) the proposed concern/problem is demonstrated (tantrums occur too frequently) and (b) the data provide sufficient demonstration of a clearly defined (e.g., predictable) baseline pattern of responding that can be used to assess the effects of an intervention. This step is represented in the Evidence Standards because if a proposed concern is not demonstrated or a predictable pattern of the concern is not documented, the effect of the independent variable cannot be assessed. The data in Figure 1 in Appendix F.2 demonstrate a Baseline 1 phase with 11 sessions, with an average of 66 percent throwing tantrums across these 11 sessions. The range of tantrums per session is from 50 percent to 75 percent with an increasing trend across the phase and the last three data points averaging 70 percent. These data provide a clear pattern of responding that would be outside socially acceptable levels, and if left unaddressed would be expected to continue in the 50 percent to 80 percent range.

The two purposes of a baseline are to (a) document a pattern of behavior in need of change, and (b) document a pattern that has sufficiently consistent level and variability, with little or no trend, to allow comparison with a new pattern following intervention. Generally, stability of a baseline depends on a number of factors and the options the researcher has selected to deal with instability in the baseline (Hayes et al., 1999). One question that often arises in single-case design research is how many data points are needed to establish baseline stability. First, the amount of variability in the data series must be considered. Highly variable data may require a longer phase to establish stability. Second, if the effect of the intervention is expected to be large and demonstrates a data pattern that far exceeds the baseline variance, a shorter baseline with some instability may be sufficient to move forward with intervention implementation. Third, the quality of measures selected for the study may impact how willing the researcher/reviewer is to accept the length of the baseline. In terms of addressing an unstable baseline series, the researcher has the options of: (a) analyzing and reporting the source of variability; (b) waiting to see whether the series stabilizes as more data are gathered; (c) considering whether the correct unit of analysis has been selected for measurement and if it represents the reason for instability in the data; and (d) moving forward with the intervention despite the presence of baseline instability. Professional standards for acceptable baselines are emerging, but the decision to end any baseline with fewer than five data points or to end a baseline with an outlying data point

should be defended. In each case it would be helpful for reviewers to have this information and/or contact the researcher to determine how baseline instability was addressed, along with a rationale.

Step 2: The second step in the visual analysis process is to assess the level, trend, and variability of the data within each phase and to compare the observed pattern of data in each phase with the pattern of data in adjacent phases. The horizontal lines in Figure 2 illustrate the comparison of phase levels and the lines in Figure 3 illustrate the comparison of phase trends. The upper and lower defining range lines in Figure 4 illustrate the phase comparison for phase variability. In Figures 2–4, the level and trend of the data differ dramatically from phase to phase; however, changes in variability appear to be less dramatic.

Step 3: The information gleaned through examination of level, trend, and variability is supplemented by comparing the overlap, immediacy of the effect, and consistency of patterns in similar phases. Figure 5 illustrates the concept of overlap. There is no overlap between the data in Baseline 1 (A1) and the data in Intervention 1 (B1). There is one overlapping data point (10 percent; session 28) between Intervention 1 (B1) Baseline 2 (A2), and there is no overlap between Baseline 2 (A2) and Intervention 2 (B2).

Immediacy of the effect compares the extent to which the level, trend, and variability of the last three data points in one phase are discriminably different from the first three data points in the next. The data in the ovals, squares, and triangles of Figure 6 illustrate the use of immediacy of the effect in visual analysis. The observed effects are immediate in each of the three comparisons (Baseline 1 and Intervention 1, Intervention 1 and Baseline 2, Baseline 2 and Intervention 2).

Consistency of similar phases examines the extent to which the data patterns in phases with the same (or similar) procedures are similar. The linked ovals in Figure 7 illustrate the application of this visual analysis feature. Phases with similar procedures (Baseline 1 and Baseline 2, Intervention 1 and Intervention 2) are associated with consistent patterns of responding.

Step 4: The final step of the visual analysis process involves combining the information from each of the phase comparisons to determine whether all the data in the design (data across all phases) meet the standard for documenting three demonstrations of an effect at different points in time. The bracketed segments in Figure 8 (A, B, C) indicate the observed and projected patterns of responding that would be compared with actual performance. Because the observed data in the Intervention 1 phase are outside the observed and projected data pattern of Baseline 1, the Baseline 1 and Intervention 1 comparison demonstrates an effect (Figure 8A). Similarly, because the data in Baseline 2 are outside of the observed and projected patterns of responding in Intervention 1, the Intervention 1 and Baseline 2 comparison demonstrates an effect (Figure 8B). The same logic allows for identification of an effect in the Baseline 2 and Intervention 2 comparison. Because the three demonstrations of an effect occur at different points in time, the full set of data in this study are considered to document a causal relation as specified in the *Standards.*

The rationale underlying visual analysis in SCDs is that predicted and replicated changes in a dependent variable are associated with active manipulation of an independent variable. The

process of visual analysis is analogous to the efforts in group-design research to document changes that are causally related to introduction of the independent variable. In group-design inferential statistical analysis, a statistically significant effect is claimed when the observed outcomes are sufficiently different from the expected outcomes that they are deemed unlikely to have occurred by chance. In single-case research, a claimed effect is made when three demonstrations of an effect are documented at different points in time. The process of making this determination, however, requires that the reader is presented with the individual unit's raw data (typically in graphical format) and actively participates in the interpretation process.

There will be studies in which some participants demonstrate an intervention effect and others do not. The evidence rating (*Strong Evidence, Moderate Evidence,* or *No Evidence*) accounts for mixed effects.

E. RECOMMENDATIONS FOR COMBINING STUDIES

When implemented with multiple design features (e.g., within- and between-case comparisons), SCDs can provide a strong basis for causal inference (Horner et al., 2005). Confidence in the validity of intervention effects demonstrated within cases is enhanced by replication of effects across different cases, studies, and research groups (Horner & Spaulding, in press). The results from single-case design studies will not be combined into a single summary rating unless they meet the following threshold:[32]

1. A minimum of five SCD research papers examining the intervention that *Meet Evidence Standards or Meet Evidence Standards with Reservations*
2. The SCD studies must be conducted by at least three different research teams at three different geographical locations
3. The combined number of experiments (i.e., single-case design examples) across the papers totals at least 20

F. EFFECT-SIZE ESTIMATES FOR SINGLE-CASE DESIGNS[33]

Effect-size estimates are available for most designs involving group comparisons, and in meta-analyses there is widespread agreement about how these effect sizes (ES) should be expressed, what the statistical properties of the estimators are (e.g., distribution theory, conditional variance), and how to translate from one measure (e.g., a correlation) to another (e.g., Hedges' g). This is not true for SCDs; the field is much less well-developed, and there are no agreed-upon methods or standards for effect size estimation. What follows is a brief summary of the main issues, with a more extensive discussion in an article by Shadish, Rindskopf, and Hedges (2008).

[32] These are based on professional conventions. Future work with SCD meta-analysis can offer an empirical basis for determining appropriate criteria and these recommendations might be revised.

[33] Prepared by David Rindskopf and William Shadish.

Several issues are involved in creating effect size estimates. First is the general issue of how to quantify the size of an effect. One can quantify the effect for a single case, or for a group of cases within one study, or across several SCD studies. Along with a quantitative ES estimate, one must also consider the accuracy of the estimate; generally the issues here are estimating a standard error, constructing confidence intervals, and testing hypotheses about effect sizes. Next is the issue of comparability of different effect sizes for SCDs. Finally the panel considers comparability of ES estimates for SCDs and for group-based designs.

Most researchers using SCDs still base their inferences on visual analysis, but several quantitative methods have been proposed. Each has flaws, but some methods are likely to be more useful than others; the panel recommends using some of these until better methods are developed.

A number of nonparametric methods have been used to analyze SCDs (e.g., Percentage of Nonoverlapping Data [PND], Percentage of All Nonoverlapping Data [PAND], or Percent Exceeding the Median [PEM]). Some of these have been accompanied by efforts to convert them to parametric estimators such as the phi coefficient, which might in turn be comparable to typical between-groups measures. If that could be done validly, then one could use distribution theory from standard estimators to create standard errors and significance tests. However, most such efforts make the erroneous assumption that nonparametric methods do not need to be concerned with the assumption of independence of errors, and so the conversions might not be valid. In such cases, the distributional properties of these measures are unknown, and so standard errors and statistical tests are not formally justified. Nonetheless, if all one wanted was a rough measure of the approximate size of the effect without formal statistical justification or distribution theory, selecting one of these methods would make sense. However, none of these indices deal with trend, so the data would need to be detrended[34] with, say, first-order differencing before computing the index. One could combine the results with ordinary unweighted averages, or one could weight by the number of cases in a study.

Various parametric methods have been proposed, including regression estimates and multilevel models. Regression estimates have three advantages. First, many primary researchers are familiar with regression so both the analyses and the results are likely to be easily understood. Second, these methods can model trends in the data, and so do not require prior detrending of the data. Third, regression can be applied to obtain an effect size from a single case, whereas multilevel models require several cases within a study. But they also come with disadvantages. Although regression models do permit some basic modeling of error structures, they are less flexible than multilevel models in dealing with complex error structures that are likely to be present in SCD data. For multi-level models, many researchers are less familiar with both the analytic methods and the interpretation of results, so that their widespread use is

[34] When a trend is a steady increase or decrease in the dependent variable over time (within a phase), such a trend would produce a bias in many methods of analysis of SCD data. For example, if with no treatment, the number of times a student is out of her seat each day for 10 days is 20, 19, 18, 17, 16, 15, 14, 13, 12, 11, this is a decreasing trend. If a "treatment" is introduced after the fifth day, so that the last 5 days' data are during a treatment phase, some methods would find the treatment very effective. For example, all of the measurements after the treatment are lower than any of the measurements before the treatment, apparently showing a strong effect. To correct for the effect of trend (i.e., to "detrend" the data), one can either subtract successive observations (e.g., 19-20, 18-19, etc.) and compile these in a vector within a phase (one cannot subtract from the final observation and so it is excluded) which is called differencing, or use statistical methods that adjust for this trend.

probably less likely than with regression. Also, practical implementation of multilevel models for SCDs is technically challenging, probably requiring the most intense supervision and problem-solving of any method. Even if these technical developments were to be solved, the resulting estimates would still be in a different metric than effect-size estimates based on between-group studies, so one could not compare effect sizes from SCDs to those from group studies.

A somewhat more optimistic scenario is that methods based on multilevel models can be used when data from several cases are available and the same outcome measure is used in all cases. Such instances do not require a standardized effect-size estimator because the data are already in the same metric. However, other technical problems remain, estimators are still not comparable with those from between-groups studies (see further discussion below), and such instances tend to be rare across studies.

The quantitative methods that have been proposed are not comparable with those used in group-comparison studies. In group studies, the simplest case would involve the comparison of two groups, and the mean difference would typically be standardized by dividing by the control group variance or a pooled within-group variance. These variances reflect variation across people. In contrast, single-case designs, by definition, involve comparison of behavior within an individual (or other unit), across different conditions. Attempts to standardize these effects have usually involved dividing by some version of a within-phase variance, which measures variation of one person's behavior at different times (instead of variation across different people). Although there is nothing wrong statistically with doing this, it is not comparable with the usual between-groups standardized mean difference statistic. Comparability is crucial if one wishes to compare results from group designs with SCDs.

That being said, some researchers would argue that there is still merit in computing some effect size index like those above. One reason is to encourage the inclusion of SCD data in recommendations about effective interventions. Another reason is that it seems likely that the rank ordering of most to least effective treatments would be highly similar no matter what effect size metric is used. This latter hypothesis could be partially tested by computing more than one of these indices and comparing their rank ordering.

An effect-size estimator for SCDs that is comparable to those used in between-groups studies is badly needed. Shadish et al. (2008) have developed an estimator for continuous outcomes that is promising in this regard, though the distribution theory is still being derived and tested. However, the small number of cases in most studies would make such an estimate imprecise (that is, it would have a large standard error and an associated wide confidence interval). Further, major problems remain to be solved involving accurate estimation of error structures for noncontinuous data—for example, different distributional assumptions that might be present in SCDs (e.g., count data should be treated as Poisson distributed). Because many outcomes in SCDs are likely to be counts or rates, this is a nontrivial limitation to using the Shadish et al. (2008) procedure. Finally, this method does not deal adequately with trend as currently developed, although standard methods for detrending the data might be reasonable to use. Hence, it might be premature to advise the use of these methods except to investigate further their statistical properties.

Until multilevel methods receive more thorough investigation, the panel suggests the following guidelines for estimating effect sizes in SCDs. First, in those rare cases in which the dependent variable is already in a common metric, such as proportions or rates, then these are preferred to standardized scales. Second, if only one standardized effect-size estimate is to be chosen, the regression-based estimators are probably best justified from both technical and practical points of view in that SCD researchers are familiar with regression. Third, the panel strongly recommends doing sensitivity analyses. For example, one could report one or more nonparametric estimates (but not the PND estimator, because it has undesirable statistical properties) in addition to the regression estimator. Results can then be compared over estimators to see if they yield consistent results about which interventions are more or less effective. Fourth, summaries across cases within studies and across studies (e.g., mean and standard deviation of effect sizes) can be computed when the estimators are in a common metric, either by nature (e.g., proportions) or through standardization. Lacking appropriate standard errors to use with the usual inverse-variance weighting, one might report either unweighted estimators or estimators weighted by a function of either the number of cases within studies or the number of time points within cases, although neither of these weights has any strong statistical justification in the SCD context.

REFERENCES

APA Publications and Communications Board Working Group on Journal Article Reporting Standards, (2008). Reporting standards for research in Psychology: Why do we need them? What might they be? *American Psychologist*, 63, 839-851.

Barlow, D. H., & Hayes, S. C. (1979). Alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. *Journal of Applied Behavior Analysis*, 12, 199–210.

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency*, 83, 460–472.

Cone, J. D. (2001). *Evaluating outcomes: Empirical tools for effective practice*. Washington, DC: American Psychological Association.

Fisher, W., Kelley, M., & Lomas, J. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis*, 36, 387–406.

Furlong, M., & Wampold, B. (1981). Visual analysis of single-subject studies by school psychologists. *Psychology in the Schools*, 18, 80–86.

Hartmann, D. P., Barrios, B. A., & Wood, D. D. (2004). Principles of behavioral observation. In S. N. Haynes and E. M. Hieby (Eds.), *Comprehensive handbook of psychological assessment (Vol. 3, Behavioral assessment*) (pp. 108-127). New York: John Wiley & Sons.

Hayes, S. C. (1981). Single-case experimental designs and empirical clinical practice. *Journal of Consulting and Clinical Psychology*, 49, 193–211.

Hayes, S. C., Barlow, D. H., Nelson-Gray, R. O. (1999). *The scientist practitioner: Research and accountability in the age of managed care* (2nd ed.). Needham Heights, MA: Allyn & Bacon.

Hersen, M., & Barlow, D. H. (1976). *Single-case experimental designs: Strategies for studying behavior change*. New York: Pergamon.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., Wolery, M. (2005). The use of single subject research to identify evidence-based practice in special education. *Exceptional Children* 71(2), 165–179.

Horner, R., & Spaulding, S. (in press). Single-Case Research Designs. Encyclopedia. Springer.

Horner, R., Swaminathan, H., Sugai, G., & Smolkowski, K. (in press). Expanding analysis of single case research. Washington, DC: Institute of Education Science, U.S. Department of Education.

Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York: Oxford University Press.

Kazdin, A. E. (in press). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York: Oxford University Press.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston: Allyn and Bacon.

Kratochwill, T. R. (Ed.). (1978). *Single subject research: Strategies for evaluating change*. New York: Academic Press.

Kratochwill, T. R. (1992). Single-case research design and analysis: An overview In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 1–14). Hillsdale, NJ: Erlbaum.

Kratochwill, T. R., & Levin, J. R. (Eds.). (1992). *Single-case research design and analysis: New directions for psychology and education*. Hillsdale, NJ: Erlbaum.

Kratochwill, T. R., & Levin, J. R. (In press). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*.

Levin, J. R. (1994). Crafting educational intervention research that's both credible and creditable. *Educational Psychology Review*, 6, 231–243.

Levin, J. R., O'Donnell, A. M., & Kratochwill, T. R. (2003). Educational/psychological intervention research. In I. B. Weiner (Series Ed.) and W. M. Reynolds & G. E. Miller (Vol. Eds.). *Handbook of psychology: Vol. 7. Educational psychology* (pp. 557–581). New York: Wiley.

Moher, D., Schulz, K. F., & Altman, D. G. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel-group randomized trials. *Annals of Internal Medicine*, 134, 657–662.

Morgan, D., & Morgan R., (2009). *Single-case research methods for the behavioral and health sciences*. Los Angles, Sage Publications Inc.

McReynolds, L. & Kearns, K. (1983). *Single-subject experimental designs in communicative disorders*. Baltimore: University Park Press.

Odom, S.L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children* 71(2), 137–148.

Parsonson, B., & Baer, D. (1978). The analysis and presentation of graphic data. In T. Kratchowill (Ed.) *Single Subject Research* (pp. 101–166). New York: Academic Press.

Richards, S. B., Taylor, R., Ramasamy, R., & Richards, R. Y. (1999). *Single subject research: Applications in educational and clinical settings*. Belmont, CA: Wadsworth.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin.

Shadish, W. R., & Cook, T. D. (2009). The renaissance of field experimentation in evaluating interventions. *Annual Review of Psychology*, 60, 607–629.

Shadish, W. R., Rindskopf, D. M. & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, 3, 188–196.

Suen, H. K., & Ary, D. (1989). *Analyzing quantitative behavioral observation data*. Hillsdale, NJ: Erlbaum.

Tawney, J. W., & Gast, D. L. (1984*). Single subject research in special education*. Columbus, OH: Merrill.

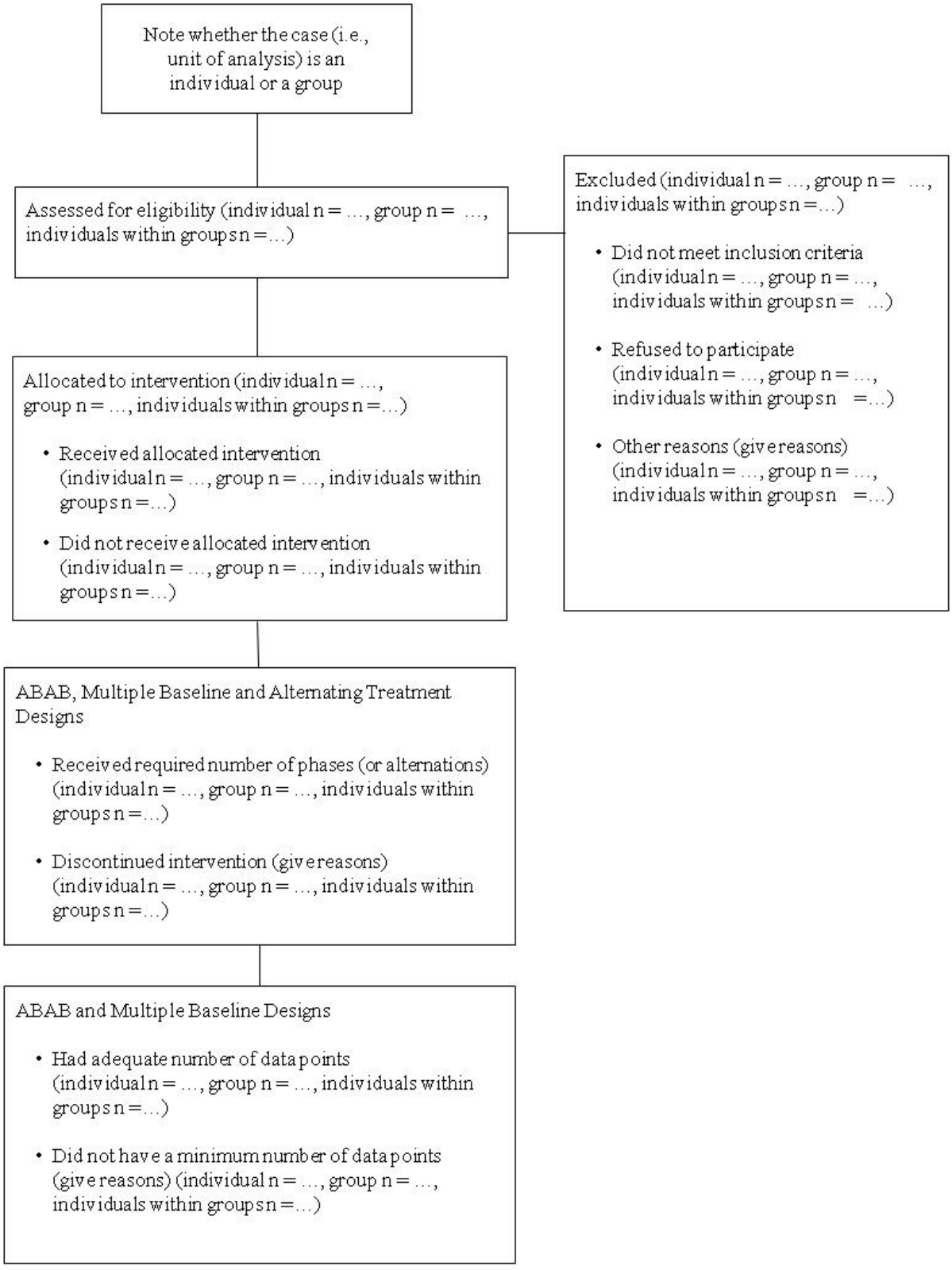
What Works Clearinghouse. (2008). *Procedures and standards handbook* (version 2.0). Retrieved July 10, 2009, from http://ies.ed.gov/ncee/wwc/references/idocviewer/doc.aspx?docid=19&tocid=1.

White, O. R., & Haring, N. G. (1980). *Exceptional teaching* (2nd ed.). Columbus, Ohio: Charles E. Merrill.

APPENDIX F.1

ATTRITION DIAGRAM

ATTRITION DIAGRAM



APPENDIX F.2

VISUAL ANALYSIS

Figure 1. Depiction of an ABAB Design

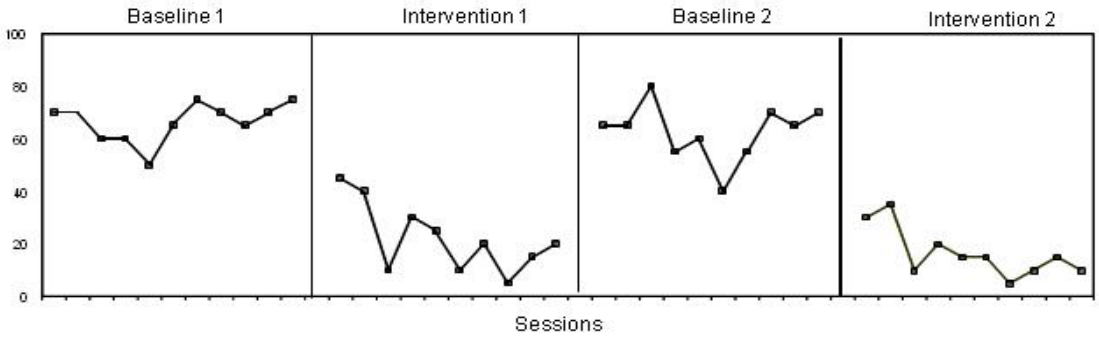


Figure 2. An Example of Assessing Level with Four Phases of an ABAB Design

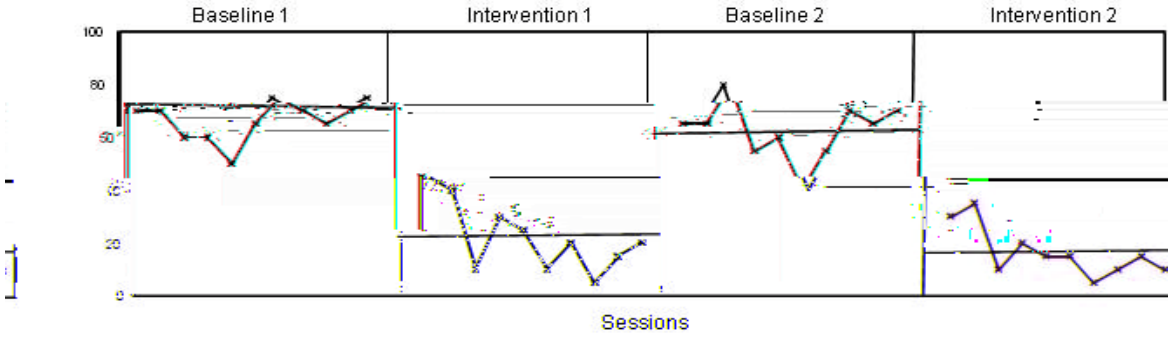


Figure 3. An Example of Assessing in Each Phase of an ABAB Design

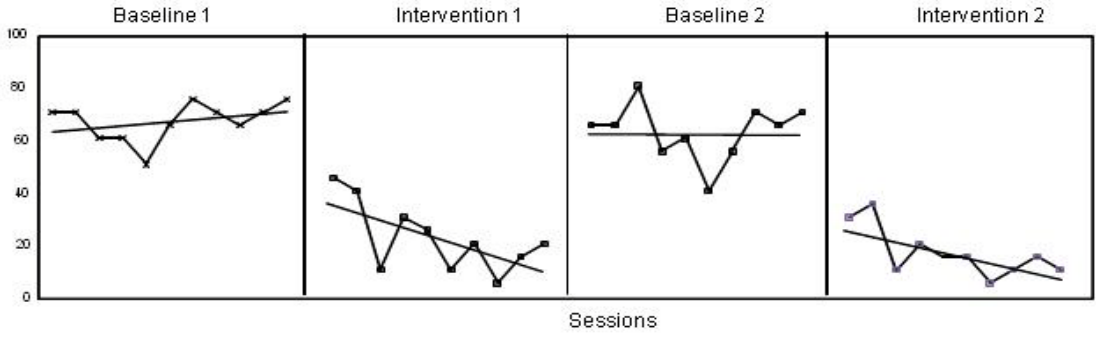


Figure 4. Assess Variability Within Each Phase

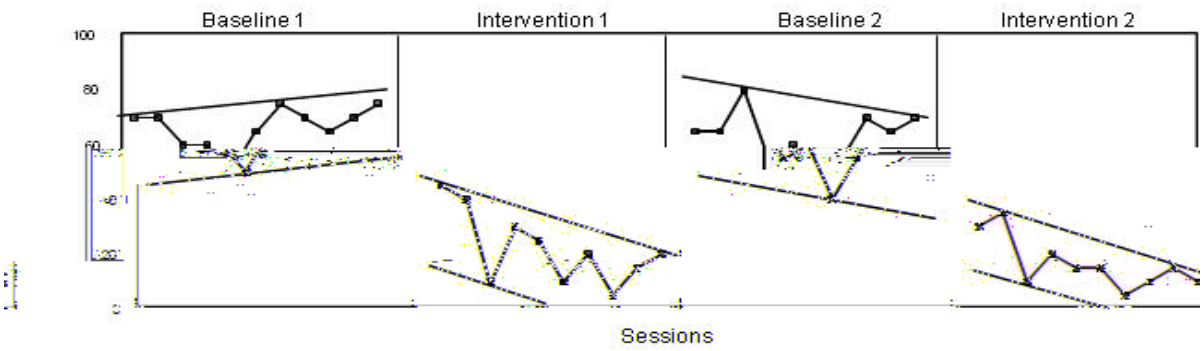


Figure 5. Consider Overlap Between Phases

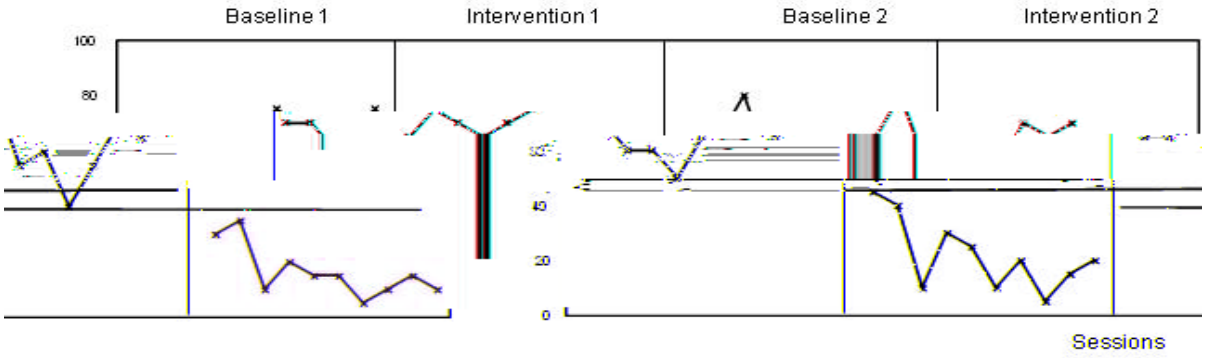


Figure 6. Examine the Immediacy of Effect with Each Phase Transition

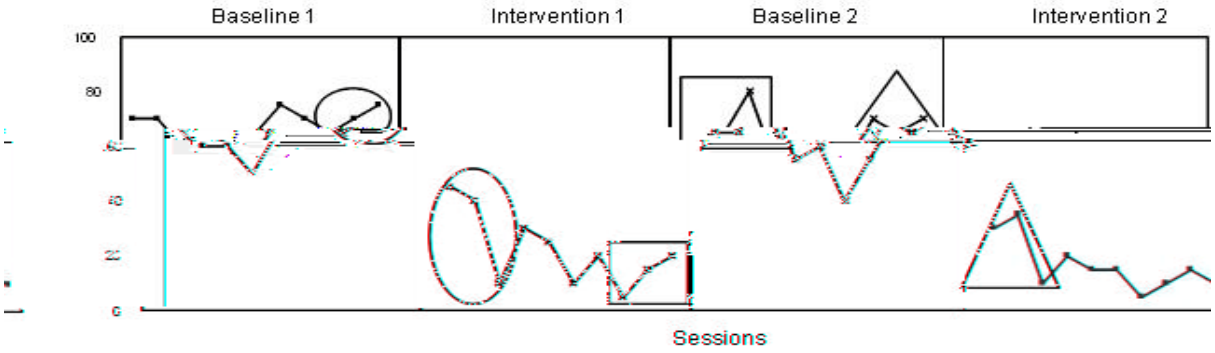


Figure 7. Examine Consistency Across Similar Phases

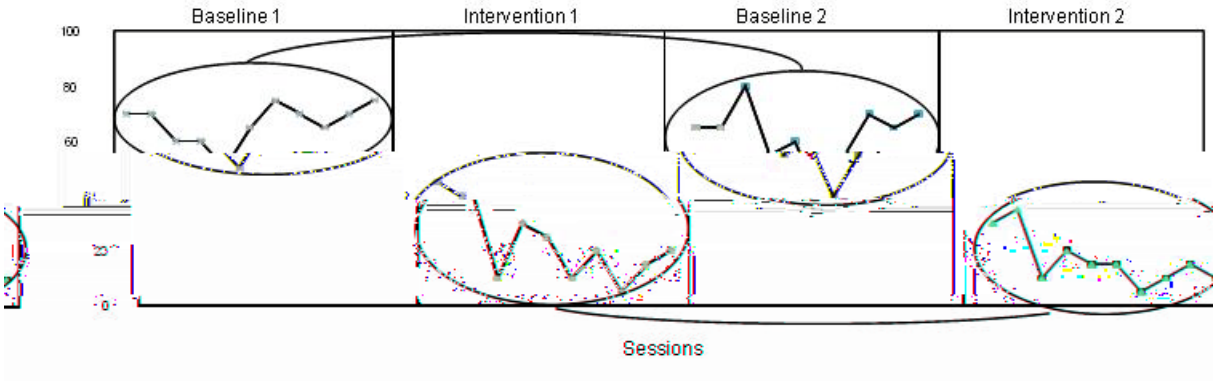


Figure 8A. Examine Observed and Projected Comparison Baseline 1 to Intervention 1

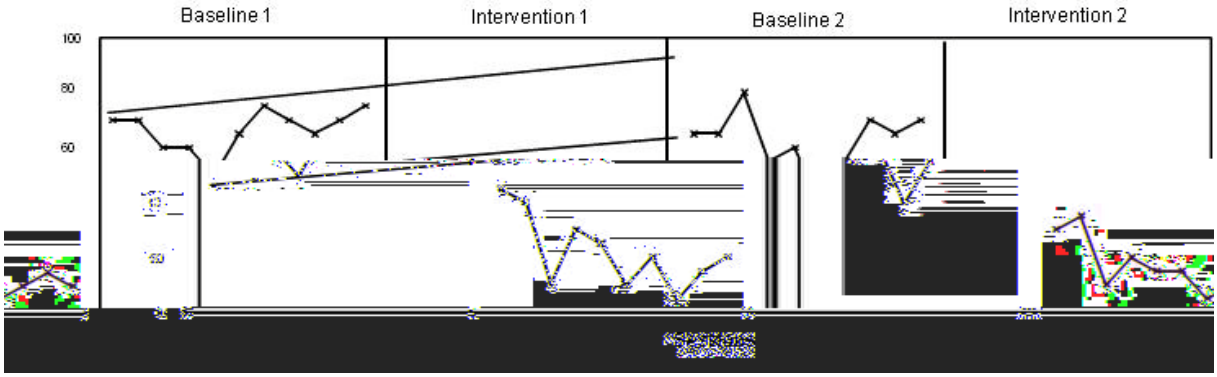


Figure 8B. Examine Observed and Projected Comparison Intervention to Baseline 2

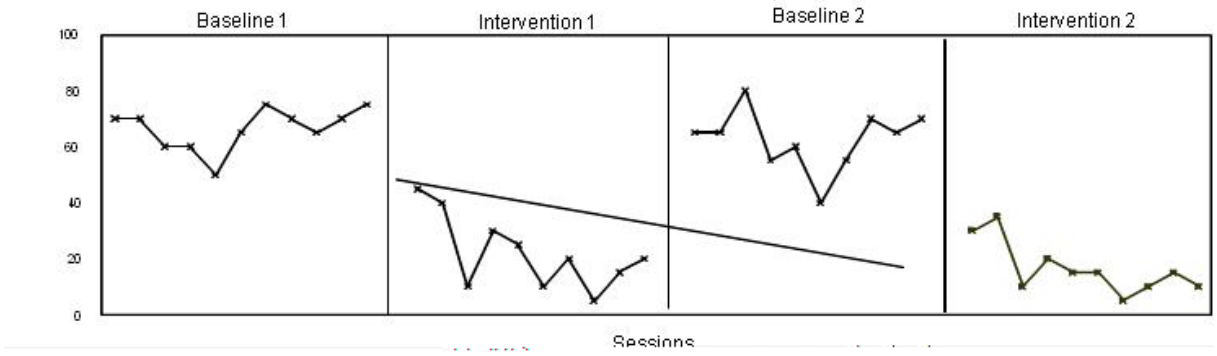
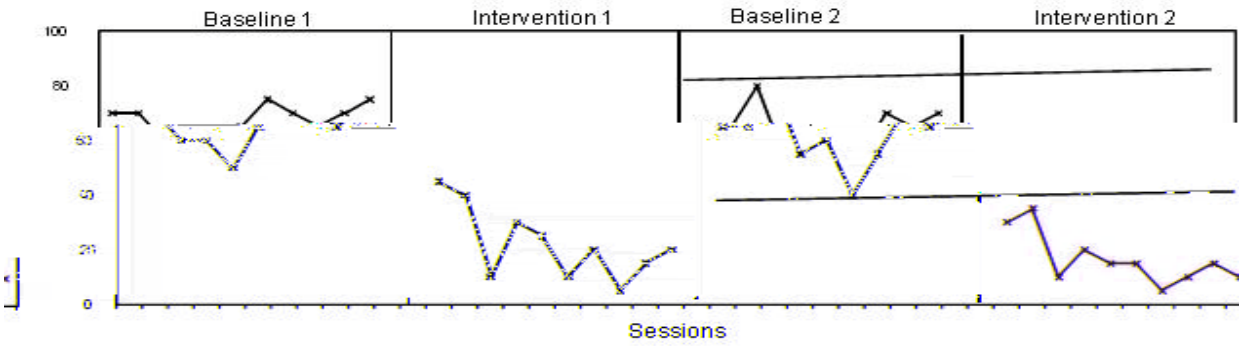


Figure 8C. Examine Observed and Projected Comparison Baseline 2 to Intervention 2



APPENDIX G. INTERVENTION RATING SCHEME

The following heuristics are applied to the outcome variable(s) identified by the principal investigator (PI) as relevant to the review. The PI may choose to ignore some variables if they are judged sufficiently peripheral or nonrepresentative and to consider only the remaining ones. Similarly, if the PI judges that there is one core variable with all the others secondary or subsidiary, only that one may be considered.

A. DEFINITIONS AND DEFAULTS

- *Strong and weak designs.* A strong design is one that Meets Evidence Standards, whereas a weak design is one that Meets Evidence Standards with Reservations.

- *Effect size*. A single effect size or, in the case of multiple measures of the specified outcome, either (1) the mean effect size or (2) the effect size for each individual measure within the domain.

- *Substantively important.* The smallest positive value at or above which the effect is deemed substantively important with relatively high confidence for the outcome domain at issue. Effect sizes at least this large will be taken as a qualified positive effect even though they may not reach statistical significance in a given study. The suggested default value is a student-level effect size greater than or equal to 0.25.[35] The PI may set a different default if explicitly justified in terms of the nature of the intervention or the outcome domain.

- *Statistical significance.* A finding of statistical significance using a two-tailed *t*-test with $\alpha = .05$ for a single measure or mean effect within each domain.

- *Accounting for clustering.* A *t*-test applied to the effect size (or mean effect size in cases of multiple measures of the outcome) that incorporates an adjustment for clustering. This procedure allows the reviewer to test the effect size directly when a misaligned analysis is reported (see Appendix C). The suggested default intra-class correlation (ICC) value is .20 for achievement outcomes and .10 for behavioral and attitudinal outcomes. The PI may set different defaults if explicitly justified in terms of the nature of the research circumstances or the outcome domain.

- *Accounting for multiple comparisons.* When multiple hypothesis tests are performed within a domain, the Benjamini-Hochberg procedure may be used to correct for multiple comparisons and identify statistically significant effects for individual measures (see Appendix D).

[35] Note that this criterion is entirely based on student-level effect sizes. Cluster-level effect sizes are ignored for the purpose of the rating scheme because they are based on a different effect size metric than the student-level effect sizes and, therefore, are not comparable to student-level effect sizes. Moreover, cluster-level effect sizes are relatively rare, and there is not enough knowledge in the field yet to set a defensible minimum effect size for cluster-level effect sizes.

B. CHARACTERIZING STUDY EFFECTS

Statistically significant positive effect if any of the following is true:

If the analysis as reported by the study author is properly aligned:

For a single outcome measure:

- The effect reported is positive and statistically significant.

For multiple outcome measures:

- Univariate statistical tests are reported for each outcome measure and at least half of the effects are positive and statistically significant and no effects are negative and statistically significant.
- Univariate statistical tests are reported for each outcome measure and the effect for at least one measure within the domain is positive and statistically significant and no effects are negative and statistically significant, accounting for multiple comparisons.
- The mean effect for the multiple measures of the outcome is positive and statistically significant.
- The omnibus effect for all the outcome measures together is reported as positive and statistically significant on the basis of a multivariate statistical test.

If the analysis as reported by the study author is not properly aligned:

For a single outcome measure:

- The effect reported is positive and statistically significant, accounting for clustering.

For multiple outcome measures:

- Univariate statistical tests are reported for each outcome measure and the effect for at least one measure within the domain is positive and statistically significant and no effects are negative and statistically significant, accounting for clustering and multiple comparisons.
- The mean effect for the multiple measures of the outcome is positive and statistically significant, accounting for clustering.

Substantively important positive effect if the single or mean effect is not statistically significant, as just described, and either of the following is true:

For a single outcome measure:

- The effect size reported is positive and substantively important.

For multiple outcome measures:

- The mean effect size reported is positive and substantively important.

Indeterminate effect if the single or mean effect is neither statistically significant nor substantively important, as described earlier.

Substantively important negative effect if the single or mean effect is not statistically significant, as described earlier, and either of the following is true:

For a single outcome measure:

- The effect size reported is negative and substantively important.

For multiple outcome measures:

- The mean effect size reported is negative and substantively important.

Statistically significant negative effect if no statistically significant or substantively important positive effect has been detected and any of the following is true:

If the analysis as reported by the study author is properly aligned:

For a single outcome measure:

- The effect reported is negative and statistically significant.

For multiple outcome measures:

- Univariate statistical tests are reported for each outcome measure and at least half of the effects are negative and statistically significant.
- Univariate statistical tests are reported for each outcome measure and the effect for at least one measure within the domain is negative and statistically significant, accounting for multiple comparisons.
- The mean effect for the multiple measures of the outcome is negative and statistically significant.
- The omnibus effect for all the outcome measures together is reported as negative and statistically significant on the basis of a multivariate statistical test.

If the analysis as reported by the study author is not properly aligned:

For a single outcome measure:

- The effect reported is negative and statistically significant, accounting for clustering.

For multiple outcome measures:

- Univariate statistical tests are reported for each outcome measure and the effect for at least one measure within the domain is negative and statistically significant, accounting for clustering and multiple comparisons.
- The mean effect for the multiple measures of the outcome is negative and statistically significant, accounting for clustering.

APPENDIX H. COMPUTATION OF THE IMPROVEMENT INDEX

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

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

Specifically, the improvement index is computed as follows:

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

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

Mechanically, the conversion of an effect size to a U3 index entails using a table that lists the proportion of the area under the standard normal curve for different values of z-scores, which can be found in the appendices of most statistics textbooks. For a given effect size, U3 has a value equal to the proportion of the area under the normal curve below the value of the effect size—under the assumptions that the outcome is normally distributed and that the variance of the outcome is similar for the intervention group and the comparison group.

Compute Improvement Index = U3 – 50%

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

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

APPENDIX I. EXTENT OF EVIDENCE CATEGORIZATION

The Extent of Evidence Categorization was developed to tell readers how much evidence was used to determine the intervention rating, focusing on the number and sizes of studies. This scheme has two categories: small and medium to large.

The extent of evidence is medium to large if all of the following are true:

- The domain includes more than one study.
- The domain includes more than one school.
- The domain findings are based on a total sample size of at least 350 students OR, assuming 25 students in a class, a total of at least 14 classrooms across studies.

The extent of evidence is small if any of the following are true:

- The domain includes only one study.
- The domain includes only one school.
- The domain findings are based on a total sample size of less than 350 students AND, assuming 25 students in a class, a total of less than 14 classrooms across studies.

Each intervention domain receives its own categorization. For example, each of the three domains in character education—behavior; knowledge, attitudes, and values; and academic achievement—receives a separate categorization.

Example: Intervention Do Good, a character education intervention, had three studies that met WWC standards and were included in the review. All three studies reported on academic achievement. There were a total of six schools across the three studies. The first study reported testing on 150 students, the second study 125 students, and the third study reported testing four classes with 15 students in each class. The extent of evidence on academic achievement for the Do Good intervention is considered "medium to large"—it met the condition for both the number of studies and the number of schools, and although the total number of students is less than 350 (150 + 125 + [4*15] =335), the number of classes exceeded 14 (150/25 + 125/25 + 4 = 15).

A "small" extent of evidence indicates that the amount of the evidence is low. There is currently no consensus in the field on what constitutes a "large" or "small" study or database. Therefore, the WWC set the indicated conditions based on the following rationale:

- With only one study, the possibility exists that some characteristics of the study—for example, the outcome instruments or the timing of the intervention—might have affected the findings. Multiple studies provide some assurance that the effects can be attributed to the intervention and not to some features of the particular place where the intervention was studied. Therefore, the WWC determined that the extent of evidence is small when the findings are based on only one setting.

- Similarly, with only one school, the possibility exists that some characteristics of the school—for example, the principal or student demographics—might have affected the

findings or were intertwined or confounded with the findings. Therefore, the WWC determined that the extent of evidence is small when the findings are based on only a single school.

- The sample size of 350 was derived from the following assumptions:
  - A balanced sampling design that randomizes at the student level
  - A minimum detectable effect size of 0.3
  - The power of the test at 0.8
  - A two-tailed test with an alpha of 0.05
  - The outcome was not adjusted by an appropriate pretest covariate.

The Extent of Evidence Categorization provided in recent reports, and described here, signals WWC's intent to provide at some point a rating scheme on the external validity, or the generalizability, of the findings, for which the extent of evidence is only one of the dimensions. The Extent of Evidence Categorization, in its current form, is not a rating on external validity; instead, it serves as an indicator that cautions readers when findings are drawn from studies with small size samples, a small number of school settings, or a single study.