IES Institute of Education Sciences

VERSION 5.0

What Works Clearinghouse

Procedures and Standards Handbook, Version 5.0

WWC 2022008
U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Evaluation at IES

WHAT WORKS CLEARINGHOUSE PROCEDURES AND STANDARDS HANDBOOK, VERSION 5.0

August 2022

U.S. Department of Education

Miguel A. Cardona
Secretary

Institute of Education Sciences

Mark Schneider
Director

National Center for Education Evaluation and Regional Assistance

Matthew Soldner
Commissioner

Elizabeth Eisner
Associate Commissioner of Knowledge Use

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

What Works Clearinghouse. (2022). *What Works Clearinghouse procedures and standards handbook, version 5.0*. U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). This report is available on the What Works Clearinghouse website at https://ies.ed.gov/ncee/wwc/Handbooks.

Alternate Formats

The Alternate Format Center (AFC) produces documents in specific alternate formats, such as braille, large print, and electronic format, for employees and members of the public with disabilities. These documents may include agendas, correspondence, course materials, regulations, and materials for public distribution.

Contact the AFC for submission guidelines and turnaround times. Hours of operation are Monday through Friday, 8:00 a.m. until 4:00 p.m., Eastern Time. For information, contact the AFC, Tena Witherspoon, Tena.Witherspoon@ed.gov, (202) 260-0818, or Tracey Flythe, Tracey.Flythe@ed.gov, (202) 260-0852.

U.S. DEPARTMENT OF EDUCATION

WWC 2022008

IES Institute of Education Sciences

WWC Contractors

Molly Cain
Sarah Caverly
Alicia Garcia
Natalya Gnedko-Berry
Daniel Hubbard
David Miller
Joshua Polanin
Jordan Rickles
Sarah Sahni

Lisa Shimmel
Joseph Taylor
Ryan Williams
American Institutes for Research

Daniel M. Swan
Emily Tanner-Smith
University of Oregon

Jeffrey C. Valentine
University of Louisville

Project Officers

Erin Pollard
Betsy Wolf
Jonathan Jacobson
Institute of Education Sciences

Statistical, Technical, and Analysis Team

The What Works Clearinghouse (WWC) standards and procedures are developed in consultation with methodological experts and WWC content contractors through the Statistical, Technical, and Analysis Team. The team meets in small groups and as a large group to discuss advances in research methods in relation to the WWC procedures and standards.

Jack Buckley
American Institutes for Research

John Ferron
university of South Florida

Michael S. Garet
American Institutes for Research

Russell Gersten
Instructional Research Group

Ben B. Hansen
University of Michigan

Fran Harmon
Development Services Group

Larry V. Hedges
Northwestern University

Wendy Machalicek
University of Oregon

Rebecca Maynard
University of Pennsylvania

Hiren Nisar
2M Research

Terri Pigott
Georgia State University

Allan Porowski
Abt Associates

James E. Pustejovsky
University of Wisconsin-Madison

David Rindskopf
City University of New York

Jessaca Spybrook
Western Michigan University

Emily Tanner-Smith
University of Oregon

Elizabeth Tipton
Northwestern University

Jeffrey C. Valentine
University of Louisville

Elias Walsh
Mathematica

Vivian Wong
University of Virginia

Contents

Tables

Figures

WELCOME TO THE *WHAT WORKS CLEARINGHOUSE PROCEDURES AND STANDARDS HANDBOOK, VERSION 5.0*

Education decisionmakers need access to the best evidence about the effectiveness of education interventions, including practices, products, programs, and policies. It can be difficult, time consuming, and costly to access and draw conclusions from relevant studies about the effectiveness of interventions. The What Works Clearinghouse (WWC) addresses the need for credible, succinct information by identifying existing research in education, assessing the quality of this research, and summarizing and disseminating the evidence from studies that meet WWC standards.

This *WWC Procedures and Standards Handbook, Version 5.0*, provides a detailed description of how the WWC reviews studies that meet eligibility requirements for a WWC review. Key differences between the current version of WWC procedures and standards (5.0) and the previous version (4.1) are summarized below. These differences focus on technical and procedural nuances and are most relevant for WWC reviewers. *Summary of Changes* on the WWC website provides a detailed summary of all changes.

Version 5.0

Version 5.0 of the *Handbook* replaces the two documents used since October 2020, the *What Works Clearinghouse Procedures Handbook, Version 4.1*, and the *What Works Clearinghouse Standards Handbook, Version 4.1*. The WWC chose to combine procedures and standards into one document to improve the Handbook's usability to education practitioners, WWC reviewers, and others who rely on the WWC to assess the quality of education research.

Major technical and procedural changes between versions 5.0 and 4.1 of WWC procedures and standards

- Under previous versions, the handbooks articulated the standards for study reviews, but topic area review teams had the ability to customize certain aspects of the standards. Under version 5.0, **the WWC no longer allows for topic-specific customization of the standards**. The WWC shifted to a uniform application of the standards because allowing topic area customization of some standards resulted in the same study having multiple and sometimes different WWC ratings, creating inconsistency and confusion. In some instances, the WWC standards include different options for application, with the choice dependent on the circumstances of individual studies. Teams reviewing individual studies will decide on the most appropriate option for each study. For example, determining joiner risk is now a review team decision based on the circumstances of individual studies but is not a topic area decision.
- Under version 5.0, **all WWC study reviews will be conducted according to the *WWC Procedures and Standards Handbook* and complemented by the *Study Review Protocol***, including individual study reviews and reviews of studies included in evidence synthesis products, such as practice guides and intervention reports. The *Study Review Protocol* articulates information to supplement the *Handbook*, such as how outcome measures should be grouped into outcome domains. All studies reviewed individually and as part of evidence synthesis products will be reviewed using the *Study Review Protocol* to increase review consistency, transparency, and efficiency. Topic area synthesis protocols will continue to be used to provide criteria for the literature search; guidance on how to identify and prioritize relevant studies for review and

inclusion in evidence synthesis products; and guidance on intervention, sample, and outcome eligibility criteria for the synthesis.

- **The WWC aligned its effectiveness ratings with U.S. Department of Education evidence definitions** for individual studies and synthesis products. The WWC modified effectiveness ratings to include evidence definitions to streamline the identification of effective interventions.
- Under version 4.1, the study or intervention effectiveness rating was the highest rating obtained from any main finding, with *main findings* usually being findings for the full study sample at the end of intervention. Under version 5.0, **the WWC determines effectiveness ratings at the outcome domain level**. If a study has multiple main findings in the same domain, the WWC creates a composite finding and reports the effectiveness rating for the domain-level composite. The WWC shifted to domain-level composites because a synthesis of multiple findings in the same domain will tend to provide a better representation of the underlying construct than will any single measurement. In addition, this change simplifies WWC procedures by eliminating the need for multiple comparison corrections.
- A new procedure under version 5.0 allows the WWC's effectiveness ratings for individual study reviews to be based on **outcome measures that are independent** of intervention developers and study authors. This procedure reflects the WWC's concern that potentially meaningful differences in effect sizes can be obtained from measures created by intervention developers or study authors, and that these measures may not be as informative to policymakers and practitioners as independent measures. Therefore, in outcome domains that have a relatively plentiful number of recognized, widely accepted, and independent measures, the WWC review will focus on those measures for reporting on an intervention's effectiveness. The *Study Review Protocol* will identify outcome domains for which the WWC will use independent measures to assess effectiveness. Studies that use nonindependent measures can still meet WWC standards. Nonindependent measures can contribute to a cross-study synthesis (for example, intervention reports, practice guide) when the need for nonindependent measures is documented in the topic area review protocol, but nonindependent measures will not contribute to effectiveness ratings in individual study reviews.
- Under version 5.0, when a cross-study synthesis includes findings rated *Meets WWC Standards Without Reservations* and *Meets WWC Standards With Reservations* and the sample size is sufficiently large, the **WWC will attempt to ensure that a majority of the meta-analytic weight is based on findings rated *Meets WWC Standards Without Reservations***. The rationale for this change is to ensure that findings from the most rigorously designed studies receive the most weight in the synthesis.
- Under version 5.0, individual-level and cluster-level **high attrition randomized controlled trials (RCTs) and high attrition regression discontinuity design (RDD) studies no longer need to demonstrate baseline equivalence** to be rated *Meets WWC Standards With Reservations* when attrition bias is assessed using the optimistic boundary. Study authors only need to use an acceptable adjustment strategy in the impact analysis. The WWC allows this flexibility because while attrition can undermine the validity of an estimated intervention effect, strong control over the assignment mechanism (through randomization in RCTs or a forcing variable cutoff in RDDs) often provides a reasonable basis for statistical procedures that attempt to adjust for the potentially biasing effects of attrition. This change also allowed the WWC to bring RCT and RDD reviews into closer alignment than in previous versions of the standards.

- Under previous versions of the standards, review protocols determined the choice between the optimistic and cautious attrition boundaries. Under version 5.0, **teams conducting WWC reviews are responsible for determining whether to use an optimistic or a cautious attrition boundary** for a specific review and for documenting their reasoning based on the principles described in the Handbooks. If review teams find that they cannot defensibly choose between the optimistic and cautious boundaries, then they should use the cautious attrition boundary because when there is doubt, a more cautious approach is warranted. The WWC standards allow for review teams' choice of the attrition boundary because the applicability of optimistic or cautious attrition assumptions depends on the circumstances of individual studies, which review teams are best positioned to evaluate against the WWCs' guidance on selecting the attrition boundary.
- The WWC has **removed procedures for WWC-applied difference-in-difference adjustments**, which had previously allowed the WWC to use reported baseline information to adjust effect sizes based on unadjusted outcome statistics. If a study requires baseline adjustment to meet WWC standards, then the study authors must be the ones to apply any required adjustment, not the WWC. The WWC will continue to report effect sizes based on unadjusted statistics if adjusted statistics are unavailable and adjustment for baseline differences was not required, such as for low-attrition RCTs. The change to remove WWC-applied difference-in-differences adjustments aligns with an overall principle in version 5.0 of the WWC *Handbook* of greater transparency by increasing correspondence between the effects reported by the study authors and the WWC.
- **The WWC no longer considers bundled—or combined—interventions a confounding factor in reviews of individual studies** because a bundled intervention can produce a valid impact estimate for the "package" of interventions, provided they are eligible for WWC review. For a topic area synthesis, a bundled intervention will remain problematic if any of the bundled interventions do not meet the definition of eligibility as articulated in the topic area synthesis protocol. Under this circumstance, a bundled intervention may be excluded from a synthesis product.
- **The WWC now classifies cluster RCTs as having either a low or high risk of bias due to compositional change from joiners** (individuals who enter intervention or comparison clusters after the clusters' assignments to conditions is known outside of the study team). The change eliminates a prior distinction about late versus early joiners, and provides explicit guidance about when including joiners in the analytic sample should not affect the study's research rating. Review teams will characterize the risk of bias due to joiners based on three factors: (a) the unit of assignment, (b) the unit of measurement, and (c) the potential for the intervention to affect joining. Review teams will use a similar set of factors to characterize the risk of bias due to leavers, which guides the choice of the attrition boundary and an acceptable reference sample for determining individual-level attrition. The WWC made these changes to simplify the cluster-level assignment standards and decrease ambiguity in applying them.
- Under version 5.0, single-case design studies (SCDs) that use multiple baseline/multiple probe, treatment reversal, and changing criterion designs need to have **at least six data points in the initial baseline phases** for their findings to be eligible for the rating *Meets WWC Standards Without Reservations*. Previous versions of the standards required at least five data points per phase for designs to be eligible for the rating

Meets WWC Standards Without Reservations. This change was intended to ensure there is sufficient opportunity to understand the initial pattern of responding in these designs.

- Version 5.0 introduces **an exception for minimum data point requirements** for SCDs that use multiple baseline/multiple probe, treatment reversal, and changing criterion designs. Any phases with three or more data points and zero within-phase variance, including the initial baseline phase, are considered to have sufficient data points to be eligible for the rating *Meets WWC Standards Without Reservations*. The WWC made this change because additional data points would likely not improve a study's design under these circumstances.
- Under version 5.0, the WWC modified **the interobserver agreement requirements for SCDs to apply to all data** in the study rather than to specific conditions as was the case under the previous version of WWC procedures and standards. The WWC made this change to better align the standards with practice in high-quality SCDs.
- The WWC **added a new "limit risk of bias" step to the review process** for multiple baseline/multiple probe, treatment reversal, and changing criterion SCDs that are eligible for the rating *Meets WWC Standards Without Reservations*. This process uses the nonoverlap of all pairs as a decision rule that is intended to be analogous to some of the visual-analytic judgments that are used to assess the internal validity of SCDs.
- The WWC provided **updated guidance on how to rate SCDs** with features from multiple design types and SCDs with more cases and/or phases than the minimum required to meet WWC standards. Under version 5.0, a study is typically eligible to receive the highest rating that any subset of cases or phases is eligible to receive. The WWC made this change to ensure that SCDs that include information above and beyond what is required by the standards are not penalized for reporting more data than studies that report the minimum data required, and to allow study authors more flexibility to design studies using a combination of design features. This change also brings SCD study ratings into closer alignment with group design study ratings.

Content and organization of the *WWC Procedures and Standards Handbook, Version 5.0*

The *Handbook* is organized such that most frequently used information appears in earlier chapters. The need for technical knowledge of research design increases in subsequent chapters.

- **Chapter I** provides a general overview of WWC procedures and standards. The overview is intended for readers who need a working knowledge of how the WWC reviews studies but who will not conduct study reviews or design studies intended to meet WWC standards.
- **Chapter II** describes procedures for screening studies for eligibility.
- **Chapter III** describes procedures and standards for reviewing findings from randomized controlled trials and quasi-experimental designs.
- **Chapter IV** describes procedures and standards for reviewing findings from studies that use a regression discontinuity design.

- **Chapter V** focuses on reviewing findings from group design studies that use advanced methodological procedures, such as randomized controlled trials that estimate complier average causal effects, and analyses that impute missing data.
- **Chapter VI** describes procedures and standards for reviewing findings from single-case design studies.
- **Chapter VII** describes procedures for synthesizing and characterizing findings from reviews of individual studies and in intervention reports and practice guides.

The *Handbook* concludes with technical appendices. These appendices provide details on the procedures underlying the review process; for example, the calculation and estimation of effect sizes and other computations used in WWC reviews. In addition, the technical appendices include information on procedures underlying the development of WWC products, such as how the WWC identifies studies to include in intervention reports and practice guides. The WWC provides technical appendices to ensure that the methods and procedures used within a given review are replicable and transparent.

As the WWC continues to refine and develop procedures and standards, *WWC Procedures and Standards Handbook, Version 5.0,* may be revised or supplemented to reflect these changes. Any written supplements for use in combination with this *Handbook* will be specified in the most recent version of the *Study Review Protocol*.

Readers who have questions about WWC procedures and standards or who want to provide feedback on the *Handbook* may contact the WWC Help Desk at https://ies.ed.gov/ncee/wwc/help.

CHAPTER I. OVERVIEW OF THE WHAT WORKS CLEARINGHOUSE AND ITS PROCEDURES AND STANDARDS

Chapter I of the *What Works Clearinghouse (WWC) Procedures and Standards Handbook, Version 5.0*, presents a high-level overview of the WWC, how it works, and the principles underlying WWC standards. This chapter is intended for a broad audience of practitioners and researchers.

What the WWC is, what it does, and why

The WWC is an initiative of the Institute of Education Sciences (IES) within the U.S. Department of Education. **The WWC's mission is to be a central and trusted source of scientific evidence for what works in education.** The WWC reviews relevant research, identifies well-designed and well-implemented impact studies, summarizes the findings from those studies, and disseminates them to the public. The goal of the WWC is to help educators, administrators, families, researchers, and policymakers make evidence-based decisions.

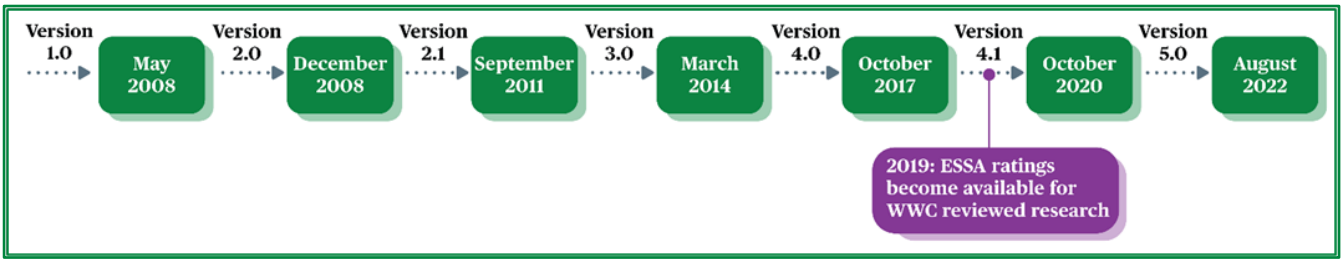
The WWC's products are available on the WWC website, including reviews of individual studies, as well as intervention reports and practice guides that synthesize evidence across multiple studies. The website includes information to help the education community identify research that is both effective and relevant, such as effectiveness ratings and grade levels in which studies were conducted. The WWC provides this information to ensure that education decisionmakers have quick access to the evidence they need.

WWC history

The WWC was founded in 2002 by the Office of Educational Research and Innovation at the U.S. Department of Education. The office formed the WWC to help schools, districts, and states identify "scientifically based research" required by the No Child Left Behind Act of 2001. In November 2002, Congress passed the Education Sciences Reform Act that created the Institute of Education Sciences, which took over responsibility for the WWC. To support its mission, the WWC developed procedures and standards for identifying and reviewing research on the impacts of education interventions on student and educator outcomes. The WWC continues to refine its procedures and standards based on improvements in education research and research synthesis methods. The WWC also refines its procedures and standards to meet the needs of education decisionmakers. For instance, in 2019 the WWC incorporated the evidence tiers introduced by the Every Student Succeeds Act (ESSA) of 2015 into its products to provide education decisionmakers a convenient place to find research aligned with at the U.S. Department of Education evidence definitions (34 CFR, Part 77).

Figure 1 shows the timeline for the revisions of the WWC procedures and standards. All previous versions of the WWC procedures and standards are available in the Handbooks and Other Resources section on the WWC website. A detailed summary of the changes in the *WWC Procedures and Standards Handbook, Version 5.0,* is available in *Summary of Changes*.

Figure 1. *Timeline of the What Works Clearinghouse Procedures and Standards Handbook*



WWC roles and responsibilities

Translating scientific evidence into accurate, accessible, and useful information for educators requires collaboration across individuals in multiple roles, including IES leadership, methodologists, study reviewers, peer reviewers of WWC products, practice guide panelists, and many more.

WWC reviewers are the backbone of the WWC: Reviewers are responsible for reviewing and interpreting research and extracting or developing information for presentation on the WWC website. A typical WWC reviewer has a graduate degree in education or the social sciences, has a strong background in educational research methodology, and demonstrates a willingness to pay attention to small study details. WWC reviewers ensure that each reviewed study receives fair and equitable treatment under WWC standards. The work of reviewers is supported by review team leadership, as well as content experts, methodologists, and IES leadership.

WWC products

When the WWC is ready to publish a product that synthesizes evidence across multiple studies, such as an intervention report or a practice guide described in the next section, external methodologists and content experts peer review the product to ensure quality. The WWC publishes each product on the WWC website. For select products, the WWC also communicates the findings through summaries, webinars, infographics, and other materials. The WWC currently produces three products: reviews of individual studies, intervention reports, and practice guides. The key characteristics of each product are in table 1.

Table 1. *What Works Clearinghouse products*

Category	Reviews of individual studies	Intervention reports	Practice guides
Purpose	Summarize evidence from an individual study.	Synthesize evidence for an intervention based on a systematic review of studies that examined the intervention. The intervention is often a "branded" program or product—that is, a commercial program or product.	Synthesize evidence to identify teaching methods, learning strategies, and other approaches to learning that may improve educational outcomes. Practice guides may use evidence from "branded" programs or products, but they do not focus on them.
Effectiveness ratings including evidence tier designation[a]	Effectiveness rating for each outcome domain in the study for which a main finding meets WWC standards (includes Tier 1, Tier 2, and Tier 3 evidence).	Effectiveness rating for each outcome domain based on studies of the intervention for which main findings meet WWC standards (includes Tier 1, Tier 2, and Tier 3 evidence).	Effectiveness rating for each practice recommendation based on the syntheses of findings meeting WWC standards and informing the recommendation (includes Tier 1, Tier 2, Tier 3, and Tier 4 evidence).
Key audience	State and district administrators; curriculum and intervention developers; and researchers.	State, district, and school administrators; curriculum and intervention developers; researchers; and instructional leaders.	State, district, and school administrators; educators; instructional leaders; and professional development providers.

a. The WWC uses evidence definitions from the Every Student Succeeds Act, or ESSA, together with the U.S. Department of Education's general administrative regulations (34 CFR, Part 77) to assess favorable effects from an intervention, taking into account sample size, multiple settings, and the absence of overriding, negative effects.

WWC training and reviewer certification

All studies are reviewed by WWC-certified reviewers. To become a WWC-certified reviewer, individuals must successfully complete WWC training. The WWC certifies individuals on three sets of standards: group designs, advanced group designs, and single-case designs (SCDs). The group design training covers an introduction to the WWC, randomized controlled trials (RCTs), quasi-experimental designs (QEDs), and previews other designs. Topics covered in this training include an overview of the WWC and its products and in-depth instruction on the WWC review standards, procedures, and review tools. This information is relevant for all reviews of research by the WWC. The advanced group design training covers regression discontinuity designs (RDDs) and certain advanced analytical approaches, including complier average causal effect estimates from RCTs. Last, the WWC offers a training on SCDs, which are the only designs without a comparison group eligible for a WWC review.

Within each training, trainees pursuing certification are expected to take and pass a multiple-choice certification exam, which includes questions about how to apply the standards using examples from studies reviewed by the WWC. The exam is separated into sections, and trainees have four chances to pass each section: an initial attempt plus three retakes. Table 2 provides a summary of the WWC's certification training and requirements for version 5.0. Individuals interested in becoming certified should check the WWC website for the most recent

training and certification requirements. The certification trainings are available online, and there is no charge to access them or to pursue WWC certification.

Table 2. *What Works Clearinghouse certification training content and requirements*

Category	Group design standards training	Advanced group design standards training	Single-case design standards training
Content	• General WWC procedures and standards, review tools, policies, and practices. • Procedures and standards for randomized controlled trials. • Procedures and standards for quasi-experimental designs.	• Procedures and standards for regression discontinuity designs. • Procedures and standards for studies using certain advanced analytical approaches, such as complier average causal effect estimates from randomized controlled trials.	• Procedures and standards for single-case designs.
Prerequisites	None.	Group design certification.	Group design certification.
Exam	Multiple-choice certification test, including questions that simulate a study review (four attempts for each section of the exam).	Multiple-choice certification test, including questions that simulate a study review (four attempts within each section of the exam).	Multiple-choice certification test, including questions that simulate a study review (four attempts within each section of the exam).
Online access (5.0 updates available early 2023)	https://ies.ed.gov/ncee/wwc/OnlineTraining	Training for advanced group design standards is forthcoming.	https://ies.ed.gov/ncee/wwc/SingleCaseTraining

In addition to training and certifying reviewers on the three sets of standards shown in Table 2, the WWC offers training on key procedures and standards for general group design studies. This training is intended for individuals who are interested in this content but do not wish to pursue a full WWC certification in group designs. The content of this training includes the core topics that underlie reviews of group designs, such as assignment to intervention and comparison conditions, compositional change, baseline adjustment, and outcome requirements. Individuals who pursue this training option will view the first series of modules of the group designs series.

How the WWC conducts study reviews

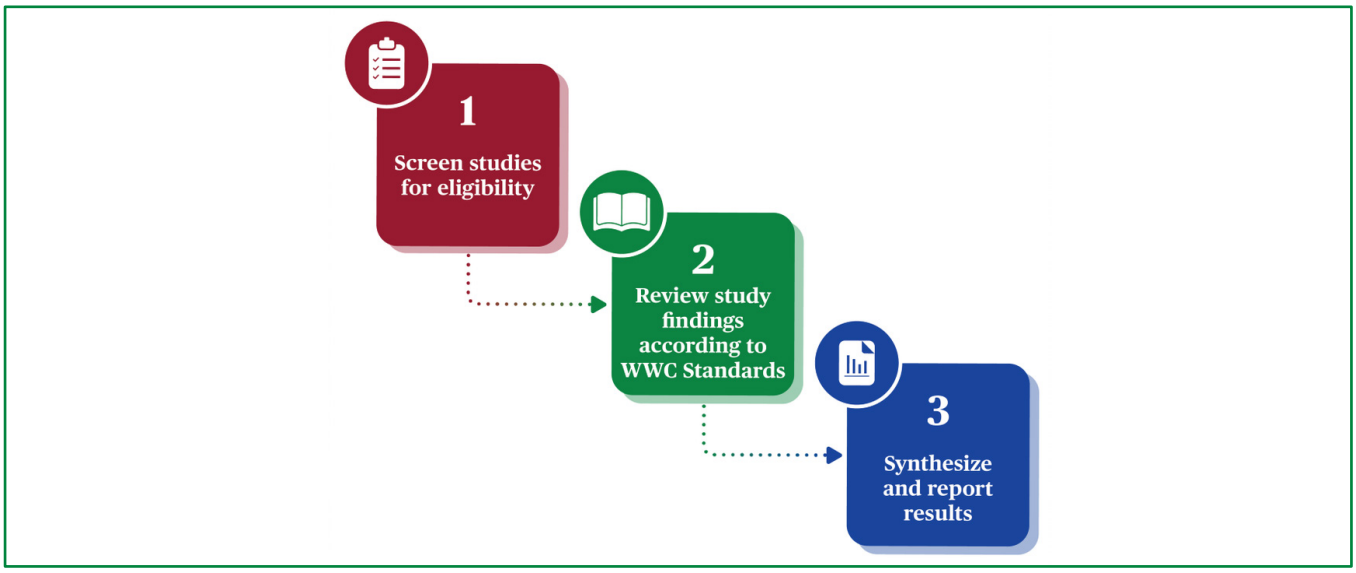
Individual study reviews are the basis for all WWC products. The review process enables the WWC to use consistent, objective, and transparent procedures and standards in its reviews. The WWC's review process consists of the three phases shown in figure 2.

All WWC study reviews are conducted using the *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0*. The WWC uses the *Study Review Protocol* to review individual studies and studies reviewed as part of synthesis products, such as practice guides. Review teams also use relevant topic areas synthesis protocols for evidence synthesis products. Topic area synthesis protocols provide criteria for the literature search, guidance on how to identify and prioritize relevant studies for review and inclusion in synthesis products, and provide guidance on intervention, sample, and outcome eligibility criteria. WWC protocols are available on the WWC website.

How the WWC decides what to review

The WWC prioritizes topics for intervention reports and practice guides based on their applicability to a broad range of students, policy relevance, potential to improve educational outcomes, and the availability of research. The WWC also reviews individual studies outside of intervention reports and practice guides, for example, when studies have the potential to provide Tier 1, Tier 2, or Tier 3 evidence as defined by the U.S. Department of Education. Appendix A includes detailed information on how the WWC prioritizes topics and studies to review.

Figure 2. *Three phases in the What Works Clearinghouse study review process*



Phase 1: Screen studies for eligibility

The studies WWC identifies for review are screened for eligibility as the first phase in the review process. The criteria include whether the study used an eligible design, was published in an eligible time frame, included an eligible sample, included an eligible outcome measure, was conducted in an eligible location, and met other relevant criteria. Studies must meet each eligibility criterion to be eligible for WWC review. Chapter II, Screening

Studies for Eligibility, describes in greater detail how the WWC determines the eligibility of education research for WWC review.

Phase 2: Review study findings according to WWC standards

Eligible studies advance to the next phase in the review process, during which the eligible findings from studies are assessed according to WWC standards. The WWC examines several research features to determine whether the study's findings can be attributed to the intervention. The process used to decide who receives the intervention and who does not is a critical study feature that affects the WWC's level of confidence in the study's claim that the intervention produced the observed findings. Sometimes, the WWC examines whether groups of participants who receive and do not receive the intervention were similar when the study began and remained similar throughout the study's duration. Whether the study authors used trustworthy outcome measures and prevented factors that could have interfered with the study's findings also contribute to the WWC's confidence in its findings.

Phase 2 of the study review carries considerable weight in the review process because the WWC makes consequential decisions about the study that will affect the research ratings of its findings (ratings are described in the next section). For these reasons, usually several reviewers contribute to each study review, supported by review team leadership, and in some instances content experts, methodologists, and external experts as described in Chapter I, WWC roles and responsibilities. The features of the study that the WWC reviews in phase 2 are described in more detail Chapter I, How research can meet WWC standards.

Phase 3: Synthesize and report results

In the final phase of the review process, the WWC synthesizes and reports two sets of results: (a) a research rating based on the strength of the research design and its execution, and (b) an effectiveness rating based on the evidence of favorable effects from the intervention.

Determining a WWC research rating based on the strength of research design and execution

Based on the WWC's confidence in the strength of the research design and execution of the design, eligible findings from the study will receive one of three research ratings: *Meets WWC Standards Without Reservations*, *Meets WWC Standards With Reservations*, or *Does Not Meet WWC Standards*. Table 3 describes each rating.

Table 3. *What Works Clearinghouse research ratings*

Rating level	Description
	The highest rating a finding can receive is ***Meets WWC Standards Without Reservations***. This rating is reserved for findings based on a strong research design that is well-executed. This rating therefore provides the highest degree of confidence that the intervention caused the observed effect.
	The second-highest rating a finding can receive is ***Meets WWC Standards With Reservations***. Because of natural limitations in research designs or because of circumstances around execution of a design, findings that receive this rating do not sufficiently rule out that something other than the intervention caused the observed effect.
Does Not Meet WWC Standards	The lowest research rating is ***Does Not Meet WWC Standards***. Findings that receive this rating are not accompanied by sufficient evidence that the intervention caused the observed effect.

WWC research ratings and the Standards for Excellence in Education Research

WWC research ratings focus primarily on the validity of education study findings for causal inferences. The Standards for Excellence in Education Research are a set of broader IES-wide principles, distinct from WWC standards, to encourage and acknowledge high-quality education research studies along several additional dimensions, such as preregistration of research questions and analysis plans, documentation of core components of the intervention and of the counterfactual condition, description of the characteristics of the analytic sample, and reporting of cost information. For more information about the Standards for Excellence in Education Research principles and their use across IES, visit https://ies.ed.gov/seer.

Determining an effectiveness rating based on the evidence of the intervention's effects

For studies rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*, the WWC will synthesize and report the corresponding findings to characterize the effectiveness of the intervention by outcome domain. To do that, the WWC relies on effect sizes and their statistical significance. In general, the WWC uses effect sizes to determine the direction (favorable or unfavorable) and magnitude (the size) of the findings. Effect sizes allow the WWC to describe the average difference between an intervention and a comparison condition in a standardized way. This is especially important for WWC systematic reviews and evidence syntheses that combine findings in the same outcome domain but that were assessed using different outcome measures. The WWC uses statistical significance of domain-level findings to highlight positive or negative effects as opposed to uncertain effects.

In addition to effect sizes and statistical significance, the WWC uses evidence definitions from the Every Student Succeeds Act, or ESSA, together with the U.S. Department of Education's general administrative regulations (34 CFR, Part 77) to assess favorable effects from an intervention, taking into account sample size, multiple settings, and the absence of overriding, negative effects. Wherever appropriate, the WWC assigns evidence tiers to findings from individual studies, intervention reports, and recommendations from practice guides. The WWC

incorporates evidence tiers into its reporting to increase the usability of findings by education decisionmakers who often need to identify evidence that aligns with the Department's evidence definitions. How the WWC synthesizes and reports evidence of effectiveness is described in detail in Chapter VII, Synthesis and Reporting of Results. The main features are summarized below.

Summary of evidence for individual studies and intervention reports. The WWC characterizes evidence from individual studies and intervention reports by outcome domain. If an outcome domain includes several main findings, which are usually findings for the full study sample at the end of intervention, the WWC will synthesize across main findings. For intervention reports, this approach allows the WWC to combine every effect size in the same outcome domain across all studies that examined the same intervention into a single average that accounts for differences between studies, such as their sample size. Then the WWC provides an effectiveness rating for the intervention for each outcome domain based on the direction of the average effect, its statistical significance, whether the findings are from large samples and multisite samples, and the research ratings of the findings meeting WWC standards. The effectiveness ratings include strong evidence (Tier 1), moderate evidence (Tier 2), promising evidence (Tier 3), uncertain effects, and negative effects.

Summary of evidence for practice guides. Practice guides provide recommendations that combine the empirical literature and expert opinion. The recommendations focus on education strategies and practices, for example writing strategies or teaching math to young students, rather than on characterizing the evidence for specific, "branded" products or programs. The WWC assigns effectiveness ratings to practice guide recommendations as supported by strong evidence (Tier 1), moderate evidence (Tier 2), promising evidence (Tier 3), or notes when evidence demonstrates a rationale for a recommendation (Tier 4). The effectiveness ratings for each recommendation are based on the evidence of statistically significant and positive effects for relevant outcome domains, whether the findings are from large samples and multisite samples, the research ratings of the findings meeting WWC standards, and relevance to the scope and context of the practice guide recommendation.

How research can meet WWC standards

Phase 2 in the WWC systematic review process involves assessing an eligible study according to WWC standards. Phase 2 is the most intensive phase of the review in which a reviewer must examine the critical features of a study's research design to determine the credibility of its findings and assign findings research ratings based on the strength of the research design and execution. Because of the weight that this phase carries in the review process, this section provides additional information on the research features that the WWC evaluates. The goal of additional information is to equip practitioners and researchers with a working knowledge of what the WWC looks for in the eligible studies it reviews. Detailed information about how the WWC evaluates studies is available in Chapters III through VI of the *Handbook*.

When reviewing an eligible study according to WWC standards, the WWC examines several of its critical features. Depending on the research design, these features include outcome measures, presence of confounding factors in the study, the process of assignment to condition, compositional change in study conditions, baseline equivalence between study conditions, and adequate baseline adjustment.

Outcome measure standards

For a study to meet WWC standards, it must have at least one eligible finding that was measured in a way that satisfies the WWC's outcome measure standards. The WWC has four standards for an outcome measure: It should have face validity (the measure appears to measure what it claims measure), it should demonstrate reliability (the measure produces consistent findings), it should not be overaligned with the intervention (give unfair advantage to participants in one condition over another), and it should be measured consistently for the groups or participants being compared. The WWC also will examine the independence of outcome measures from study authors and from the developers of the intervention, who may have conflicts of interest regarding the success of the intervention. The *Study Review Protocol* specifies outcome domains in which the WWC will consider measure independence. Detailed information on how the WWC reviews outcome measures is available in Chapter III, Outcome measures.

The WWC examines all findings in a study using the *Handbook* and the *Study Review Protocol* to determine whether any of the findings in the study are eligible for WWC review and are assessed using an outcome measure that meets the WWC's outcome measure standards, such as face validity, reliability, not being overaligned with the intervention, and being measured consistently for the groups or participants being compared. A study must have at least one eligible finding assessed using an outcome measure that meets WWC standards for the study to be rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*. If a study has eligible findings but none of them are assessed using measures that meet WWC outcome measure standards, then the study will be rated *Does Not Meet WWC Standards,* and the review will stop.

Confounding factors

The goal of the studies reviewed by the WWC is to identify the effect of the intervention on outcomes. The ideal study, perhaps only feasible in a lab setting, isolates the intervention's effect by eliminating the possibility that factors outside the intervention could cause a change in outcomes. Although it is impossible to eliminate every outside factor when conducting research in most educational settings, it is possible to observe the characteristics of the participants and their settings to determine whether one or both could have caused the observed change. The presence of such a characteristic is considered a confounding factor. If the WWC determines a confounding factor is present, the study will receive a rating of *Does Not Meet WWC Standards* and the review will stop.

The WWC looks for three confounding factors in group design studies briefly described next. Detailed information about confounding factors in group design studies is available in Chapter III, Confounding factors.

- **Intervention or comparison group contains a single unit, such as teacher, classroom, school, or district (also known as an *N* = 1 confounding factor).** This confounding factor involves situations in which the intervention or comparison condition contains a single teacher, classroom, school, or district, and that unit is aligned with only one of the study's conditions—that is, represented in only one condition. An example is a study that assigns one school to the intervention condition but two schools to the comparison condition. However, if the study is conducted in one school but students within the school are assigned to intervention and comparison conditions, one school will not be a confounding factor because it is aligned with both study's conditions.

- **Characteristics that can affect outcomes differ between conditions with no overlap.** A second type of confounding occurs when the characteristics of the students in the intervention or comparison group differ systematically in ways that are associated with the outcome. For example, if an intervention is delivered to students in grade 3 and the comparison group is composed of students in grade 2, and if grade level is related to the outcome, then it is difficult to disentangle the effects of the intervention from the effects of grade level. However, if the intervention is delivered to students in grade 3 and some students in grade 2, then the study will not have a confound, because the grade level overlaps with the study conditions.
- **Time is aligned with one condition with no overlap with the other condition.** Time can be a confounding factor when it is aligned completely with only one of the study conditions. For instance, if an intervention group of students in grade 3 and a comparison group of students in grade 3 are observed during different academic years, then time could be a confounding factor. However, when the time period of the intervention condition overlaps with the time period of the comparison condition, the WWC does not consider it a confound.

Confounding factors in RDDs and SCDs are similar to those in group design studies but with some differences described in Chapter IV, Reviewing outcome measures and checking for confounding factors in RDD studies and Chapter VI, Confounding factors (SCDs).

Assignment to conditions

If the study has at least one finding assessed using an outcome measure aligned with the WWC standards and does not contain a confounding factor, the WWC will review additional aspects of its design. A key factor in the WWC research rating is how participants were assigned to conditions–that is, how the intervention (those who receive the intervention) and comparison (those who do not receive the intervention) conditions were formed. How the conditions were formed determines the WWC's initial confidence in the study's claims regarding the intervention's effect on the observed outcomes.

The research design reflects how participants were assigned to conditions. The WWC currently only reviews studies that use one the following designs: RCTs, QEDs (including cross-sectional group designs, comparative interrupted time series, difference-in-difference designs, and growth curve designs), RDDs, and SCDs.[1] RCTs, QEDs, and RDDs are group designs in which participants, usually students, are assigned to intervention and comparison conditions. SCDs are the only non-group designs currently eligible for WWC review. In SCDs, each case, usually a student, serves as its own intervention and comparison condition. The study authors may observe a student several times before the intervention and several times during the intervention to determine whether the intervention changed outcomes. SCDs usually include a handful of participants, whereas group design studies often include groups of several dozen or more participants in each condition. Table 4 shows each research design eligible for a WWC review and the key features of its assignment process, along with the highest possible rating that each design could receive.

[1] The WWC's definition of QED differs from other uses of this term. Shadish et al. (2002) defined QEDs as experiments that do not employ random assignment, which includes RDDs.

When the WWC does not review certain QEDs, it is typically for one of two reasons.

- **Contain a confounding factor.** The WWC does not review single group pretest-posttest designs, which compare outcomes for the same group of participants before and after the introduction of an intervention. Single group pretest-posttest designs include a confounding factor between time and study condition (time confounding factor is discussed in the previous section). These designs are ineligible for WWC review as group design studies, as they are unable to provide trustworthy evidence of effectiveness. If the WWC were to make these group design studies eligible for review, they would always receive the research rating of *Does Not Meet WWC Standards*.

- **Review standards are currently unavailable.** Some QEDs, such as interrupted time series designs without a contemporaneous comparison group, could provide credible evidence of effectiveness. However, the WWC has not yet developed standards for reviewing studies using these designs. The only designs eligible for WWC review in which outcomes need not be measured at the same time under both study conditions are SCDs, as these designs include the control of assignment to condition by researchers and the monitoring of outcomes before and after the introduction of the intervention.

How intervention and comparison groups are formed—that is, the methods or processes used to create the intervention and comparison conditions—is a critical component of a study. Because researchers and administrators planning RCTs, RDDs, and SCDs exercise strong control over participant assignment to conditions, these designs are eligible for the WWC's highest research rating, *Meets WWC Standards Without Reservations*. U.S. Department of Education evidence definitions (34 CFR, Part 77) consider all three designs–RCTs, RDDs, and SCDs–to be "experimental" designs because of their use of *controlled* assignment.

Table 4. *Research designs reviewed by the What Works Clearinghouse and the highest research rating they are eligible to receive*

Research design	Assignment to condition	Highest possible research rating
Randomized controlled trial (RCT)	Assignment to intervention and comparison conditions is determined through random assignment before the intervention begins.	*Meets WWC Standards Without Reservations*
Regression discontinuity design (RDD)	Study authors create intervention and comparison conditions using a forcing variable, such as a test score, measured before the intervention.	*Meets WWC Standards Without Reservations*
Quasi-experimental design (QED)	Assignment to intervention and comparison conditions is determined through mechanisms other than random assignment or using a forcing variable (for instance, student assignment to conditions may be based on teacher referrals).	*Meets WWC Standards With Reservations*
Single-case design (SCD)	Study authors determine when a participant receives the intervention and observe changes within a participant and sometimes across participants. The WWC standards are primarily focused on treatment effects within participants.	*Meets WWC Standards Without Reservations*

Compositional change of the intervention and comparison groups

An additional aspect of group designs that could undermine the WWC's confidence in the study's findings and, consequently, could affect the study's research rating is compositional change.[2] This term refers to changes in the membership of intervention and comparison groups after participants have been assigned. Compositional change is a unique threat to RCTs and RDDs, in which, by definition, assignment to conditions is controlled. The WWC examines two sources of compositional change in RCTs and RDDs: attrition and sample joiners.

Attrition

Attrition refers to the lack of outcome data for participants who were assigned to either the intervention or the comparison group. The three basic sources of attrition are:

- Participants who have been assigned to a condition either do not participate or stop participating in that condition and are not available for outcome measurement. For example, a family might move out of the district after their child was enrolled in a study, so the child leaves the analytic sample.
- Participants are not available when outcome measurements were assessed. For example, a student might be absent the day that a posttest was given.
- Participants are removed from the study or the data analysis. For example, researchers may exclude participants from the data analysis because their scores on the outcome measure appear to be unusual.

High attrition could signal that the participants for whom outcomes are measured at the end of the study are too different from participants who were initially included in the study. This change in the composition of study groups could be an important source of bias.

The WWC developed an algorithm (described by Deke & Chiang, 2017) to determine whether the pattern of attrition suggests reason for concern. If the pattern in an RCT or RDD suggests that attrition is too high for the WWC to be confident in the study's findings, then the highest rating it can receive is *Meets WWC Standards With Reservations*. A detailed discussion about how the WWC reviews attrition in RCTs is included in Chapter III, Compositional change, and the description of attrition in RDDs in included in Chapter IV, Reviewing Findings From Regression Discontinuity Designs.

Sample joiners

Sample joiners are individuals who enroll in the intervention or comparison condition after researchers have assigned the students, classrooms, or schools to conditions and these assignments are known to anyone who could plausibly influence individuals' placement into clusters. This type of compositional change only occurs in cluster-level assignment studies, in which the whole classrooms, schools, districts, or other entities are assigned to the intervention or comparison group. Many of these studies have the potential for joiners.

[2] The WWC does not presently consider issues of compositional change for SCDs. Because measurements are repeated over time in SCDs, an absence of individuals from some observations is less impactful once the results are aggregated for analysis compared with typical group design studies.

Whether sample joiners present a risk of bias in the study's findings depends on whether the study authors include joiners in the analysis of outcomes.

- **Sample joiners not included in the analysis of outcomes.** If joiners are not included in the analysis, then joiners are unlikely to bias the study's results. This is because while joiners entered the classrooms or schools, they did not actually join the study.
- **Sample joiners included in the analysis of outcomes.** If study authors include joiners in the analyses, then the sample composition has changed, which could bias the study's results.

  Example. Consider an intervention that provides math tutoring to students in some classrooms at a school but not others. If math tutoring is highly valued by parents and students, some families will be motivated to transfer their children to the intervention classrooms. If parents succeed, then the composition of intervention and comparison classrooms may change in ways that are associated with the study's outcomes: Students with already advanced math skills or those who are motivated to study math could be disproportionately represented in the intervention classrooms. In this case, if joiners are included in the analysis of outcomes, then the WWC will have less confidence in the study's findings.

If data from joiners are not analyzed, then an RCT or an RDD with low sample attrition can receive the highest research rating of *Meets WWC Standards Without Reservations*. If joiners are included in the analytic sample, and the WWC concludes that they pose a low risk of bias, a study also can receive the rating of *Meets WWC Standards Without Reservations*. If joiners pose an elevated risk of bias, then the highest rating a study can receive is *Meets WWC Standards With Reservations*, provided that study authors satisfy other standards described in Chapter III, Reviewing Findings From Randomized Controlled Trials and Quasi-Experimental Designs, and Chapter IV, Reviewing Findings From Regression Discontinuity Designs.

Baseline equivalence

The comparability of intervention and comparison conditions before the start of the intervention—that is, at baseline—is a critical consideration for the WWC's confidence in the study's findings. RCTs, RDDs, and SCDs address comparability through the design of the research. RCTs, for instance, use random assignment to conditions to create groups that are expected to be similar at baseline on observable and unobservable characteristics. When a study is a QED, study authors do not have this level of control over assignment to conditions. For this reason, authors of QEDs must demonstrate baseline equivalence of the intervention and comparison conditions on relevant covariates, such as outcome pretest, and if needed address differences between conditions by applying an acceptable statistical adjustment. When RCTs and RDDs experience high attrition that makes the groups dissimilar when the outcomes are measured, the WWC also requires that study authors address baseline differences between the intervention and comparison conditions. Depending on the extent of concern over attrition, study authors of RCTs and RDDs may need to satisfy baseline equivalence by demonstrating that groups were similar at baseline or only apply an acceptable statistical adjustment to account for baseline group differences. Detailed information about the WWC's standards for baseline equivalence is summarized in Chapter III, Baseline equivalence standard.

When study authors demonstrate baseline equivalence or use acceptable adjustment strategies they reduce, but likely do not eliminate, the potential bias associated with the group assignment procedures. Therefore, the highest possible WWC research rating for QEDs, and for RCTs and RDDs with a high risk of bias due to compositional change, is *Meets WWC Standards With Reservations*.

WWC PROCEDURES AND STANDARDS FOR STUDY REVIEWS

Study reviews are the foundation of all WWC products. The WWC reviews multiple individual studies as part of evidence synthesis products such as intervention reports and practice guides. The WWC also reviews individual studies outside of evidence syntheses, for example, when studies have the potential to provide Tier 1, Tier 2, or Tier 3 evidence as defined by the U.S. Department of Education.

Each study review begins with screening the study for eligibility. Eligible studies are reviewed according to WWC standards. The final step in the study review process is synthesizing and reporting results. WWC procedures and standards for each phase of the study review are described in the chapters that follow. Presented next are several concepts that apply across chapters: how to use the *Handbook* and review protocols, how the WWC defines a study, and how to determine the study's unit of assignment.

How to use the *Handbook* and review protocols

This *Handbook* guides all study reviews conducted by the WWC. Reviewers will use the *Handbook* with the *Study Review Protocol* to review individual studies. Review teams conducting systematic reviews and syntheses of evidence will use the *Handbook*, the *Study Review Protocol*, and a topic area synthesis protocol as described below.

- **When to use the *Study Review Protocol***. Reviewers should use the *Handbook* supplemented by the *Study Review Protocol* to review all studies. The *Study Review Protocol* contains additional information needed to complete study reviews, in particular the definition of eligible outcome domains.
- **When to use a topic area synthesis protocol**. Review teams conducting systematic reviews should use a topic area synthesis protocol in addition to the *Handbook* and the *Study Review Protocol*. The studies included in the synthesis products should be reviewed using the *Handbook* and the *Study Review Protocol*. A topic area synthesis protocol will guide reviewers on what literature to search for and where to search for it (appendix A in the technical appendices outlines general guidelines) as well as how to prioritize studies for review by their expected relevance, quality, or usefulness. Prioritization may account for factors such as the release date of study manuscripts, the relevance and usefulness of the research for addressing the questions motivating the systematic review, the existence of additional studies of the same intervention in the WWC database of findings, and the expected number of experimental versus quasi-experimental designs available for review. For example, one topic area team may prioritize the review of randomized experiments, while another topic area team may prioritize release date. A topic area synthesis protocol also outlines criteria that are unique to the topic area, including, but not limited to, the outcome domains and populations relevant to the WWC publications that will be based on the systematic reviews of evidence, and whether supplemental findings, for example on particular subgroups, should be used in those syntheses of evidence.

The WWC's review protocols are available on the WWC website at https://ies.ed.gov/ncee/wwc/Protocols.

How the WWC defines a study

The core of the WWC evidence review process is the assessment of eligible studies against WWC standards. The definition of a study is important, given how the WWC reports on and summarizes evidence. The WWC defines a study as an examination of the effect of an intervention on a group of participants in which assignment to conditions was coordinated.

A manuscript may contain a single study or multiple studies in the same document, such as reports of multiple investigations with unique samples. Likewise, one study may be published across multiple manuscripts, such as by adding additional cohorts to the sample. It is not uncommon for the WWC to review two or more manuscripts that assess the same intervention and use the same sample. To establish whether findings from multiple manuscripts addressing the same intervention should be classified as the same study, the WWC establishes whether sample overlap is present. When a study is eligible for review, the review team should establish whether the study sample is independent or not from other studies based on sample overlap.

- **Independent samples** do not share any participants.
- **Dependent samples** share study participants; that is, the study sample has overlapping or identical samples.

Samples are defined in terms of their units of analysis (usually students or teachers) rather than broader, aggregated units, such as school districts or states.

It may be unclear whether two studies use dependent samples. To facilitate this determination, review teams should compare the following fields across the samples in question:

- Full reference.
- Each author and their Open Researcher and Contributor ID (ORCID), up to 7.
- Intervention name.
- Year(s) in which students received the intervention.
- Year(s) in which outcome data were collected.
- Each location in which a study was performed. The WWC defines locations as a district or state because exact school locations are rarely available in manuscripts.
- Grade level of participants in the sample.
- Number of participants in the sample.

If the contents of some of these fields for a manuscript overlap with or are the same as those of another manuscript, the review team may query the study authors to determine whether the studies have overlapping samples. If study authors do not respond to the query, then review team leadership will make a judgment based on the available information.

If the samples overlap, the WWC will treat the findings from the overlapping samples as multiple effects from the same study for the purposes of synthesis.

In the case of blocked or multi-site studies, the WWC will consider block- or site-specific effects reported in separate manuscripts different studies. If block- or site-specific effects are reported in one manuscript, the WWC will consider it a single study. Appendix F describes the WWC's procedure for aggregating findings from studies that examine the same intervention.

In the case of SCD manuscripts, each manuscript typically contains one or more designs (sometimes called experiments) intended to examine the effect of an intervention. While multiple designs or experiments may receive independent ratings, most SCD manuscripts are a single study because they are the examination of an intervention's effects on a single group of participants.

How to determine the unit of assignment

Before assessing the study according to WWC standards, the reviewer will need to determine the level of assignment to conditions: individual or cluster.

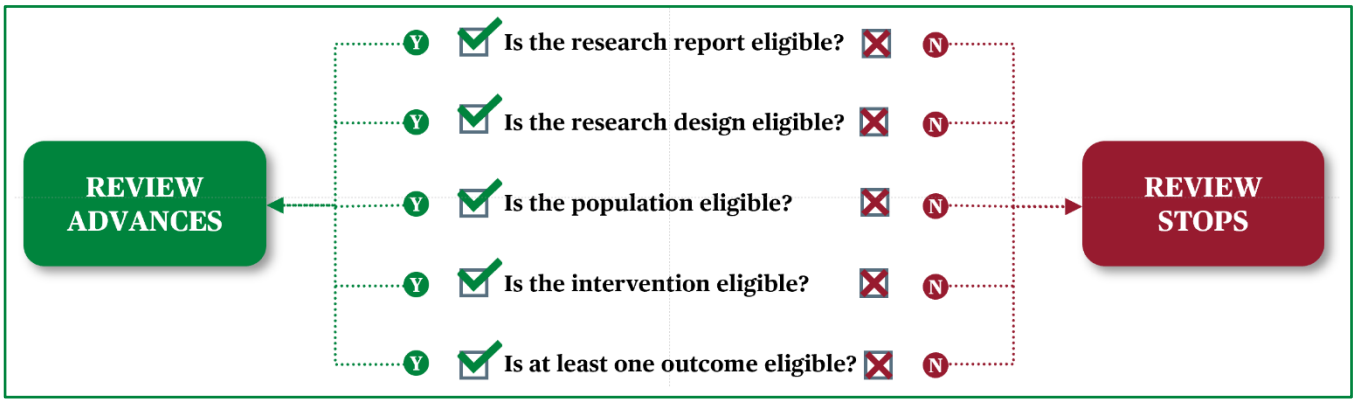
- In **individual-level assignment** studies, individuals such as students are assigned to study conditions.
- In **cluster-level assignment** studies, groups of individuals such as classrooms or schools are assigned to conditions as intact units.

More information about how to determine the unit of assignment is included in Chapter III, Assignment to conditions.

CHAPTER II. SCREENING STUDIES FOR ELIGIBILITY

Each study review begins by determining its eligibility for a WWC review. To be eligible, the study must satisfy all requirements shown in figure 3 and described next. If a study is determined to be ineligible for review, then the review stops, and the study does not receive a research rating.

Figure 3. *Study eligibility requirements for a What Works Clearinghouse review*



Study eligibility requirements

To be eligible for a WWC review, studies must meet all requirements described next.

Eligible research report

To be eligible for WWC review, studies must meet the availability, completeness, timeframe, and language requirements that follow.

- **Availability.** Studies and findings must be publicly available, that is, available on the internet with or without a subscription. However, when conducting systematic reviews, the WWC will only search for dissertations and theses in ProQuest Dissertations & Theses Global, EBSCO Open Dissertations, or ERIC, or another database specified in the topic area review protocol (see appendix A, Principles for Prioritizing and Searching for Studies to Review).
- **Completeness.** The manuscript must describe, in paragraph form, the intervention under study; the implementation of the study including design, data, and methods; and findings of the study. The manuscript must contain sufficient detail to warrant a WWC review, including descriptive statistics of the study sample and inferential statistics about the findings. Elements such as data tables, graphics, and references should be clearly and sufficiently labeled.
    - ***Usually eligible for WWC review*** are working papers and pre-published versions of articles, provided they are complete, free of track changes or edits, and are not watermarked as drafts.
    - ***Usually ineligible for WWC review*** are design documents, conference papers, and presentation slides from conference presentations because they typically lack sufficient information for WWC review.

- **Timeframe.** Studies must have been released or made public within the 20 years preceding the year of the review. For example, for reviews being conducted in 2022, the study must have been published in 2002 or later.[3]
- **Language.** Studies must be available in English, or have a full English translation, to be eligible for WWC review.

Eligible research designs

The study must use one of four research designs: randomized controlled trial (RCT), including RCTs that estimate a complier average causal effect (CACE); quasi-experimental design (QED); regression discontinuity design (RDD); or single-case design (SCD). Under the umbrella of QEDs are cross-sectional group designs, comparative interrupted time series, difference-in-difference designs, and growth curve designs.[4] Other QEDs are currently not eligible for WWC review for one of two reasons.

Minimum implementation duration

The WWC does not have a minimum implementation duration requirement, nor must the intervention be "branded" for the study to be eligible for review. For studies that meet WWC standards, the WWC documents study characteristics, including implementation duration, if reported by study authors.

1. **Designs that always contain a confounding factor.** Single-group pretest-posttest designs in which pretest observations are collected at one time point (for instance, fall semester) and posttest observations are collected at another time point (for instance, spring semester) are not eligible for WWC review because they always contain a time confound (Chapter III, Confounding factors, provides additional information about confounding factors).
2. **Review standards are currently not available.** Certain QEDs, such as interrupted time series designs without a contemporaneous comparison group, may produce credible causal estimates, but the WWC has not yet developed standards for reviewing studies using these designs.

Eligible populations

The study must examine the effectiveness of an intervention administered to the following types of participants:

- Students and other learners in early intervention programs for infants and toddlers; in preschool education programs for children ages 3 through 5; or in elementary, secondary, postsecondary, or adult education programs. "Student" and "learner" may be used interchangeably.
- Teachers, school leaders, other educators, or home or school-based service providers.

The majority of the study's analytic sample must include participants in the United States, in its territories or tribal entities, at U.S. military bases overseas, or in Organisation for Economic Cooperation and Development

[3] A topic area synthesis protocol may limit the scope of review to studies released fewer than 20 years from the time of the review. For example, in topic areas with an abundance of eligible research, or in which the context for implementation of interventions has changed substantially over the past 20 years, the topic area synthesis protocol may focus reviews on studies released in the past 15 years.

[4] A topic area synthesis protocol may, with justification, exclude QEDs from the scope of a topic area review for publications other than intervention reports because these designs are ineligible to receive the research rating of *Meets WWC Standards Without Reservations*.

member countries in which English is the primary or most used language—that is, Australia, Canada, Ireland, New Zealand, or the United Kingdom.

The WWC does not have a minimum sample size or sample composition requirements for studies to be eligible for review. However, the WWC will not prepare intervention reports unless it has reviewed findings including at least 20 individuals from studies meeting WWC standards.[5]

Eligible interventions

The study must examine an educationally relevant or school-based intervention. The WWC defines the term "intervention" broadly, and this term may comprise education practices, products, policies, and programs. Therefore, the following types of interventions may be included, which are not mutually exclusive:

- **Practices**. Education practices are discrete, clearly defined activities focused on improving student learning and related outcomes. Practices may be used with a broad and diverse range of participants to address a wide range of learning goals. Practices may be targeted to address a specific learning goal, skill, or population. An example of a practice is teaching new vocabulary.
- **Products**. Education products are "branded" or commercial interventions such as curricula or software. Products may be used as the primary instructional tool in the classroom or to supplement classroom material with differentiated instruction, remediation, or enrichment. Products may possess a trademark or copyright and generally are supported by a developer who provides technical assistance and sells or distributes the intervention.
- **Policies**. Education policies involve structural changes that are intended to improve student outcomes directly or indirectly. Examples of education policies include modifying the academic calendar and changing the number of credits required for graduation.
- **Programs**. Education programs are combinations of practices, products, or policies. For example, a charter school program may combine teacher practices with policies regarding school uniforms and total days of instruction.

Eligible outcomes

To be eligible for a WWC review, the study should include at least one outcome from a domain relevant to the education community. These domains include academic readiness, educational attainment, educational progress (for example, applying to or attending college), labor market outcomes, school attendance and progress (for example, promotion to next grade), school environment (for example, school climate and equity), social-emotional outcomes, behaviors and skills, school leader outcomes, and teacher outcomes. The full list of outcome domains and their descriptions is available in the *Study Review Protocol*.

[5] The minimum sample requirement of 20 individuals corresponds with the minimum sample size of cases for assigning effectiveness ratings to interventions included in SCD studies reviewed under version 3.0 and version 4.0 of the WWC *Handbook*.

After the study is found eligible

After the study is found eligible for a WWC review, the study advances through the review process and is assessed according to WWC standards. Depending on how the study is designed and how the design is executed, the review may include up to four steps: (a) reviewing outcome measures and checking for confounding factors, (b) evaluating the process of assignment to conditions, (c) evaluating compositional change, and (d) evaluating baseline equivalence. Each step has its own standards for evaluating study findings, and the standards vary based on the research design. Therefore, the following chapters describe WWC standards by design: Chapter III describes RCTs and QEDs, Chapter IV describes RDDs, Chapter V describes advanced group designs, and Chapter VI describes SCDs. The final chapter, Chapter VII, outlines the synthesis and reporting of results.

When reviewing findings from the study according to WWC standards, **the WWC evaluates each eligible outcome for each sample separately.** A reviewer reviewing a study with multiple eligible outcomes will repeat the steps described in the chapters that follow for each eligible outcome/sample combination. How the WWC presents findings across multiple outcomes is described in Chapter VII, Synthesis and Reporting of Results.

CHAPTER III. REVIEWING FINDINGS FROM RANDOMIZED CONTROLLED TRIALS AND QUASI-EXPERIMENTAL DESIGNS

Randomized controlled trials (RCTs) and quasi-experimental designs (QEDs) are the most common research designs reviewed by the WWC. Some of these studies assign individuals to intervention and comparison groups, whereas others assign clusters of individuals, such as entire schools of students. This chapter details how to review RCTs and QEDs that use individual-level and cluster-level assignment.

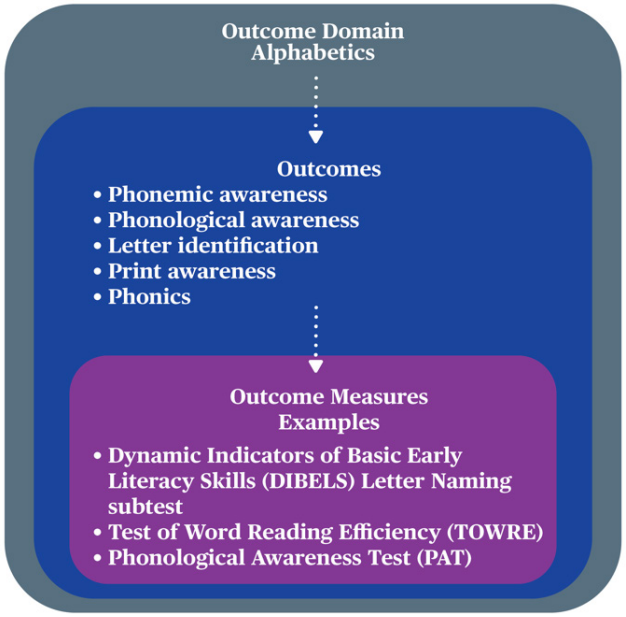
Step 1. Reviewing outcome measures and checking for confounding factors

The first step in assessing findings in an RCT or a QED according to WWC standards is to review the outcome measures used in the study and to examine the study for the presence of confounding factors. This step is usually completed prior to the in-depth review of the research design because this step will determine whether the review advances. If there is not a single finding that was measured using an outcome measure consistent with the WWC standards, or if the study contains a confounding factor that affects all findings, the study will receive a rating of *Does Not Meet WWC Standards* and the review will stop.

Outcome measure standards

The WWC defines an outcome as the knowledge, skills, attitudes, behaviors, or other measurable characteristic that researchers measure to learn about the impact of an intervention. An outcome domain is a group of closely related outcomes that provide information on the same underlying construct. An outcome measure is an instrument, device, or method that provides data on the desired outcome. Figure 4 shows an example of outcome domain for alphabetics, with possible outcomes and outcome measures.

Figure 4. *Example of outcome domain, outcomes, and measurements*



The WWC requires that at least one outcome found eligible for review when the study was screened for eligibility is assessed using an outcome measure that meets four standards: (1) face validity, (2) reliability, (3) not overaligned with the intervention, and (4) consistent data collection procedures. If the study does not include at least one eligible outcome measured in a way aligned with the WWC standards 1 through 4, it will receive a rating of *Does Not Meet WWC Standards* and the review will stop.

Standard 1: Face validity

To show evidence of face validity, an outcome measure must appear to measure what it claims to measure. To demonstrate face validity, a measure must have a clear definition of what it measures, such as a skill, an event, a condition, or an object, and assess what it claims to measure. For instance, a measure described as a test of

reading comprehension that only assesses reading fluency does not demonstrate face validity. Dichotomized measures must preserve the natural ordering of the latent variable to demonstrate face validity.

Standard 2: Reliability

The reliability of a measure captures whether it would yield similar scores from different administrations. The reliability standard aims to set principles for maximum allowable measurement error. Higher reliability indicates lower measurement error. Internal consistency and test-retest reliability can capture measurement error that results from poor question wording, for example. Inter-rater reliability can capture measurement error that results from coder judgment. Although this measurement error does not create bias in the impact estimate, the error reduces precision and therefore the likelihood of detecting an impact if one exists. Some measures WWC considers valid and reliable without study authors reporting reliability statistics as summarized in box 1.

Minimum WWC reliability standards for an outcome measure

1. Internal consistency (for example, Cronbach's alpha) of at least .60.
2. Temporal stability and test-retest reliability of at least .40.
3. Inter-rater reliability (correlation) of at least .50.
4. Inter-rater agreement (percent agreement and kappa) of at least .80 (for percent agreement) and at least .60 (for kappa), based on at least 20 percent of the judgments.

The ***Study Review Protocol*** may specify higher reliability standards in specific domains.

Box 1. Outcome measures WWC considers valid and reliable without reliability statistics

Administrative records:

- Administrative records, such as **standardized tests** routinely administered by states or districts, or records collected by school districts or institutions of higher education—such as **attendance, school enrollment, graduation, degree attainment, course credits, or disciplinary incidents**—do not need to demonstrate face validity and reliability.
- **Grade point average** is presumed to be reliable, assuming consistent collection across the sample included in the study. When different school districts or different institutions of higher education are included in the study, the review team will need to confirm that the same scale of grade point average has been used across sites for that measure to be considered reliable.

Outcome measure that can be scored with very low error by a single coder

Review team leadership may make an exception for an outcome measure that can be scored with very low error by a single coder. In this case, the measure can meet WWC review standards even if the reliability of the measure is not demonstrated. Examples of measures that might be scored with very low error include words read correctly per minute, the ability to recite the alphabet, and the ability to count to a predetermined number.

Standard 3: Not overaligned

A third standard of outcome measures is that they not be overaligned with the intervention. Overalignment occurs when the outcome measure provides an unfair advantage to one group or condition over another. When outcome measures are closely aligned with or tailored to the intervention condition, the study findings may not be an accurate indication of the effect of the intervention.

Example. An outcome measure based on an assessment that relied on reading materials or vocabulary words used in the intervention condition but not in the comparison condition likely would provide an advantage to students in the intervention condition. The WWC is likely to consider such a measure overaligned.

The overalignment rule does not apply when material covered by an outcome measure must be taught. For example, reciting the alphabet requires being taught the alphabet, but improving reading comprehension does not require focusing on a specific set of reading passages. Put another way, an outcome measure is overaligned when the content or materials provided to participants in a single condition might affect scores through tailoring of outcome measure for participants in that condition, developing familiarity with the format, or other means besides learning educationally relevant material.

The decision about whether a measure is overaligned is made by the review team leadership. Content experts can provide guidance on whether the content assessed in an outcome measure is broadly educationally relevant, and thus not overaligned.

Standard 4: Consistent data collection procedures

A fourth standard of outcome measures is that data were collected in the same manner for the intervention and comparison conditions. The WWC assumes data were collected in the same manner if no information to the contrary is provided in the study. However, a reviewer should look for comments in the study indicating that different modes, timing, or personnel were used to collect data for the intervention and comparison conditions for each time period included in the analysis or indicating that measures were constructed differently for the conditions. When outcome data are collected differently for the intervention and comparison conditions, study-reported impact estimates cannot be isolated from differences in the data collection methods.

Example 1. Measuring dropout rates based on program records for the intervention group and school administrative records for the comparison group will result in impact estimates that are difficult to interpret, because dropouts may be defined differently in different data sources The impact estimates due to the intervention will be impossible to isolate from any effect of the different methods for assessing the outcome.

Example 2. If intervention and comparison students are in different districts, grade point average might be calculated differently or be based on different courses in the two groups. For instance, districts may differ in how Advanced Placement courses are handled and on the weighting of specific courses. If so, the impact estimates due to the intervention will be impossible to isolate from any effect of different procedures for the collection of grade point average.

Additional consideration: Independence of outcome measure

The WWC also will consider, depending on the outcome domain specified in the *Study Review Protocol* and the purpose of the review, whether the measure is independent of the intervention.

A measure will be considered nonindependent if either it was developed by study authors and is not in broader use, or if it was developed by the intervention's developers.

Measures created by study authors are nonindependent unless there is evidence of their broader use. Measures created by developers to accompany an intervention are always nonindependent when used to estimate the impact of the same developer's interventions. Measures in administrative records, such as those from a local or state education agency, are considered independent because they are typically not developed for a specific intervention. For certain outcome domains identified in the *Study Review Protocol*, findings based on nonindependent measures cannot be designated as main findings for individual study reviews.

In making the independence determination, study review teams should refer to the definition of nonindependence and should consult the list of known independent measures that is compiled by the WWC and appended to the *Study Review Protocol*. Study review teams should be aware that a measure's independence designation may change over time, moving from nonindependent to independent with wider use by different study authors. Changes in the independence designation will be reflected in the list of independent measures. Also, the list of known independent measures should not be considered exhaustive; a measure might meet the definition of nonindependence and not be on the list.

Confounding factors

A reviewer conducting a WWC review should determine whether a study contains a confounding factor. A confounding factor is an aspect of a study that is always present for members of one group and never present for members of the other group. A confounding factor has the following characteristics:

- It is observed.
- It aligns completely with only one study condition—that is, present for all participants in one condition and absent for all in the other condition.
- It is not part of the intervention the study is testing.

Confounding factors make it impossible to isolate the effect of the intervention from the effect of a confounding factor. A finding that contains a confounding factor receives a rating of *Does Not Meet WWC Standards* and the review stops for that finding. If the study has other findings that are not confounded, the review will continue for those findings. However, when the study has a confounding factor, it usually affects all findings.

Intervention or comparison group contains a single study unit, such as teacher, classroom, school, or district (also known as an *N* = 1 confounding factor)

N = 1 is the most common type of confounding factor among studies reviewed by the WWC that have a confound (table 5). It occurs when either the intervention or comparison condition contains a single study unit—such as a teacher, classroom, school, or district—and that unit is not present in the other condition.

Table 5. *Examples of an* N = *1 confounding factor*

Examples of an *N* = 1 confounding factor	Similar circumstances without an *N* = 1 confounding factor
A study includes two schools, one in the intervention condition and one in the comparison condition. A single school in each condition is a confounding factor.	A study includes two schools. All students in one school are in the intervention condition. Students in the other school are in the intervention and comparison conditions.

Continued on next page

Table 5. *Examples of an* N = *1 confounding factor (continued)*

Examples of an *N* = 1 confounding factor	Similar circumstances without an *N* = 1 confounding factor
A study has two intervention classrooms and two comparison classrooms. Both intervention classrooms had the same teacher, who had no interaction with the comparison classrooms. A single teacher is a confounding factor.	A study has two intervention classrooms and two comparison classrooms. The same teacher taught all classrooms.

A single unit is not always a confounding factor. A critical distinction between what does and does not constitute an *N* = 1 confounding factor is the alignment of study unit and condition: The unit must be aligned with only one condition to be a confounding factor. If one school is included in the study but students from that school are assigned to the intervention and comparison conditions, the single school is not a confounding factor because the school is not aligned solely with either condition. Likewise, if one instructor interacts with intervention and comparison students, then the single instructor is not a confounding factor because the instructor is not aligned solely with either condition.

The WWC also does not consider a single unit a confounding factor if that unit is the focus of intervention. For example, when the intervention of interest is attending a school with unique organization and governance, the WWC will not consider a single such school to be a confounding factor (it must be compared to at least two other schools to avoid an *N* = 1 confound in the comparison condition). The single unique school is not a confounding factor because the school and the intervention are the same. That is, attending the school is the same as receiving the intervention.

Is volunteering a confounding factor?

Volunteering occurs when individuals choose to participate in an activity or provide a service usually free of charge. For example, teachers may volunteer to implement an intervention. The WWC does not consider volunteering a confounding factor in RCTs if volunteers are first recruited and then randomly assigned to conditions.

Study design in QEDs allows for volunteering, which–ideally–is mitigated through statistical adjustment. Volunteering is one of WWC's several concerns with QEDs, which is why the highest possible rating for QEDs is *Meets WWC Standards With Reservations*.

Characteristics that can affect outcomes differ between conditions with no overlap

Confounding also can occur when characteristics of the study units–for example, participating students or teachers delivering the intervention–are aligned with only one condition (table 6). In such cases, differences between groups could be due to the intervention, characteristics of study units, or a combination of the intervention and characteristics.

> *Example*. If students' preintervention level of achievement is aligned with only one study condition and achievement can affect the outcome of the intervention, then student characteristics will be a confounding factor.

If a certain characteristic is a requirement for the intervention, the WWC will not consider it a confounding factor. For instance, if teachers delivering the intervention must have a master's degree to deliver the

intervention, then master's degree will not be a confounding factor even if this characteristic is aligned with only the intervention condition.

When reviewing studies that use regression discontinuity designs (RDD), the WWC does not consider the forcing variable, which is a variable used to assign participants to conditions, a confounding factor. This is because the WWC standards for RDDs are designed to ensure that the impact estimates from these studies account for the forcing variable as described in Chapter IV, Reviewing Findings From Regression Discontinuity Designs.

Time is a confounding factor if it is completely aligned with conditions, with no overlap between conditions

Several research designs have time as a confounding factor (see table 6).

> *Example*. A design in which groups are defined by cohort, often labeled a successive-cohort design or cohort design, has a time confound. Consider an intervention group consisting of a cohort of grade 3 students in 2010/11 and a comparison group consisting of the previous cohort of grade 3 students in 2009/10. Usually, both cohorts are observed in one school or the same set of schools. In this cohort design, the intervention and comparison conditions are completely aligned with different time periods, and the estimated impact is confounded with any changes that may have occurred between those time periods. These changes—such as new district policies, new personnel, or new state tests—could plausibly affect outcomes.

When overlap in the participant characteristic or time between the conditions is imperfect—even by a small degree—the WWC does not consider the characteristic a confound.

Table 6. *Examples of characteristics or time as a confounding factor*

Examples of characteristics or time as a confounding factor	Similar circumstances where characteristics or time are not a confounding factor
All teachers in the intervention group—and none in the comparison group—have a PhD. Advanced qualifications for all teachers in the intervention group is a confounding factor.	All teachers in the intervention group—and none in the comparison group—have a PhD. A PhD is a requirement for delivering the intervention and therefore is not a confounding factor.
The comparison group is grade 4 students in 2010/11 and the intervention group is grade 4 students in 2011/12. Time is a confounding factor.	The comparison group is grade 4 students in 2010/11 and 2011/12. The intervention group is grade 4 students in 2011/12. Time is not a confounding factor because grade 4 students in the comparison group overlap with the grade 4 students in the intervention group.

Step 2. Assignment to conditions

Studies that have no confounding factors and have at least one outcome aligned with the WWC's standards will advance through the review. Next, a reviewer will examine the process used to determine assignment to the intervention and comparison conditions. The reviewer will examine how the assignment was carried out and whether any factors interfered with it. Assignment to conditions is a critical component of the research design; it affects the WWC's confidence in claims that the intervention caused the observed effect and, subsequently, the study's rating, as described below.

RCTs—Group assignment via a random process

Group assignment refers to the process of placing study units, such as students, classrooms, or schools, into intervention and comparison conditions. Group assignment via a random process, or random assignment, is a method of assignment carried out using chance procedures. The WWC considers random assignment well executed if it meets two criteria:

1. Units are assigned to conditions entirely by chance.
2. Units have a nonzero probability of being assigned to each condition.

Methods of accounting for assignment probabilities

The WWC accepts three methods of accounting for assignment probabilities. Studies can:

1. Use inverse probability weights,
2. Include an indicator (or dummy) variable in the analysis for each subsample with a different probability, or
3. Combine impacts estimated separately for each subsample.

Well-executed randomization creates groups that are similar on observable and unobservable characteristics on average. RCTs can claim that differences in outcomes are due to the intervention, not pre-existing differences between groups. RCTs with a well-executed randomization have the potential of demonstrating a causal effect of an intervention and to receive the highest research rating of *Meets WWC Standards Without Reservations*.

Researchers may use several possible methods to conduct random assignment, which may include using blocked random assignment, matched pair random assignment, random subsampling, predetermined probabilities to assign units to conditions, and rerandomization methods (such as those developed by Morgan & Rubin, 2012).

Probability of assignment to condition

Every unit in a well-executed RCT must have a nonzero probability of being assigned to each study condition. However, the probability of being assigned to a particular condition can differ across study units. If units are assigned to a condition with different probabilities—that is, if the chance of being assigned to a group differs for subjects within the same assigned condition—then the analysis must adjust for the different assignment probabilities. This requirement also applies if the probability of assignment to a group varies across blocks in a stratified random assignment.

Assignment that is not random

Assignment based on factors such as last name, birthday, or class schedule is not random; these factors do not rely solely on chance.

Purposeful group assignment made to accommodate specific factors, such as students' needs, does not meet a nonzero probability criterion.

Using rerandomization

If study authors use a rerandomization method for group assignment, they must additionally satisfy the following two requirements:

1. **The criteria used for defining acceptable rerandomizations must be specified prior to assigning units to conditions**. This requirement would be satisfied if, for example, the research team identified an empirical threshold (or thresholds) on a measure of group differences for acceptable randomization prior to assigning units to conditions.

2. **Study authors must provide a peer-reviewed and published methodological source for the rerandomization method used.** This requirement would be satisfied, for example, if study authors cited a peer-reviewed journal article that was the basis of the rerandomization method used.

Additionally, the review team methodologist may need to review the rerandomization procedures that were used to ensure that the study authors did not inadvertently conduct a deterministic assignment process, thereby creating an $N = 1$ confound. Scenarios like these may arise when using rerandomization with especially small samples and or especially stringent rerandomization criteria.

How to determine the unit of random assignment

A reviewer will need to determine the level of assignment to conditions: individual or cluster.

- In **individual-level RCTs**, individuals such as students are randomly assigned to study conditions.
- In **cluster RCTs**, groups of individuals such as classrooms or schools are randomly assigned to conditions as intact units.

Distinguishing types of study units in RCTs

Unit of assignment: Study unit at which random assignment occurred.

Unit of measurement: Study unit at which outcomes were measured. For a study that analyzes school-level means of student test scores, the unit of measurement is students.

Unit of analysis: Study unit at which study authors analyzed the data. For a study that analyzes school-level means of student test scores, the unit of analysis is schools.

A defining characteristic of cluster RCTs is that the unit of assignment is larger than the unit of measurement.

> *Example*. Consider an RCT that randomly assigns teachers to conditions and measures student and teacher outcomes. Student outcomes, such as student achievement, will be considered as arising from cluster-level assignment because the unit of assignment (teachers) is larger than the unit of measurement (students). Student behavior based on classroom observations also will be considered as arising from cluster-level assignment (with students as the unit of measurement), even if the study authors recorded their observations at the classroom level. In contrast, the analysis of teacher outcomes, such as classroom practices, will be considered as arising from individual-level assignment because the unit of assignment (teachers) matches the unit of measurement (teachers).

Determining whether to classify an outcome as having arisen from cluster-level or individual-level assignment does not depend on the method of impact estimation or level of data aggregation (unit of analysis).

> *Example*. An RCT might assign schools to conditions and then carry out the statistical analysis as if the school means were the data. Even though the analysis was based on school-level means, the unit of measurement would still be students, making the study a cluster-level assignment study.

A study may appear to have more than one level of cluster membership.

> *Example*. An RCT may assign grade 1 teachers to conditions. The schools, however, may have only one grade 1 teacher for the entire school. Hence, assigning teachers to conditions would be the same as assigning schools to conditions. In this case, the WWC would define the largest study unit (schools, not teachers) as the unit of assignment.

While uncommon, studies may assign both clusters and individuals to conditions.

> *Example*. An RCT might assign classrooms to conditions and randomize students to classrooms, in either order. In this case, the WWC defines the study as individual-level assignment (students assigned to conditions).

How to determine the integrity of random assignment

A reviewer conducting a WWC review of an individual-level or a cluster RCT needs to determine whether the integrity of random assignment has been maintained. When an RCT does not meet the two criteria for a well-executed randomization—units are assigned entirely by chance and each unit has a nonzero probability of being assigned to each condition—random assignment is considered compromised. A compromised RCT will be reviewed using the standards for QEDs. The highest possible research rating for a compromised RCT is *Meets WWC Standards With Reservations*.

An individual-level or cluster RCT can be compromised in four ways, as described next. Additional examples of compromised individual-level and cluster RCTs are in table 7.

1. **Analyses include units not randomly assigned**. An RCT is compromised when individuals or clusters in the analytic sample used to estimate findings were not subject to random assignment.

   > *Example*. Consider a situation in which a student with special needs always receives the intervention. The study examining the impact of the intervention would be a compromised RCT if it included this student in the analytic sample. Study authors would need to exclude the student from analyses to preserve the integrity of the random assignment.

   Cluster RCTs also can be compromised if the analyses include clusters that were not randomly assigned. An example is replacing intervention schools that left the study with schools that were not randomly assigned. However, an example that does not compromise a cluster RCT is including individuals who were not present in the clusters at the time of random assignment. These individuals are called joiners. Including joiners in analyses can sometimes reduce the maximum possible WWC research rating. Chapter III, Compositional change in cluster RCTs addresses considerations regarding joiners.

2. **Analyses do not account for differing assignment probabilities**. An RCT is compromised if units are randomly assigned to the study conditions with different probabilities, but the findings are based on an analysis that does not account for the different assignment probabilities.

   > *Example*. Consider a study that conducts random assignment separately within two districts of students. The study includes the same number of students in both districts, but students in district A are high performing at baseline, while students in district B are low performing at baseline. The study assigns 70 percent of district A students to the intervention condition but assigns only 30 percent of district B students to the intervention condition. In this case, the intervention group includes 70 percent high-performing students, while the comparison group includes 70 percent low-performing students. Study authors should take this imbalance into account by controlling for district membership so that any positive impacts are due to the intervention and not due to the groups' dissimilarity from the start.

3. **A randomized unit's assigned condition is not the same as the unit's analyzed condition**. As a general rule, the WWC prefers estimates of program effects from an intent-to-treat analysis. An intent-to-treat analysis uses the participant's assignment to a condition as the independent variable–regardless of whether the participants remained in the condition.[6]

   An RCT is compromised when the investigator changes a participant or cluster's group membership after they have been randomly assigned to a condition. This can occur within both the intervention and comparison conditions.

   Example. Some participants may be assigned to the intervention but participate as comparison group members. Should the study authors change their assigned status from intervention to comparison, the RCT is considered compromised. Likewise, if some participants assigned to the comparison condition receive the intervention, and then the study authors change the condition from comparison to intervention, the RCT is considered compromised.

4. **Analyses exclude units randomly assigned based on reasons with a clear link to group status**. An RCT is compromised when study authors manipulate the analytic sample in a way that systematically excludes certain individuals or clusters based on events that occurred after the introduction of the intervention and where there is a clear link between study condition and the reason for the exclusion. A clear link is present when the exclusion is based on a measure that may have been affected by the intervention status. As a result, those excluded from the intervention condition may differ systematically from those excluded from the comparison condition and the remaining participants in the study conditions may no longer be comparable.

   Example. Consider an RCT focused on an intervention designed to improve student attendance. If study authors exclude from the analysis students with high levels of absenteeism, the RCT is compromised. The sample exclusion is based on a measure (level of absenteeism) that may be affected by the intervention status, indicating a clear link between study condition and the reason for sample exclusion.

In randomized block designs, including randomized pair designs, excluding blocks or pairs from the analysis does not compromise the RCT unless there is clear evidence that the exclusions were based on postintervention criteria that may have been affected by intervention status. In randomized pair designs, specifically, if pairs are excluded from analyses because one or both members of the pair are missing outcomes, the exclusions count as attrition and do not compromise the RCT. The RCT is only considered compromised if the pair exclusion is clearly based on additional postintervention criteria that may have been affected by intervention status, such as excluding pairs based on completion of the intervention. The four concerns listed above are summarized in table 7. Sample exclusions that do not compromise an RCT are described in Chapter III, Attrition in individual-level RCTs.

A well-executed random assignment procedure should result in groups that are similar on observable and unobservable characteristics. Sometimes valid randomization procedures produce intervention and comparison

[6] Studies that address noncompliance by reporting complier average causal effects (CACE) or fuzzy RDD standards may be eligible for review using the standards described in the sections on advanced group designs and RDDs.

groups that appear dissimilar based on chance. The WWC does not consider these chance differences to compromise the RCT, and such studies are reviewed using the usual review process for RCTs.

Table 7. *Examples of compromised individual-level and cluster-level randomized controlled trials*

Types of compromised RCT	Examples of compromised individual-level RCT	Example of compromised cluster RCT
1. Analyses include units not randomly assigned	Study assigns some students with special needs to always receive the intervention and analyzes them along with randomly assigned students.	Study replaces an intervention school that left the study with another school that was not randomly assigned.[a]
2. Analyses do not account for differing assignment probabilities	Students with low socioeconomic status have a higher probability of being in the intervention group, but the study does not control for these probabilities in the analyses.	Study assigns schools within geographic regions, yielding different assignment probabilities across regions, and the study does not control for geographic region in analyses.
3. Change a randomized unit's assigned condition in analyses	Study includes a student assigned to the intervention condition in the comparison condition after the student switches classrooms following random assignment.	Study analyzes a randomly assigned intervention school as if it were a comparison school (perhaps because the school did not implement the intervention).
4. Exclude based on reasons with a clear link to group status	Study excludes students from the intervention condition who did not participate in the intervention at intended thresholds.	Study excludes intervention schools (or students) with low implementation fidelity.

a. However, including individuals who were not present in clusters at the time of random assignment—that is, joiners—does not compromise cluster RCTs. See "Analyses include units not randomly assigned" above and Chapter III, Compositional change for additional discussion.

QEDs—Assignment via an uncontrolled process

In QEDs, groups are formed using an uncontrolled assignment process. The groups must be distinct (nonoverlapping). A QED can involve group formation before or after collecting outcomes.

> *Example*. Study authors are interested in evaluating the impact of participation in student athletics on class attendance. All students who currently participate in athletics are designated as the intervention group, and the students who are not currently enrolled in athletics are designated as the comparison group.

When reviewing the group assignment process for a QED, a reviewer needs to assess the level at which assignment occurred: individual or cluster. For QEDs, the WWC generally defines the unit of assignment as the largest study unit that contains members of only one condition. Consider if study authors compared student outcomes in schools implementing a dropout prevention program versus comparison schools. The unit of assignment is schools because each school has only intervention students or only comparison students. This study is a cluster-level assignment study. In contrast, if some schools have both intervention and comparison

students, then the unit of assignment is students (assuming there is no other intermediate study unit such as classrooms that could be the unit of assignment). This study is an individual-level assignment study.

In a QED, groups could differ on observable and unobservable characteristics.

- **Observable characteristics** are those that researchers can measure, for example test scores.
- **Unobservable characteristics** are those that researchers either could not or did not measure (for example, motivation).

Examples of group formation in QEDs

- **Convenience samples**: nonparticipants who are nearby and available.
- **Groups formed for another purpose**, such as using a separate district or district average as the comparison group, as long as the groups do not overlap.
- **Nonparticipants matched** to participants on baseline data using statistical techniques.

Even with equivalence on observable characteristics, there may be differences between groups in unobservable characteristics that could introduce bias into an estimate of the effect of the intervention. Bias is a systematic difference between the true impact of the intervention and the estimated impact that can lead to incorrect conclusions about the effect of the intervention. For this reason, the highest research rating that QEDs can receive is *Meets WWC Standards With Reservations*.

Step 3. Compositional change

In RCTs, changes to the composition of the intervention and comparison groups after participants have been assigned can introduce bias because the research design may no longer account for differences in observable and unobservable group characteristics. Therefore, the WWC considers compositional change a key issue. The WWC assesses issues of compositional change for RDDs as well, as detailed in Chapter IV, Sample attrition and baseline equivalence.

Compositional change can occur in QEDs, but the WWC already assumes that QEDs cannot fully account for differences in observable and unobservable group characteristics. Therefore, compositional change is not assessed for QEDs. A reviewer reviewing a QED should skip the step on compositional change and instead examine whether study authors satisfied the baseline equivalence standard described in Chapter III, Baseline equivalence standard.

In individual-level RCTs, compositional change occurs through sample attrition. Attrition refers to instances in which units (such as students, teachers, or principals) that were assigned to intervention or comparison conditions leave the study or are otherwise unavailable for outcome measurement. For example, students who went through randomization may leave study schools or may not provide information for follow-up surveys conducted for the study.

In cluster RCTs, the WWC considers three types of compositional change: cluster-level attrition, individual-level attrition, and joiners.

1. Cluster-level attrition refers to instances where entire cluster-level assignment units (such as classrooms or schools) contribute no outcome data to the analytic sample. For example, this type of compositional change could happen if the school decided not to administer an outcome measure to students, stopped participating in the data collection, or if the school closed.
2. Individual-level attrition refers to (a) individuals who leave the clusters at some point after random assignment (leavers) and (b) individuals who are present in the clusters at follow-up but have missing outcome data.
3. Joiners are individuals who enter clusters after the results of random assignment are known to anyone who could plausibly influence individuals' placement into clusters. An example of this type of compositional change is the guidance counselor, who knows that a new student is behind in math, placing the student in an intervention classroom that is testing a new math program.

The WWC uses these considerations regarding compositional change to determine the highest possible research rating for RCTs:

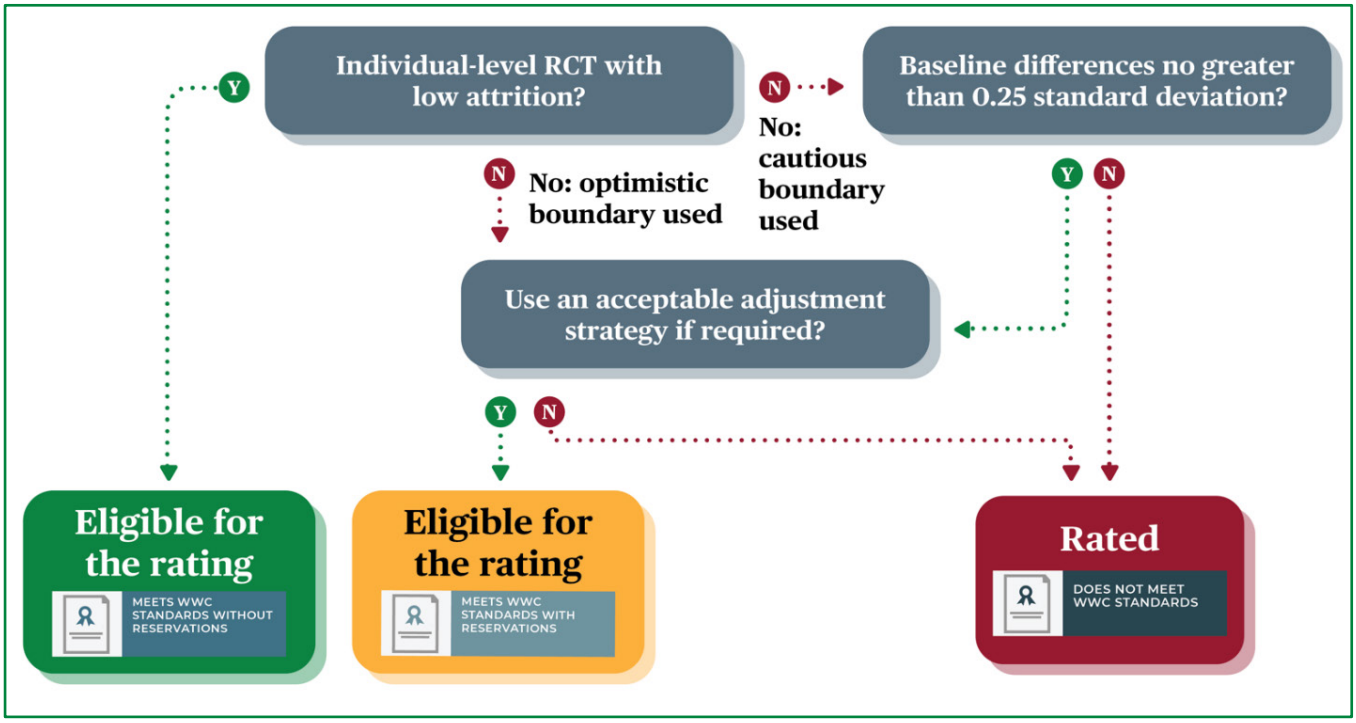
- **RCTs eligible to be rated *Meets WWC Standards Without Reservations*:**
    - Individual-level RCTs with low attrition are eligible to be rated *Meets WWC Standards Without Reservations*, reflecting a low risk of bias due to compositional change.
    - Cluster RCTs are eligible for this rating if they have low cluster-level attrition, have low individual-level attrition, and exclude any high-risk joiners from analyses, also reflecting a low risk of bias due to compositional change.
- **RCTs eligible to be rated *Meets WWC Standards With Reservations*:**
    - The highest possible research rating for individual-level RCTs with high attrition is *Meets WWC Standards With Reservations*, reflecting a high risk of bias due to compositional change.
    - The highest possible research rating is also *Meets WWC Standards With Reservations* for cluster RCTs that have high cluster-level attrition, high individual-level attrition, or include high-risk joiners in analyses. To receive this research rating, these RCTs generally must satisfy the baseline equivalence standard described later in step 4. However, one exception is that cluster RCTs with low cluster-level attrition that demonstrate follow-up data are representative of the individuals in the clusters do not need to satisfy the baseline equivalence standard, as detailed in step 3d.

How the WWC assesses compositional change varies by the level of assignment. A reviewer reviewing an RCT with individual-level assignment to conditions should continue to the next section. A reviewer reviewing an RCT with cluster-level assignment to conditions should skip to Compositional change in cluster RCTs.

Attrition in individual-level RCTs

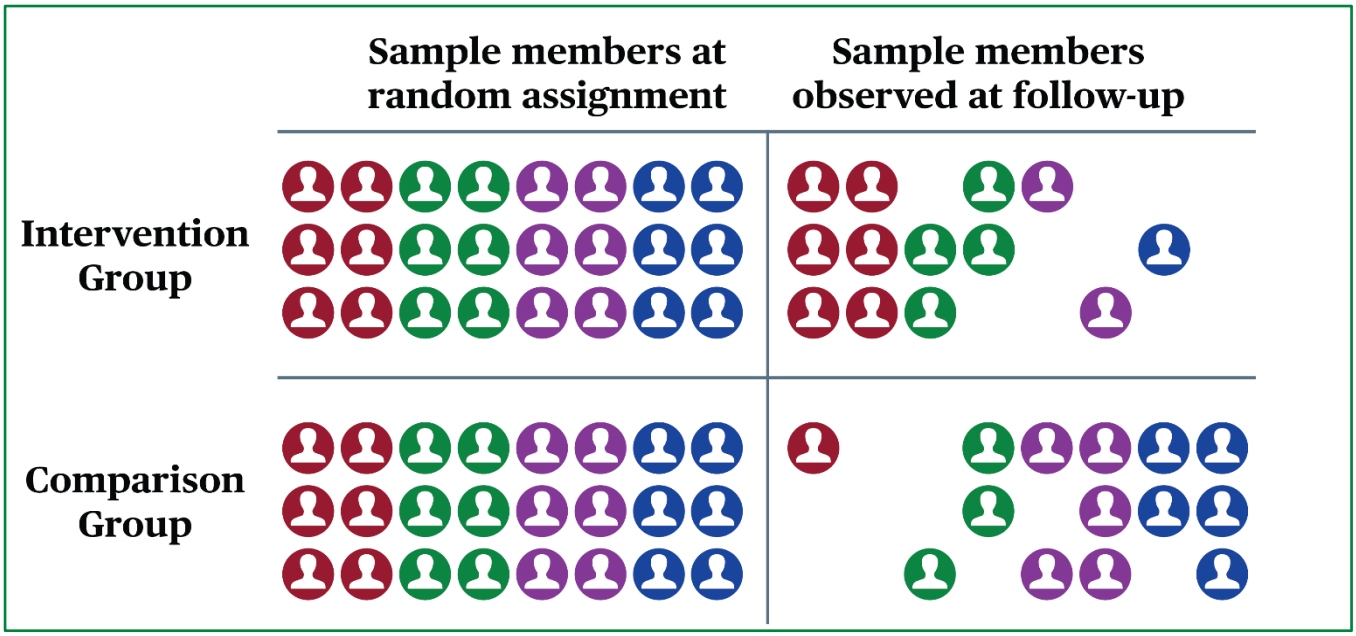
Figure 5 maps the critical components of how attrition in individual-level RCTs affects the research rating. RCTs with low attrition are eligible to receive the research rating *Meets WWC Standards Without Reservations*, whereas RCTs with high attrition are eligible to receive the research rating *Meets WWC Standards With Reservations* if the study satisfies the baseline equivalence standard. For high-attrition RCTs, the choice of attrition boundaries detailed in this following section affects the available options for satisfying the baseline equivalence standard.

Figure 5. *Ratings flowchart for individual-level assignment studies*



Attrition in individual-level RCTs occurs when sample members who were randomly assigned have missing outcome data or were otherwise not included in analyses (see figure 6). The main concern with attrition is that the intervention and comparison conditions are no longer similar on observable and unobservable characteristics. (Recall that when an RCT includes participants who were not randomized in the analysis, it becomes compromised and is reviewed as a QED.)

Figure 6. *Example of differential attrition rates resulting in dissimilar groups*



The WWC's attrition standard is based on a theoretical model for attrition bias and on empirically based assumptions.[7] The model depicts expected bias as a function of the rates of overall and differential attrition.

- **Overall attrition** is the percentage of randomly assigned units for which the study authors do not observe outcome data—that is, the level of attrition for the whole sample.
- **Differential attrition** is the absolute value of the percentage point difference between attrition rates for the intervention group and the comparison group—that is, the level of attrition for the group with the highest rate of attrition minus the level of attrition for the group with the lowest rate of attrition.

The WWC's attrition algorithm attempts to assess the amount of bias that could result from attrition.

- **High attrition** is defined as potential bias ≥ 0.05 standard deviation.
- **Low attrition** is defined as potential bias < 0.05 standard deviation.

When reviewing an RCT, a reviewer needs to determine whether a study has a high or low level of attrition.

Determining optimistic or cautious attrition boundary when assessing attrition

The WWC has estimated the levels of expected bias associated with different combinations of overall (for the entire sample) and differential (for different conditions) attrition rates and identified the combinations that

[7] To determine reasonable values to use in assessing the extent of potential attrition bias in a study, the WWC made assumptions about the relationship between attrition and outcomes that are consistent with findings from several randomized trials in education. More information on the model and the development of the attrition standard can be found in the WWC Technical Paper (Deke & Chiang, 2017).

generate low and high levels of attrition. Attrition boundaries are described in appendix C in the technical appendices along with an example of how the WWC uses these boundaries to compute attrition.

Teams conducting WWC reviews are responsible for determining whether to use optimistic or cautious assumptions when assessing sample attrition, and should document their reasoning. In general, choice of the optimistic boundary indicates the review team's assessment that attrition (or lack thereof) in the sample is unlikely to be related to the intervention. If review teams find that they cannot defensibly choose between optimistic and cautious assumptions, they should articulate this and should use the cautious assumptions. Examples where the optimistic attrition boundary could be used include interventions unlikely to influence retention in schools, such as a supplemental K–12 curriculum or other targeted intervention delivered during regular school hours. Examples where the cautious attrition boundary typically should be used include dropout prevention programs, school choice programs, programs delivered outside of regular school hours, elective or selective courses such as Advanced Placement courses, and postsecondary interventions that could affect student course or college enrollment decisions.

Imputed outcomes count as attrition

For example, if a study analyzes data from 100 units, including 90 with measured outcome data and 10 with imputed outcome data, then the overall attrition rate is 10 percent. See Chapter V, Procedures and standards for analyses with imputations for missing data for more information on how the WWC reviews studies with missing or imputed baseline or outcome data.

Sample exclusions that do not count as attrition

In some instances, sample exclusions in RCTs after random assignment do not constitute attrition. These exclusions include:

- **Exclusions due to acts of nature if both conditions are affected.** Losing sample members after random assignment because of acts of nature, such as hurricanes, earthquakes, or pandemics, is not considered attrition when the loss is likely to affect intervention and comparison group members in the same manner. However, when the loss due to an act of nature was concentrated in one group, the loss will be considered attrition.
- **Exclusions based on random selection**. Excluding a subset of the initial sample from the analysis is not considered attrition if the subsample of the intervention or comparison group was randomly selected.
- **Exclusions based on preintervention characteristics**. Excluding a subset of the initial sample from the analysis is not considered attrition if exclusions are based on characteristics determined prior to the introduction of the intervention and applied consistently across the intervention and comparison conditions. For example, students who were excluded from data analysis because they had individualized education programs prior to the study would not be counted as attrition if they were excluded from both conditions. The WWC considers characteristics that are unlikely to change over time, including sex and race or ethnicity, as having been determined prior to the introduction of the intervention, even when the researchers collected these data later.

The WWC presumes that sample exclusions arising from sources other than those described above could be related to outcomes and, therefore, could constitute attrition.

Compositional change in cluster RCTs

A reviewer will follow a different process for assessing compositional change in cluster RCTs. The risk of bias in cluster RCTs can arise from two sources:

- A change in the **sample of clusters** (unit of assignment), and
- A change in **individuals within clusters** (unit of measurement).

Figure 7 shows the steps involved in reviewing compositional change in cluster RCTs. Cluster QEDs and compromised cluster RCTs should skip the steps on compositional change (steps 3a, 3b, 3c, 3d). and proceed to the steps described in Chapter III, Baseline equivalence standard (steps 4a, 4b).

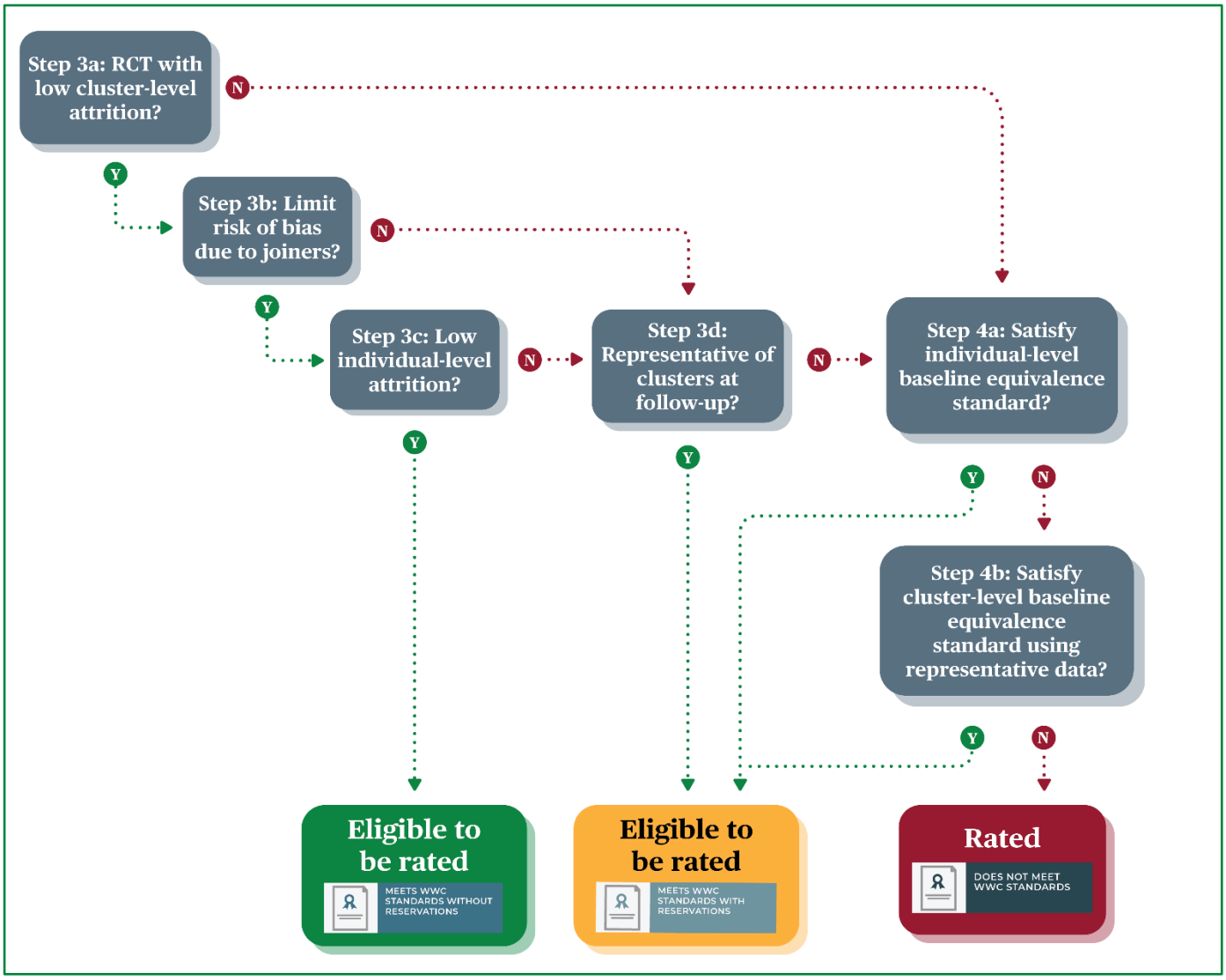
To receive the highest rating of *Meets WWC Standards Without Reservations*, a cluster RCT must:

- Have low cluster-level attrition (step 3a),
- Limit risk of bias due to joiners (step 3b), and
- Have low individual-level attrition (step 3c).

Passing all three steps means that the RCT has a low risk of bias due to compositional change. Not passing any one of these steps will reduce the maximum rating for the study to *Meets WWC Standards With Reservations* due to concerns about compositional change in clusters (step 3a) or individuals within clusters (step 3b or 3c), reflecting a high risk of bias due to compositional change. A reviewer reviewing a cluster RCT needs to evaluate whether a study passes all steps by following the guidelines that follow.

For RCTs that need to satisfy the baseline equivalence standard, the choice of attrition boundaries (steps 3a and 3c) and presence of high-risk joiners in the analytic sample (step 3b) can affect the available options for satisfying the individual-level or cluster-level baseline equivalence standard, as described later in Chapter III, When cluster RCTs can satisfy the baseline equivalence standard via adjustment only.

Figure 7. *Ratings flowchart for cluster-level assignment studies*



Step 3a. Does the finding have low cluster-level attrition?

Findings from cluster RCTs must have low cluster-level attrition to be eligible to receive the rating *Meets WWC Standards Without Reservations*. A cluster is lost when no individuals from it contribute outcome data to the analytic sample. Review teams must first choose an appropriate attrition boundary, as described above. Review team leadership should decide the attrition boundary based on whether the loss of clusters may be related to intervention status; this determination may differ from the choice of attrition boundary for assessing individual-level attrition, as detailed later. To determine whether cluster-level attrition is high or low, a reviewer should use the same tables for determining low versus high attrition as for individual-level RCTs described in appendix C in the technical appendices.

Step 3b. Does the finding limit the risk of bias due to joiners?

The WWC defines joiners as individuals who enter clusters after the results of random assignment are known to anyone who could plausibly influence individuals' placement into clusters, including family members,[8] students, teachers, principals, or other school staff. Joiners change the composition of the study sample and, therefore, introduce uncertainty—or bias—about the causes of the intervention effect.

> *Example*. Consider an RCT that randomly assigns classrooms to receive an advanced mathematics curriculum or business-as-usual instruction before student classroom rosters are set. School staff could view the curriculum as better suited to higher-performing students, preferentially placing such students into intervention classrooms. This differential placement could bias intervention effect estimates for student outcomes; students in intervention classrooms could outperform those in comparison classrooms solely due to preexisting student differences rather than improvements due to the intervention.

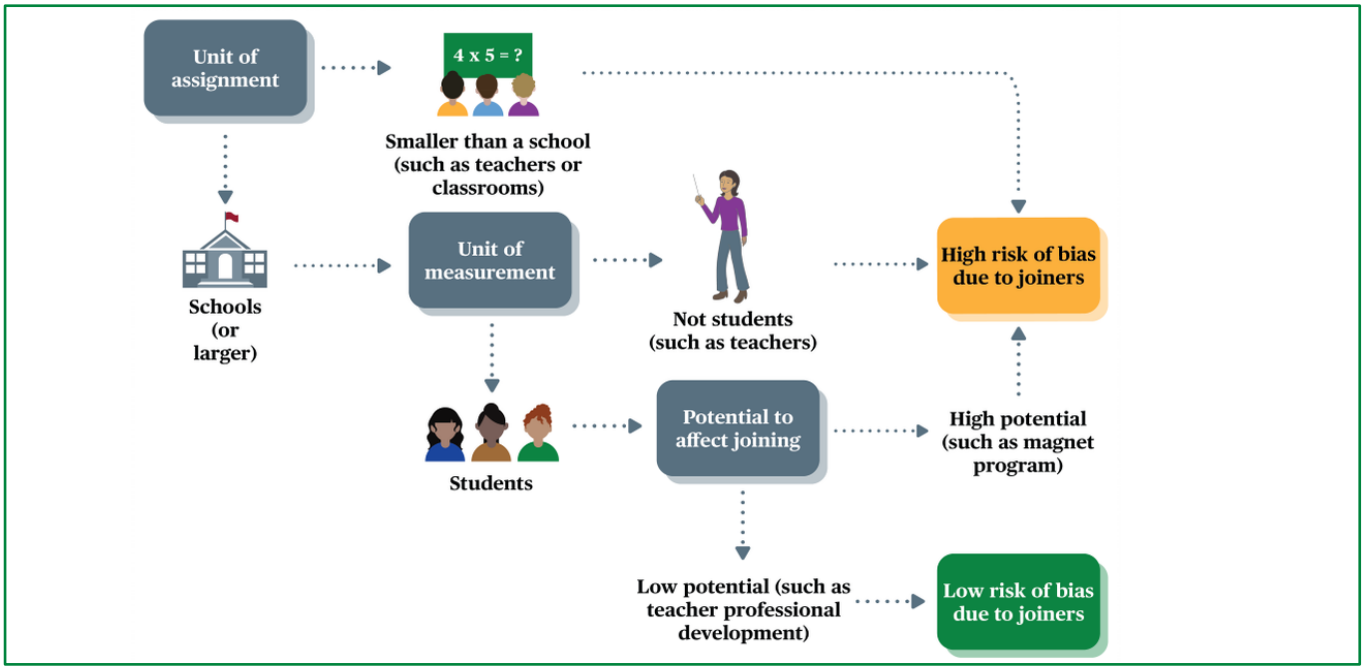
However, the risk of bias due to joiners may be low in other cases. Joiners in RCTs that randomly assign neighborhood schools to conditions often pose a low risk of bias because switching schools is usually much more difficult than switching classrooms or teachers within schools. Individuals who influence students' school placements—such as family members and students themselves—may be unaware of the intervention. Even if they are aware, they may decide that switching schools for the intervention may require too much effort or disruption. Including joiners in the analytic sample would pose a low risk of bias in this type of case.

Some individuals may enter clusters after random assignment but before anyone who could plausibly influence individuals' placement into clusters knew about the results of random assignment. The WWC never considers these individuals to pose a risk of bias because the intervention could not have affected their placements.

Review guidelines. A reviewer needs to assess the risk of bias due to joiners based on three factors: (a) unit of assignment, (b) unit of measurement, and (c) potential to affect joining (see figure 8). The WWC generally considers the risk of bias to be high for units of assignment smaller than a school or units of measurement other than students. These studies must limit that risk by excluding joiners from analyses; otherwise, such RCTs cannot receive the highest research rating, *Meets WWC Standards Without Reservations*. In contrast, the WWC considers the risk of bias to be low for school-level RCTs assessing student outcomes with low potential to affect joining, as detailed in the following sections. These studies may either include or exclude joiners from analyses and still be eligible for the highest research rating.

[8] The WWC uses the terms family and family member to include all caregivers.

Figure 8. *Judging the risk of bias due to joiners in cluster randomized controlled trials (step 3b)*



Unit of assignment. The WWC considers the risk of bias due to joiners to be high when the unit of assignment is smaller than a school, such as for teacher-level or classroom-level RCTs. Switching teachers or classrooms is typically easier than switching schools. School staff also may assign students to teachers or classrooms based on knowledge of the intervention.

The unit of assignment determines who could plausibly influence enrollment decisions. Often, more people may influence within-school placements than between-school placements (see table 8). The study authors have the burden for showing what the relevant people could not have known and when.

> *Example*. Consider an RCT in which random assignment of schools was announced to teachers in summer 2019, but the intervention was not known to families until fall 2019. Teachers' knowledge of the intervention is likely irrelevant because teachers typically do not influence families' school choice decisions. Students who entered schools in summer 2019 would not pose a high risk of bias in this case.

The WWC defines joiners as individuals who enter clusters that are the unit of assignment. Individuals who move within the units of assignment are not joiners. For example, if the unit of assignment is a school, students are not joiners if they move between classes within the school. Students are joiners only if they enroll in a new school.

Table 8. *People influencing students' placement into clusters by level of assignment*

Level of assignment	Risk of bias due to joiners	People typically influencing placement into clusters
School level or higher	Typically low	Students, family members
Smaller than a school (such as classrooms)	High	Students, family members, teachers, other school staff

Unit of measurement. The WWC considers the risk of bias due to joiners to be high when the unit of measurement is not students, such as for analyses of teacher outcomes in school-level RCTs. Compared with students and their family members, teachers may have greater knowledge of educational interventions offered at a school and may select a school based on factors such as the school's participation in an intervention.

The WWC only assesses joiners for the unit of measurement. The WWC does not consider individuals as joiners if they are not the unit of measurement.

> *Example*. A school-level RCT of student outcomes may evaluate the effects of a teacher incentive program intended to attract high-performing teachers to schools. Although teachers may choose a school because of the intervention, the WWC does not consider them joiners for analyses of student outcomes.

Potential to affect joining. The WWC generally considers school-level RCTs of student outcomes to pose a low risk of bias due to joiners, unless the intervention has a high potential to affect joining. To have a high potential to affect joining, an intervention must have both high intervention visibility and high enrollment flexibility, as defined in the following guidance. Interventions with low visibility or limited enrollment flexibility have a low potential to affect joining.

- **Intervention visibility: Are those with the ability to manipulate enrollment aware of the intervention and could that knowledge plausibly influence placement decisions?** Intervention visibility depends on whether the individuals who could manipulate enrollment both (a) are aware of the contrast in educational services provided across groups and (b) view those differences as a key factor in enrollment decisions. Even highly publicized programs could appear similar across conditions, such as when comparing alternative versions of an intervention. Other interventions might be visible to some individuals (such as teachers) but not the individuals who could plausibly affect students' enrollment decisions. A teacher professional development program, for example, likely has low visibility for families selecting schools because (a) students and parents are probably unaware of the intervention and (b) they might not view professional development as an important enough factor to switch schools. In contrast, examples of highly visible interventions may include magnet programs with a science and technology focus, afterschool programs, highly publicized school turnaround initiatives, and highly publicized programs for struggling students. Interventions might be more visible to postsecondary students and adult learners, who may select institutions or learning programs based on the educational services provided, as opposed to many K–12 students, who may be receiving interventions during normal school hours in neighborhood schools.
- **Enrollment flexibility: Is it relatively easy to join a study cluster after random assignment?** Four considerations for determining enrollment flexibility are (a) grade level transitions with regular cross-school changes, (b) afterschool or other supplemental programs that may draw students enrolled in other schools, (c) longer versus shorter time spans between random assignment and follow-up data collection, and (d) differences for postsecondary students and adult learners versus K–12 students. First, many school systems have grade level transitions with cross-school changes such as from grade 8 in middle school to grade 9 in high school, increasing enrollment flexibility. For instance, a rising grade 9 student may partly decide their high school choice based on Advanced Placement opportunities at the school. Second, afterschool or other supplemental programs may sometimes draw students who are enrolled in other schools, increasing

enrollment flexibility. Third, joining schools may be easier with longer time spans between random assignment and follow-up data collection, especially for outcomes collected two or more years after random assignment. Hence, review team leadership may allow the risk of bias determination to differ based on when the outcomes that were measured. Fourth, enrollment flexibility may be greater for postsecondary students and adult learners, compared with K–12 students, since older students may exercise greater control over their educational experiences. This point especially applies if institutions or programs were randomly assigned to conditions before individuals in the analytic sample enrolled into them.

Step 3c. Does the finding have low individual-level attrition?

The WWC defines individual-level attrition in cluster RCTs as (a) individuals who leave the clusters after random assignment—called leavers—and (b) individuals who are present in the clusters at follow-up but have missing outcome data. Both types of individual-level attrition can pose a risk of bias.

- Leavers can pose a risk of bias if the intervention makes some individuals more or less likely to leave the clusters. The intervention could impose a burden that increases leaving. In a classroom-level RCT, receiving a challenging mathematics curriculum could make lower-performing students move to other classrooms. The intervention also could provide benefits or supports (such as a dropout prevention program) that make individuals more likely to stay, decreasing leaving. Interventions that affect the likelihood of leaving (in either direction) could pose a risk of bias for some types of cluster RCTs.
- Individuals with missing outcome data (but who are present in the clusters at follow-up data collection) could pose a risk of bias due to aspects of the data collection process that could differ across conditions. For instance, researchers could obtain higher student consent rates in intervention schools than comparison schools. Other common missing data issues could include differential teacher compliance, student absences, sampling design, or individual-level recruitment procedures. These issues could pose a risk of bias if they differ across study conditions. These concerns are especially relevant for researcher-collected data (as opposed to administrative data).

Review guidelines. WWC reviewers need to evaluate individual-level attrition in cluster RCTs by comparing the analytic sample size with a reference sample size, such as the number of students present in the clusters at random assignment. The allowable reference samples depend on the WWC's determination of the risk of bias due to leavers, which can be low or high (figure 9). The risk of bias due to individual leavers can differ from the risk of bias due to joiners, depending on the nature of the intervention and review team leadership's judgment. These judgments regarding the risk from various types of compositional change need to be documented in the study review.

Choosing attrition boundaries. In addition to informing the allowable reference samples, the risk of bias due to leavers should inform the review team's choice of the cautious versus optimistic boundary for assessing individual-level attrition. If the risk of bias due to leavers is high, review team leadership should always use the cautious boundary for assessing individual-level attrition. If the risk of bias due to leavers is low, the optimistic boundary is usually appropriate for assessing individual-level attrition, but review team leadership should still justify its use, as described in Chapter III, Determining optimistic or cautious attrition boundary when assessing attrition. This choice for individual-level attrition is separate from choosing the attrition boundary for cluster-

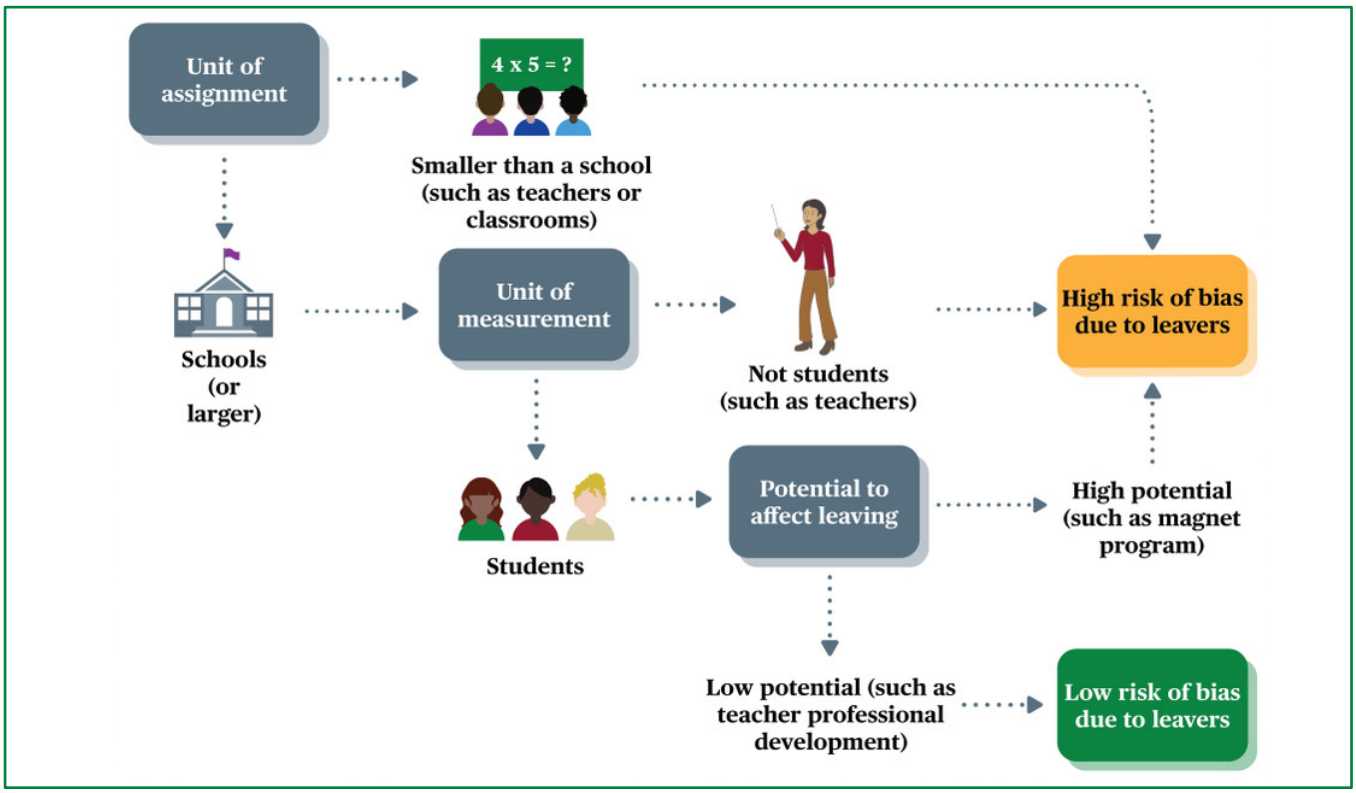
level attrition (see step 3a). Even when the optimistic boundary is appropriate for assessing individual-level attrition, the cautious boundary could remain appropriate for assessing cluster-level attrition.

Judging the risk of bias due to leavers. The initial categorization into high versus low risk of bias due to leavers focuses only on leavers, not other types of individual-level attrition such as individuals who remain in the clusters but have missing outcome data. For RCTs with a low risk of bias due to leavers, there are more options for allowable reference samples than for RCTs with a high risk of bias due to leavers. For RCTs with a high risk of bias due to leavers, the WWC will assess individual-level attrition using an early reference sample and the cautious attrition boundary. RCTs with a high risk of bias due to leavers can still receive the highest research rating of *Meets WWC Standards Without Reservations*, but findings from these studies need to meet more stringent individual-level attrition requirements than RCTs with a low risk of bias due to leavers.

In general, units of assignment smaller than a school and units of measurement other than students are more likely to pose a high risk of bias due to leavers (see the previous step about joiners). In contrast, school-level RCTs of student outcomes will often have a low risk of bias due to leavers.

One exception for school-level RCTs of student outcomes, however, is interventions with a high potential to affect leaving. Having a high potential to affect leaving means that (a) the intervention could plausibly influence student mobility and (b) enrollment flexibility is high. If both conditions are met, review team leadership should classify the risk of bias due to leavers as high and should use the cautious attrition boundary.

Figure 9. *Assessing risk of bias due to leavers in cluster randomized controlled trials (step 3c)*



Assessing the risk of bias due to leavers focuses on two questions:

- **Intervention focus: Could the intervention plausibly influence student mobility?** This question focuses on the ways students might experience receiving the intervention. Important considerations are whether the intervention (a) provides benefits and supports that may induce students to stay or (b) imposes a significant burden that may motivate students (or their families) to leave. A dropout prevention program for high school students, for example, directly targets keeping students in school. Such a program could affect who stays in intervention clusters, even if the program was not widely publicized. An intervention also could have unintended consequences on student mobility. For example, a highly challenging mathematics curriculum might lead students to leave intervention clusters. In contrast, a teacher professional development intervention focused on teaching fractions for elementary students is unlikely to affect student mobility.
- **Enrollment flexibility: Is it easy to leave a study cluster after random assignment?** The previous step about joiners (step 3b) introduced four general considerations for determining enrollment flexibility that also apply here: (a) grade level transitions with regular cross-school changes, (b) afterschool or other supplemental programs that may draw students enrolled in other schools, (c) longer versus shorter time spans between random assignment and follow-up data collection, and (d) differences for postsecondary students and adult learners versus K–12 students. In some cases, however, enrollment flexibility may be greater for leaving than for joining. For instance, dropping out of a college or a high school may be easier than entering a new one. Review team leadership has the discretion to consider aspects of the study context (such as grade level) in making such determinations.

Measuring individual-level attrition. The WWC defines individual-level attrition as the number of individuals in an allowable reference sample minus the number in the analytic sample. The risk of bias due to leavers determines whether reference sample (1), (2), or (3) in table 9 is allowable. In all cases, the WWC measures individual-level attrition within clusters that had at least one individual who contributed outcome data. These clusters are called nonattriting clusters. Clusters that did not contribute any outcome data are considered cluster-level attrition; students in those attriting clusters are not considered individual-level attrition.

The reference sample counts should generally include the total number of individuals within the units of assignment, regardless of whether individuals received the intervention. For example, if the unit of assignment is a school, the reference sample should include all students within that school. However, studies may apply certain exclusions that do not count as attrition, such as excluding based on pre-randomization characteristics. The reference sample should not include these types of exclusions. For instance, if the study only analyzed grade 3 students, then the reference sample should only include grade 3 students, assuming student grade level is a pre-randomization characteristic. These types of exclusions are described in more detail later in this section.

Table 9. *Attrition boundaries and allowable reference samples for measuring individual-level attrition in cluster randomized controlled trials (step 3c)*

Risk of bias due to leavers	Attrition boundary	Allowable reference samples
High risk	Cautious is always appropriate	(1) Individuals in nonattriting clusters prior to the intervention announcement, or (2) Individuals in nonattriting clusters in an early period.
Low risk	Optimistic is usually appropriate	Either (1), (2), or (3) Individuals in nonattriting clusters at follow-up.

The WWC will assess individual-level attrition in cluster RCTs with a high risk of bias due to leavers using the cautious attrition boundary and either of the following options for an early reference sample:

- One option is reference sample (1): individuals present prior to the intervention announcement. The WWC defines this time point as before the actors who could plausibly influence individuals' placement into clusters knew the results of random assignment. For instance, a study that randomly assigned schools in summer 2019 could use spring 2019 enrollment counts as the reference sample sizes.
- Another option is reference sample (2): individuals present in an early period. As a default, the WWC defines the early period as the first six weeks of the first school year for intervention implementation. For instance, a study that randomly assigned schools in summer 2019 could use early fall 2019 enrollment counts as the reference sample sizes.

Using these early reference samples incorporates both leaving and missing outcome data into the assessment of individual-level attrition. Both reference samples (1) and (2) will generally be acceptable, but review team leadership may override this default if even early leaving may pose a risk of bias. For instance, review team leadership may instead allow only reference sample (1), but not (2), for interventions focused on enrolling or recruiting students (such as magnet programs or charter schools).

For cluster RCTs with a low risk of bias due to leavers, the WWC also may assess individual-level attrition using reference sample (3): individuals present in clusters at follow-up. This reference sample assesses individual-level attrition due to missing outcome data as opposed to a student "leaver," capturing whether the analytic sample is representative of the individuals in clusters at follow-up data collection. This late reference sample is acceptable because empirically assessing leaving is not needed if the risk of bias due to leavers is low. However, the earlier reference samples, (1) or (2), are also allowable for cluster RCTs with a low risk of bias due to leaving.

Five other considerations for computing individual-level attrition include the following:

- **Multiple allowable reference samples:** Multiple reference samples may be allowable, especially for RCTs with a low risk of bias due to leavers. In practice, however, study authors may report sufficient information to use only one specific reference sample, narrowing the options for WWC study reviewers. When multiple allowable reference samples are provided, the WWC will prioritize the one that best aligns with how the study authors defined their sampling of individuals. For instance, if the study authors analyzed joiners and stayers, then the WWC will prefer using reference sample (3), if it is allowable. If the study authors analyzed

only stayers (excluding joiners), then the WWC will instead prefer (1) or (2), even if (3) is provided and allowable.

- **Analyses of administrative data:** The WWC provides further flexibility for analyses of administrative data sources such as district records of student test scores. The WWC assumes that administrative data have low rates of missing outcome data for individuals in clusters at follow-up, unless review team leadership concludes that patterns of missing administrative data have a high risk of differing across intervention versus comparison groups. In those cases, WWC reviewers do not need to compute individual-level attrition rates for RCTs that analyzed administrative data and have a low risk of bias due to leavers. The WWC instead assumes the individual-level attrition is low for reference sample (3). However, for RCTs with a high risk of bias due to leavers, the WWC will still compute individual-level attrition using samples (1) or (2), even for analyses of administrative data.
- **Joiners in the analytic sample:** In some cases, the analytic sample may include individuals who were not in the reference sample, which poses complications for assessing whether individual-level attrition is a concern. In a school-level RCT with a low risk of bias due to joiners, the analyses could include joiners who were not present in reference samples (1) or (2). As noted, the WWC prefers using reference sample (3) in this scenario if it is allowable, avoiding that type of complication by including joiners in both the numerator and the denominator of the individual-level attrition calculation. However, if sample (3) is unallowable or unavailable, then the WWC will attempt to calculate individual-level attrition excluding such joiners. For instance, consider a study that has 100 students present in the clusters at random assignment based on reference sample (1). The analytic sample had 110 students, including 40 students who joined between random assignment and follow-up data collection. In this case, the WWC will first adjust the analytic sample size by subtracting the individuals who were not present in the reference sample (yielding a corrected analytic sample size of 110 – 40 = 70 stayers in this example). The individual-level attrition would then be the difference between the reference sample size (100 students) and the analytic sample size of stayers (70 students), yielding an individual-level attrition rate of 30 percent in this case. The individual-level attrition rate will be missing—and presumed to be high—if the number of joiners in the analytic sample is required information for calculating individual-level attrition but cannot be determined.
- **Sample exclusions that do not count as individual-level attrition:** Like individual-level RCTs, cluster RCTs may exclude some individuals based on (a) random subsampling, (b) preintervention characteristics like race and gender, or (c) acts of nature that likely affect intervention and comparison groups equally. These types of sample exclusions should not count as individual-level attrition of individuals if applied consistently across intervention and comparison groups. In this case, the reference sample should include only eligible individuals. For instance, if the analyses included only English learner students, then the reference sample sizes should include only English learner students. Exclusions also do not count as attrition if they were based on characteristics determined after random assignment but before the results of random assignment were known to anyone who could plausibly influence individuals' placement into clusters.
- **Selective or elective enrollment interventions:** Studies may assign clusters to receive selective or elective enrollment interventions such as Advanced Placement courses that apply to only a subset of individuals within clusters. As noted earlier, the reference sample should generally include all students in the clusters, unless the study applied an exclusion that does not count as attrition. Consider an RCT of student outcomes

that randomly assigned schools to receive different versions of an Advanced Placement calculus curriculum. If the study analyzed outcome data from only students who enrolled in the course, then students who did not enroll in the course would generally count as individual-level attrition. However, if course enrollments were determined prior to random assignment, then excluding non-enrollees would not count as attrition.

The assessment of individual-level attrition rates will follow the same attrition requirements detailed in Chapter III, Attrition in individual-level RCTs. One minor difference is that the risk of bias due to leavers will generally guide the choosing of the optimistic versus cautious attrition for individual-level attrition (see the earlier step 3a for cluster-level attrition). If individual-level attrition exceeds the attrition boundary or cannot be determined (based on information in the study article or from an author query), then the highest possible research rating is *Meets WWC Standards With Reservations*.

Step 3d. Was the analytic sample of individuals representative of clusters at follow-up?

Some cluster RCTs may have low cluster-level attrition, but the WWC may still have concerns about the compositional change of individuals within clusters, such as for joiners (do not pass step 3b) or individual-level attrition (do not pass step 3c). In these cases, a low cluster-level attrition RCT may still receive the rating *Meets WWC Standards With Reservations* if the analytic sample is representative of individuals in clusters at follow-up. This review route provides an alternative to meeting WWC standards for low cluster-level attrition RCTs that are unable to collect baseline data but can demonstrate representativeness.

Representativeness in cluster-level assignment studies

Demonstrating representativeness means that missing outcome data rates were low for the population of individuals present in clusters at follow-up. High representativeness is the same as low individual-level attrition when attrition is assessed using the reference sample of individuals in nonattriting clusters at follow-up.

Demonstrating representativeness addresses concerns about potential bias in data collection procedures such as differential sampling or consent across conditions. But representativeness does not address concerns that the intervention could have affected who joined clusters (a reason for not passing step 3b) or left clusters (a reason for not passing step 3c), which is why the study cannot receive a higher research rating than *Meets WWC Standards With Reservations*.

Assessing representativeness differs from the previous step on individual-level attrition (step 3c) in that the reference sample is always the individuals in nonattriting clusters at follow-up. Otherwise, the relevant calculations are the same, using either the cautious or optimistic attrition boundary as described previously. Similarly, the WWC assumes that analyses of administrative data satisfy the representativeness requirement, unless review team leadership concludes that patterns of missing administrative data have a high risk of differing across intervention and comparison groups.

Low cluster-level attrition RCTs with demonstrated representativeness do not need to satisfy the baseline equivalence standard to be eligible to receive a research rating of *Meets WWC Standards With Reservations*. These studies have already addressed several concerns about cluster-level and individual-level compositional change, even if some reservations remain regarding joiners and leavers. Otherwise, RCTs at this step but without demonstrated representativeness must satisfy the baseline equivalence standard. A reviewer reviewing a low

cluster-level attrition RCT without demonstrated representativeness should examine baseline differences described in Chapter III, Baseline equivalence standard.

Step 4. Baseline equivalence standard

Differences between the intervention and comparison groups at baseline can bias the estimated impact of the intervention. RCTs with a low risk of bias due to compositional change avoid this problem by design. In such RCTs, groups are expected to be equivalent on all observed and unobserved characteristics on average. Any characteristics that do vary across groups are presumed to do so by chance. As a result, such RCTs are eligible to receive the research rating of *Meets WWC Standards Without Reservations*. In contrast, all QEDs, all compromised RCTs, individual-level RCTs with high attrition, and some cluster RCTs as specified in figure 7 must satisfy the baseline equivalence standard to be eligible to receive the research rating of *Meets WWC Standards With Reservations*.

Studies can satisfy the baseline equivalence standard in three possible ways, as detailed in the following sections (see the box to the right for an overview). These options share the following common characteristics for satisfying the standard:

Three ways to satisfy the baseline equivalence standard

- Demonstrate baseline differences are 0.05 standard deviation or smaller
- Demonstrate baseline differences are between 0.05 to 0.25 standard deviation and apply an acceptable adjustment for baseline differences
- Apply an acceptable adjustment for baseline differences only (this option is available for only a subset of RCTs and RDDs)

- **At baseline**–that is, prior to or early in an intervention's implementation, and ideally, at the time of assignment.[9]
- **For an acceptable baseline sample**–in the case of individual-level assignment studies, the acceptable baseline sample only includes participants from the intervention and comparison groups used to estimate findings. In the case of cluster-level assignment studies, an acceptable baseline sample may not always be the analytic sample.
- **On relevant characteristics**–that is, study authors must satisfy the baseline equivalence standard on characteristics that are relevant to the outcome. The following sections specify relevant baseline characteristics for student outcomes as well as teacher and school leader outcomes.

When studies must demonstrate that baseline differences are no greater than 0.25 standard deviations

For all QEDs and compromised RCTs, satisfying the baseline equivalence standard requires baseline group differences, or effect sizes, no greater than 0.25 standard deviation on key covariates. Where differences are between 0.05 and 0.25 standard deviation (see table 10), study authors must statistically adjust for those baseline covariates in their impact analyses. Baseline differences less than or equal to 0.05 standard deviations in absolute value automatically satisfy the baseline equivalence standard and do not require statistical adjustment. Appendix E in the technical appendices describes the formulas the WWC uses to calculate these effect sizes.

[9] The WWC does not strictly require that baseline measurement occur prior to the start of intervention services.

Some RCTs without compromised random assignment must also demonstrate baseline differences are no greater than 0.25 standard deviation, as detailed in the next section.

Table 10. *Satisfying the baseline equivalence standard based on the absolute effect size at baseline*

0.00 ≤ |ES at baseline| ≤ 0.05	0.05 < |ES at baseline| ≤ 0.25	|ES at baseline| > 0.25
Satisfies the baseline equivalence standard	Requires statistical adjustment to satisfy the baseline equivalence standard	Does not satisfy the baseline equivalence standard

ES is effect size.

When RCTs can satisfy the baseline equivalence standard via adjustment only

Some RCTs only need to apply a statistical adjustment for baseline differences to satisfy the baseline equivalence standard. For individual-level RCTs, this option is acceptable for uncompromised RCTs with high attrition assessed under the optimistic attrition boundary. Such RCTs only need to statistically adjust for key covariates to satisfy the baseline equivalence standard and be eligible for the research rating *Meets WWC Standards With Reservations*. In contrast, individual-level RCTs with high attrition assessed under the cautious attrition boundary must follow the guidelines in table 10 for satisfying the baseline equivalence standard. Analogous guidelines apply to cluster RCTs but are more complex because cluster RCTs have multiple types of compositional change, as detailed in Chapter III, When cluster RCTs can satisfy the baseline equivalence standard via adjustment only.

Baseline characteristics for student outcomes

For student outcomes, the baseline equivalence standard may be satisfied using either a pretest in the same domain as the outcome, or a pretest in a broader domain in the same content area and for which the outcome domain is a subset. Alternatively, if a single pretest in a broader domain is unavailable, study authors may include multiple pretest measures that collectively represent a broader domain.

> *Example*. For an outcome measure in the Algebra domain, the baseline equivalence standard may be satisfied using a measure of General Mathematics Achievement. However, for an outcome measure in the General Mathematics Achievement domain, the baseline equivalence standard cannot be satisfied using only a measure in the Algebra domain.

If a pretest is not available. If a pretest in the same outcome domain—or in a broader domain in the same content area encompassing the outcome—was not given or is not available for certain outcomes, then the following may be used to satisfy the baseline equivalence standard:

1. A broad, approximately continuous,[10] and standardized measure of student academic readiness, knowledge, or skills, AND

[10] Examples of measures that do not qualify as approximately continuous include pass/fail status, proficiency or grade-level benchmark status, and letter grades.

2. Baseline measures of at least two of the following for learners in the analytic sample:
    a. A measure of socioeconomic status, such as parental or caregiver level of education or eligibility for need-based assistance or financial aid
    b. Race or ethnicity
    c. Dual language or English learner student status
    d. Disability status
    e. Disciplinary measures such as frequency of suspensions or referrals
    f. Grade level, for students between kindergarten and grade 12, or else age
3. AND, for studies of interventions implemented for students younger than kindergarten or for adult education programs, the age of the learners.

For the first set of baseline measures above, the broad measure of student academic readiness, knowledge, or skills should be drawn from one of the following domains: Cognition, Academic Achievement, General Literacy Achievement, General Mathematics Achievement, or Postsecondary Academic Achievement.[11] Broad measures within the Academic Achievement domain include both standardized tests and continuous measures of student grade point average in grades 6-12. Broad measures within the Postsecondary Academic Achievement domain include both standardized tests and measures of student grade point average in college courses.

Baseline measures must satisfy standards

Measures that study authors use to satisfy the baseline equivalence standard must satisfy the same standards as the outcome measures as described in Chapter III, Outcome measure standards.

If the outcome is a broad measure of knowledge or skills, such as standardized academic achievement measures, then using a broad baseline measure of knowledge or skill is sufficient for satisfying the baseline equivalence standard, and therefore authors would not need to satisfy the standard for additional student or contextual characteristics, even if the baseline measure is in a different outcome domain. For example, if the outcome were a measure of science achievement in grade 8, then the study authors could use a measure of math achievement from grade 7 to satisfy the baseline equivalence standard. However, if the outcome in this example were high school completion, then the study authors would also need to satisfy the baseline equivalence standard for additional baseline characteristics from the second set of baseline measures.

When study authors need to satisfy the baseline equivalence standard for a broad measure of achievement and additional characteristics outlined in the second set of baseline measures above, adjustments for these additional baseline characteristics may occur as part of an acceptable statistical adjustment—as described below—or through the sampling design. For example, if a study sample was composed of 100 percent English learner students in grade 2, that study effectively adjusts for English learner student status and grade level. If study authors

[11] Baseline measures in the Proficiency in the English Language domain are acceptable for early childhood education studies.

demonstrate that the baseline differences in the analytic sample are less than 0.05 standard deviation for each required covariate, then no further statistical adjustments are required to satisfy the baseline adjustment requirement.

Baseline characteristics for teacher and school leader outcomes

For teacher and school leader outcomes, study authors should satisfy the baseline equivalence standard for relevant baseline characteristics summarized in table 11.

Table 11. *Baseline adjustment measures for teacher and school leader outcomes*

Outcome domains	Acceptable baseline measures
Teacher or School Leader Practice	The same measure as the outcome or another measure from the same domain as the outcome (and at the same unit of analysis)
Teacher or School Leader Retention Teacher Attendance Other Teacher of School Leader domains (consult *Study Review Protocol*)	• The same measure as the outcome or another measure from the same domain as the outcome (and at the same unit of analysis) OR • Average years of teacher or school leader experience or the experience categories used in the study AND one of the following: • An aggregate measure from the same domain as the outcome for the cluster or other cohorts OR • A broad, approximately continuous, standardized measure of student academic readiness, knowledge, and skills and baseline measures of at least two learner characteristics in the analytic sample as specified above for student outcomes when a pretest is not available

Other considerations for baseline equivalence

Additional considerations regarding assessing and satisfying the baseline equivalence standard include the following:

- The baseline equivalence standard must be satisfied separately for each analytic sample. Satisfying the baseline equivalence standard on one analytic sample does not affect the requirement for other analytic samples, even for outcome measures in the same domain. For example, consider a QED that measured impacts using both the full sample and a sample that excluded one student. In this example, it is necessary to assess baseline equivalence on each sample separately. However, as detailed later, acceptable baseline samples for cluster-level assignment studies do not always have to be the analytic sample of individuals.
- A difference larger than 0.25 standard deviation for any specified preintervention measure in a domain means that all outcomes in the domain fail to satisfy the baseline equivalence standard because domains are typically defined to include outcomes that are thought to be highly correlated.
- Preintervention measures used to satisfy the baseline equivalence standard must satisfy the same reliability criteria specified for outcomes, as described in Chapter III, Outcome measure standards. If reliability

information for a preintervention measure is required but unavailable, or if the reliability is below the acceptable level, then the measure cannot be used to assess baseline equivalence.

- If a significant portion of the intervention occurred prior to the assessment of a baseline measure used to satisfy the baseline equivalence standard, then the WWC will note in its reporting that the study measures the effect of the portion of the intervention that occurred after the measure was assessed and until the time of the follow-up assessment. If both preintervention and intermediate measures are available, then the WWC will use the preintervention measure to assess baseline equivalence.
- Chapter V, Procedures and standards for analyses with imputations for missing data discusses additional considerations for assessing baseline equivalence in studies with missing or imputed data. First, while all QEDs must satisfy the baseline equivalence standard, high-attrition, individual-level RCTs (but not cluster RCTs) that impute outcome data and analyze the full sample that was randomized to conditions do not need to satisfy the baseline equivalence standard to be eligible to be rated *Meets WWC Standards With Reservations*, as described in Chapter V, Is the study a high-attrition RCT that analyzes the full randomized sample using imputed data?. These studies must, however, demonstrate that the risk of bias due to analyzing missing or imputed outcome data is low. Second, if the analytic sample for individual-level assignment study includes missing or imputed data for a specified preintervention measure, then it must satisfy the baseline equivalence standard using the largest baseline difference under different assumptions about how the missing data are related to measured or unmeasured factors, as described in Chapter V, Are data in the analytic sample missing or imputed for any baseline measure specified in the Handbook?. Finally, all studies must use one of the acceptable approaches listed in table H.1 in appendix H in the technical appendices to address missing data in the analytic sample to be eligible to meet WWC standards.
- If the study used variable unit weights in the impact analysis, where units contribute more or less to the impact estimate than other units, then the baseline means also must be calculated using the same weights.
- If the study conducted random assignment within blocks or matching within strata, and the analysis includes dummy variables that differentiate these blocks or strata, then the baseline means also may be adjusted using these same dummy variables (Wolf et al., 2017).

Acceptable adjustment strategies for QEDs and for high-attrition or compromised RCTs, and for studies with individual-level versus cluster-level assignment are summarized separately in the following section.

Acceptable baseline adjustment strategies

For studies that must adjust for baseline differences to satisfy the baseline equivalence standard, the WWC considers any of these adjustment methods appropriate: regression adjustment (such as with ordinary least squares or analysis of covariance, hierarchical linear models, or generalized linear models), matching or weighting (such as propensity score matching or inverse propensity score weighting), difference-in-differences estimation, analysis of simple gain scores, assignment unit and intervention period fixed effects, and bounding techniques (figure 10). The key to each of these adjustment strategies is that they are implemented with respect to the baseline covariates specified in Chapter III, Baseline equivalence standard.

The WWC does not require further adjustment for baseline differences if study authors demonstrate baseline equivalence between the groups on relevant covariates. Baseline equivalence is defined as differences less than or equal to a Hedges' *g* effect size of 0.05 standard deviation.[12]

Figure 10. *Acceptable methods for baseline adjustment*

Acceptable methods for any baseline measure

- Regression covariate adjustments in ordinary least squares models.
- Regression covariate adjustments in hierarchical linear models.
- Analysis of covariance.
- Other approaches to regression covariate adjustments, including generalized linear models.
- Matching and weighting methods.
- Bounding techniques.

Acceptable methods when the baseline and outcome measures have the same measurement scale and the pretest-posttest correlation is .60 or greater

- Simple gain scores.
- Difference-in-differences adjustment.
- Fixed effects for units of assignment and intervention periods.

Three acceptable adjustment strategies—simple gain scores, difference-in-differences estimation, and assignment unit and intervention period fixed effects estimation—must satisfy two additional criteria for the WWC to consider the adjustments acceptable:

- **The baseline and outcome measures must be in the same outcome domain and on the same measurement scale.** For example, this condition would be satisfied if the researchers administered the same measure, using the same scoring procedures, as a pretest and as a posttest. Analyzing pretest-posttest gain scores would be an acceptable adjustment strategy in this case. If the baseline and outcome measures differ, study authors could subtract the standardized baseline effect size from the standardized outcome effect size, which would address this requirement.
- **The baseline measure must have a correlation of 0.60 or higher with the outcome.** In general, the correlation must be estimated using the study data. Review teams may waive this requirement for a measure or outcome domain if the protocol documents evidence that the correlations between pretests and the posttests of the measure typically exceed 0.60, and the exception is applied consistently for all studies within the review.

[12] Using the pooled sample of intervention and comparison group members, differences of less than or equal to 0.05 standard deviation in absolute value on the specified baseline characteristic are considered to have satisfied the requirement of adjustment for baseline differences.

Additional considerations for statistical adjustments in some common analytic approaches, such as analyses in which outcomes are collected multiple occasions, often referred to as time series or repeated measures designs, are described in Chapter V.

Special considerations for cluster-level assignment studies

The guidance in the previous sections apply to both individual-level and cluster-level assignment studies regarding assessing the baseline equivalence standard. Some special considerations apply to cluster-level assignment studies, as detailed in this section. Therefore, a reviewer of a cluster-level assignment study should follow the guidance in this and previous section, unless otherwise noted in this section.

All cluster QEDs and compromised cluster RCTs must satisfy the baseline equivalence standard to be eligible for the research rating *Meets WWC Standards With Reservations*. The process in figure 7 determines which cluster RCTs must satisfy the baseline equivalence standard. For example, RCTs with high cluster-level attrition must satisfy the baseline equivalence standard. In contrast, RCTs with a low risk of bias due to compositional change (passes steps 3a, 3b, and 3c) and low cluster-level attrition RCTs with demonstrated representativeness (passes steps 3a and 3d) do not need to satisfy the baseline equivalence standard, assuming the random assignment was not compromised.

When cluster-level assignment studies must demonstrate that baseline differences are no greater than 0.25 standard deviation

All cluster QEDs and all compromised cluster RCTs must demonstrate that baseline differences are no greater than 0.25 standard deviation to satisfy the baseline equivalence standard, along with adjusting for baseline differences in the 0.05–0.25 standard deviation range. These guidelines are the same as for individual-level assignment studies (table 10).

Some cluster RCTs without compromised random assignment must also demonstrate that baseline differences are no greater than 0.25 standard deviation, as detailed in the next section.

When cluster-level RCTs can satisfy the baseline equivalence standard via adjustment only

Compared to individual-level RCTs, unique considerations apply to when cluster RCTs can satisfy the baseline equivalence standard via statistical adjustment only. This option applies to cluster RCTs that meet four conditions: (a) the optimistic attrition boundary was used for assessing at least one type of attrition (cluster-level or individual-level), (b) attrition was low when using the cautious boundary for individual-level or cluster-level attrition, if applicable, (c) the analytic sample did not include any high-risk joiners, and (d) random assignment was not compromised. All other cluster RCTs that need to satisfy the baseline equivalence standard instead must follow the guidelines noted in table 10, meaning that, at a minimum, baseline differences must be 0.25 standard deviation or smaller. The following example illustrates the application of conditions (a) and (b).

> *Example*. Consider a school-level RCT for which the cautious attrition boundary applies to assessing cluster-level attrition, the optimistic attrition boundary applies to assessing individual-level attrition, and the risk of bias due to joiners was low. The two attrition boundaries reflect the review team's judgment that more cautious assumptions should apply to assessing cluster-level attrition than individual-level attrition. If the study has high cluster-level attrition, then the study needs to follow the guidelines in

table 10 by, at a minimum, demonstrating that baseline differences are 0.25 standard deviation or smaller. If the study has low cluster-level attrition but high individual-level attrition, then the study would only need to apply an acceptable adjustment for baseline differences to satisfy the baseline equivalence standard; in this case, condition (b) was met as attrition was low for the type of attrition that required using the cautious boundary (cluster-level attrition in this example).

When cluster-level assignment studies can satisfy the individual-level versus cluster-level baseline equivalence standard

The options of demonstrating baseline differences no larger than 0.25 standard deviation and only adjusting for baseline differences apply to two versions of the baseline equivalence standard for cluster-level assignment studies:

- **Individual-level baseline equivalence standard (step 4a).** Satisfying the individual-level baseline equivalence standard means that the individuals who contributed outcome data are similar at baseline across intervention and comparison groups. The individuals contributing baseline assessment data and outcome data must be the same for evaluating this standard.
- **Cluster-level baseline equivalence standard (step 4b).** Satisfying the cluster-level baseline equivalence standard means that the clusters which contributed outcome data are similar at baseline across intervention and comparison groups. The individuals contributing baseline assessment data and outcome data could differ for evaluating this standard.

If the study cannot satisfy the individual-level baseline equivalence standard, satisfying the cluster-level baseline equivalence standard is sufficient instead (figure 7). If the sample of individuals used to assess baseline equivalence differs from the sample of individuals used in the analysis of an outcome, then a reviewer should start at step 4b, skipping step 4a. These samples can differ for multiple reasons, including due to missing or imputed data, as noted in Chapter V, Procedures and standards for analyses with imputations for missing data.

The following sections detail unique considerations that apply to the individual-level and cluster-level versions of the baseline equivalence standard, but a reviewer should also keep the guidance in previous sections in mind, such as when adjusting for baseline differences is required to satisfy the baseline equivalence standard.

Step 4a. Does the cluster-level assignment study satisfy the individual-level baseline equivalence standard?

Satisfying the individual-level baseline equivalence standard involves special considerations regarding (a) using the analytic sample of individuals to calculate the size of baseline differences and (b) applying adjustments that account for the individual-level correlation between the baseline measure and the outcome measure.

Calculating baseline differences for individuals in the analytic sample. The calculations for individual-level baseline differences must use the analytic sample of individuals included in the analysis of a specific outcome. For studies that analyze cluster-level outcome means, the baseline equivalence calculations must use the same individuals who contribute data to the outcome measure. The calculations can use either individual-

level or cluster-level standard deviations.[13] Baseline means calculated using either cluster- or individual-level data are acceptable as long as the weighting is consistent with the weighting in the impact analysis.

Acceptable baseline adjustments for studies satisfying the individual-level baseline equivalence standard. Acceptable adjustments for satisfying the individual-level baseline equivalence standard must account for the individual-level correlation between the baseline measure and the outcome measure, using individual-level baseline data from the same individuals who contribute outcome data. Study authors who apply an acceptable individual-level adjustment do not need to also apply a cluster-level adjustment because adjustments for individual factors (such as mathematics achievement) also can adjust for cluster-level differences. Studies that apply a cluster-level adjustment only, but not a required individual-level adjustment, do not satisfy the individual-level baseline equivalence standard but can satisfy the cluster-level baseline equivalence standard, as described next.

A reviewer should consult Chapter III, Acceptable baseline adjustment strategies for general guidelines on acceptable adjustment strategies.

Step 4b. Does the cluster-level assignment study satisfy the cluster-level baseline equivalence standard?

If the cluster-level assignment study cannot satisfy the individual-level baseline equivalence standard for individuals included in analyses, the study can instead satisfy the cluster-level baseline equivalence standard. That is, studies can demonstrate that intervention and comparison clusters are similar at baseline even if the individuals used for baseline assessment differ from the individuals included in analyses.

Demonstrating cluster-level baseline equivalence has three requirements:

1. Cluster-level baseline means must be computed using acceptable baseline samples.
2. Baseline differences must be accounted for using acceptable adjustments.
3. The outcome and baseline samples must demonstrate representativeness of individuals in the clusters.

Other than these three requirements, the review principles for evaluating cluster-level baseline equivalence in step 4b are the same as for individual-level baseline equivalence in step 4a.

Acceptable baseline samples for studies satisfying the cluster-level baseline equivalence standard. The WWC allows the following three types of samples for computing the cluster-level baseline means for cluster-level baseline equivalence:

1. Individuals in the analytic sample from any preintervention period.
2. Individuals from the same cohort, within the same clusters as the individuals in the analytic sample. The baseline data may be obtained at the time that clusters were assigned to conditions or during the year before clusters were assigned to conditions.

[13] Because cluster-level standard deviations are smaller than individual-level standard deviations therefore resulting in larger effect sizes, demonstrating baseline equivalence using cluster-level standard deviations is an acceptable approach to demonstrating baseline equivalence.

3. Individuals from the immediately preceding cohort, within the same grades and clusters as individuals in the analytic sample.

If study authors report multiple baseline samples, reviewers should prioritize the baseline samples in the order listed. For instance, WWC reviewers should prioritize cluster-level equivalence based on baseline sample 1, whenever possible, in which the individuals contributing baseline data are the same individuals contributing outcome data. Note that baseline sample 1 is the same sample used in calculating individual-level baseline equivalence; this sample could be used for cluster-level baseline equivalence as well if the study failed to apply a required individual-level adjustment but did apply a required cluster-level adjustment.

Consider the following baseline samples in table 12 for an example RCT that randomly assigned schools to conditions in summer 2014 and measured outcomes for grade 12 students at the end of the 2014/15 academic year. The numbering of samples within this table refers to the three types of acceptable baseline samples listed previously, such as samples 1a and 1b referring to the first type of sample listed previously.

Table 12. *Example baseline samples for an outcome sample of grade 12 students in 2014/15 (step 4b)*

Sample	Data source	Acceptable (yes or no)
1a. Same individuals as in the analytic sample	Grade 11 students in 2013/14 (analytic sample only)	Yes
1b. Same individuals as in the analytic sample (multiyear gap)	Grade 9 students in 2011/12 (analytic sample only)	Yes
2a. Same cohort, prior grade	Grade 11 students in 2013/14	Yes
2b. Same cohort, prior grade (multiyear gap)	Grade 9 students in 2011/12	No
3a. Same grade, prior cohort	Grade 12 students in 2013/14	Yes
3b. Same grade, prior cohort (multiyear gap)	Grade 12 students in 2012/13	No

Note: This table is for an example RCT that randomly assigned schools to conditions in summer 2014 and measured outcomes for grade 12 students at the end of 2014/15. Samples 3a and 3b are distinct from scenarios involving time confounds. For instance, a time confound is present if the intervention group was grade 12 students in 2014/15 and the comparison group was grade 12 students in 2013/14. Differences between these groups could be attributed to a naturally occurring change over time rather than an intervention effect. This time confound would yield a rating of *Does Not Meet WWC Standards* as detailed previously. In contrast, the scenarios in this table refer to where the outcome data for both intervention and comparison groups come from grade 12 students in 2014/15, eliminating that confound.

These examples illustrate the different types of acceptable baseline samples:

- Baseline samples that use the analytic sample of individuals to compute the cluster-level baseline means (samples 1a and 1b) are always acceptable, regardless of the preintervention period, because the individuals contributing outcome data and baseline data are the same.
- Baseline samples of the same cohort and prior grade (samples 2a and 2b) can differ from the analytic sample of individuals due to students who joined or exited the schools. A one-year gap in data collection between 2014 and 2015 (sample 2a) is acceptable because the data were obtained within one year before clusters were assigned to conditions. In contrast, baselines sample of the same cohort with a multiyear gap (sample 2b) is not acceptable due to the longer gap between baseline and outcome data collection. This longer gap weakens the degree of overlap with the outcome sample of individuals due to the longer time frame for joining and leaving clusters to occur.

- Baseline samples of the same grade and prior cohort have no overlap with the individuals in the outcome sample (except for students who repeated grade levels). Although the individuals do not overlap, this type of sample can be acceptable if it comes from the immediately preceding cohort (sample 3a), as the individuals in each cluster may be similar from one cohort to the next on average. In contrast, baseline samples with the same grade and prior cohort with a multiyear gap (sample 3b) is not acceptable because the longer gap weakens the degree to which cluster composition may be similar.

Acceptable baseline adjustments for studies satisfying the cluster-level baseline equivalence standard. Unlike step 4a, acceptable baseline adjustments for studies satisfying the cluster-level baseline equivalence standard do not need to account for the individual-level correlation between the baseline and outcome measure. Acceptable baseline adjustments for step 4b include those mentioned in Chapter III, Acceptable baseline adjustment strategies and apply only when the samples meet the other two requirements of step 4b: (a) the sample used to calculate the cluster-level means for adjustment was one of three acceptable baseline samples as defined earlier and (b) the outcome and baseline samples demonstrate representativeness of individuals in the clusters.

Demonstrating representativeness for studies satisfying the cluster-level baseline equivalence standard. Studies satisfying the cluster-level baseline equivalence standard must demonstrate that both the baseline data and outcome data are representative of individuals in clusters. This assessment therefore involves two separate calculations of representativeness, unlike a previous step that was only about outcome representativeness (step 3d).

- **Outcome representativeness:** The individuals contributing outcome data must be representative of the clusters at follow-up data collection, as defined earlier in step 3d in Chapter III, Compositional change. Although only cluster RCTs are encountered in step 3d, the same guidelines apply to both cluster RCTs and cluster QEDs when assessing cluster-level adjustments.
- **Baseline representativeness:** The individuals contributing baseline data also must be representative of the clusters at baseline data collection (following the same guidelines in step 3d). Consider a school-level QED that measures grade 5 student outcomes in 2015 and adjusts for cluster-level means of grade 5 students from the same schools in 2014. Baseline representativeness would be based on (a) the number of grade 5 students enrolled in the schools in 2014 who contribute baseline data versus (b) the total number of grade 5 students enrolled in 2014 (who did and did not contribute baseline data).

Both representativeness calculations should be based on only the nonattriting clusters that contributed at least one individual to the outcome analytic sample. Like the guidelines in step 3d, using administrative data can satisfy both representativeness requirements unless review team leadership concludes that patterns of missing administrative data have a high risk of differing across intervention versus comparison groups.

After assessing the assignment process, compositional changes, and baseline requirements for a study finding, the reviewer will determine the finding's research rating based on the strength of the research design and its implementation. For composite findings at the outcome domain level, the reviewer also will determine an effectiveness rating based on the evidence of the intervention's effects. How the WWC determines these ratings and synthesizes results is described in Chapter VII, Synthesis and Reporting of Results.

CHAPTER IV. REVIEWING FINDINGS FROM REGRESSION DISCONTINUITY DESIGNS

Regression discontinuity design (RDD) is a group design eligible for WWC review, which can have individual-level or cluster-level assignment to conditions like other group designs. Education researchers use an RDD when interventions are made available to individuals or groups based on how they compare with a cutoff value on some known measure. The groups, therefore, are not formed randomly–they are formed purposefully, by design, using a cutoff on a continuous "forcing" or "assignment" variable. In RDD studies, an intervention effect is estimated by comparing two regression lines: one that represents the relationship between the forcing variable and the outcome in the intervention group and similarly for the comparison group. The difference in those two regression lines, at the cutoff value of the forcing variable, represents the estimated intervention effect. The choice of the forcing variable and the specific cutoff used for intervention assignment is usually determined by policymakers, institution administrators, or study authors. For example, district administrators may assign students to a summer school program if they score below a cutoff value on a standardized test, or schools may be awarded a grant based on their score on a proposal. Depending on the extent to which findings from RDDs meet WWC standards, they can be rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*.

The process the WWC uses to review RDDs is the same as the process used for randomized controlled trials (RCTs) and quasi-experimental designs (QEDs), including screening studies for eligibility, assessing studies according to WWC standards, and reporting results. However, the mechanism of review differs, in some instances considerably, from the WWC's review of other group designs. For this reason, procedures and standards for reviewing RDDs are presented separately in the *Handbook*.

Screening RDD studies for eligibility

The eligibility criteria for WWC review of RDDs are the same as described in Chapter II, Screening Studies for Eligibility. That is, RDDs must be made publicly available and released within the 20 years preceding the review, use eligible populations, examine eligible interventions, and have eligible outcomes. In addition, to be reviewed as an RDD, a study must meet three criteria pertaining to the forcing variable summarized in table 13 and described below.

Table 13. *Criteria for the forcing variable in a regression discontinuity design study*

Eligibility criteria for the forcing variable in an RDD study	Eligible for review example of forcing variable	Ineligible for review example of forcing variable
1. Intervention assignments are based on a *numerical forcing variable.*	Study authors use students' scores on a math achievement test as the forcing variable, with possible scores ranging from 0 to 100.	Study authors use results from a pass/fail course as the forcing variable, with possible values being pass or fail.
2. The forcing variable has at least *four unique values* on either side of the cutoff.	Study authors use students' grade point average as the forcing variable.	Study authors use students' letter grade (A, B, C, D, F) as the forcing variable.

Continued on next page

Table 13. *Criteria for the forcing variable in a regression discontinuity design study (continued)*

Eligibility criteria for the forcing variable in an RDD study	Eligible for review example of forcing variable	Ineligible for review example of forcing variable
3. The forcing variable is the *actual* variable; it is neither a proxy nor an estimated forcing variable.	Study authors use students' math achievement scores or a linear transformation of scores as the forcing variable.	Study authors use students' predicted math achievement scores based on a regression model as the forcing variable, even though the actual forcing variable used for intervention assignment was students' actual math achievement scores.

1. **Intervention assignments are based on a *numerical forcing variable***. Units are assigned to the intervention or comparison group based on their values on a numerical forcing variable that has a unique ordering of values from low to high. Units with values above a cutoff value are assigned to one condition, and those with values below a cutoff value are assigned to the other condition.

   Example 1. An evaluation of a tutoring program could be classified as an RDD 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.

   Example 2. A study examining the impacts of grants to improve teacher training in local areas could be reviewed as an RDD if grants are awarded to only those sites with grant application scores of at least 70.

   In some instances, RDDs may use multiple criteria to assign participants to conditions. For example, a student may be assigned to an afterschool program if the student's reading score is below 30 or the student's math score is below 40. Studies that use multiple forcing variables or cutoffs with the same sample are eligible for review under these standards only if they use a method described in the literature to reduce those variables to a single forcing variable or analyze each forcing variable separately (for example, see Reardon & Robinson, 2012; Wong et al., 2013).

2. **The forcing variable has at least four unique values below and above the cutoff**. This is required for eligibility because at least eight data points are required to credibly select bandwidths or functional forms when modelling the relationship between the outcome and the forcing variable.

3. **The forcing variable used to calculate impacts must be the *actual* forcing variable, not a proxy or an estimate of the forcing variable**. A variable is a proxy or an estimate of the forcing variable if its correlation with the actual forcing variable is less than 1.

Reviewing findings from RDD studies according to WWC standards

If a study is eligible for WWC review as an RDD, the WWC will review the study by applying five RDD standards (table 14). Standards 1 through 4 apply to "sharp" and "fuzzy" RDDs (fuzzy RDDs are RDDs where some units do not participate in the condition to which they were assigned, referred to as noncompliance or crossover). Standard 5 only applies to fuzzy RDDs under certain circumstances. Findings from an RDD study can receive one of three ratings based on the set of criteria described below and summarized in table 14.

- *Meets WWC Standards Without Reservations*. To qualify, findings from an RDD study must completely satisfy each of the five standards listed in table 14.

- *Meets WWC Standards With Reservations*. To qualify, findings from an RDD study must at least partially satisfy standards 1 through 4 and, when applicable, standard 5.

  Note: If findings from an RDD study fail to at least partially satisfy RDD standards, review teams have an option of reviewing findings under QED standards for the findings to be eligible to receive the rating *Meets WWC Standards With Reservations*. This option is not available for fuzzy RDDs.

- *Does Not Meet WWC Standards*. Findings from an RDD study will receive this rating if they fail to at least partially satisfy RDD standards or if they fail to satisfy QED standards when review team decide to pursue a QED route.

Table 14. *Ratings for findings from regression discontinuity design studies*

Standard	To be eligible to receive the rating *Meets WWC Standards Without Reservations*, findings from the RDD study must:	To be eligible to receive the rating *Meets WWC Standards With Reservations*, findings from the RDD study must:
1. Integrity of the forcing variable	Completely satisfy this standard.	Partially satisfy this standard.
2. Sample attrition and baseline equivalence	Completely satisfy this standard.	Partially satisfy this standard.
3. Continuity of the relationship between the outcome and the forcing variable.	Completely satisfy this standard.	Partially satisfy this standard.
4. Functional form and bandwidth	Completely satisfy this standard.	Partially satisfy this standard.
5. Fuzzy RDD (if applicable)	Completely satisfy this standard.	Partially satisfy this standard.

Like RCTs and QEDs, RDDs also must meet the WWC outcome measure standards and be free of confounds. Procedures and standards for each step of reviewing RDDs according to WWC standards are described in the following sections.

Reviewing outcome measures and checking for confounding factors in RDD studies

The steps for reviewing outcome measures in RDD studies and checking RDD studies for confounding factors are similar to other group designs reviewed by the WWC.

Outcome measure standards

An RDD study should have at least one finding measured using a measure that meets the WWC's outcome measure standards: (1) face validity, (2) reliability, (3) not overaligned with the intervention, and (4) consistent data collection procedures as described in Chapter III, Outcome measure standards.

Confounding factors

An RDD study should be free of confounding factors, including $N = 1$ confounds, participant characteristics aligned with only one condition, or time aligned with only one condition as described in Chapter III, Confounding factors.

When reviewing an RDD study for confounding factors, a reviewer should consider whether the cutoff value of the forcing variable represents a confound. This can happen when the cutoff value of the forcing variable is used to assign members of the study sample to intervention services other than those being studied, creating a confound.

> *Example*. If an RDD study uses the income cutoff that determines free or reduced-price lunch eligibility as the cutoff value on the forcing variable, that likely creates a study confound. This is because income cutoff could be used as an eligibility criterion for many additional supplementary services that could also affect outcomes. In such cases, the study will not be able to isolate the impact of the intervention being tested from the other interventions delivered based on this family income cutoff score.

Standard 1. Integrity of the forcing variable

For an RDD to produce unbiased estimates of intervention effects, study authors should demonstrate that there was no systematic manipulation of the forcing variable. In an RDD study, manipulation means that forcing variable scores for some participants were systematically changed from their true values to influence intervention assignments. Thus, the true forcing variable values are unknown. With nonrandom manipulation, the true relationship between the outcome and forcing variable can no longer be identified, which could lead to biased impact estimates.

When manipulation is unlikely

Manipulation is less likely to occur if the forcing variable is a standardized test score than if it is a student assessment conducted by teachers who also have input into intervention assignment decisions.

Manipulation is unlikely in cases where the researchers 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.

Manipulation of the forcing variable is possible if the participants or individuals who determine forcing variable scores have knowledge of the cutoff value and have incentives and an ability 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 intervention assignment processes are not independent. Manipulation of the forcing variable is different from intervention status noncompliance, which occurs if some intervention condition members do not receive intervention services, or some comparison condition members receive embargoed services. The likelihood of manipulation will depend on the nature of the forcing variable, the intervention, and the research design.

For RDDs, the *integrity of the forcing variable* should be established (a) institutionally, (b) statistically, and (c) graphically as shown in table 15 and described below. To be eligible for the research rating of *Meets WWC Standards Without Reservations*, findings from RDD studies must satisfy the three criteria that comprise Standard

1. To be eligible for the research rating of *Meets WWC Standards With Reservations*, findings from RDD studies must satisfy any two of the three criteria that comprise Standard 1.

Table 15. *Regression discontinuity design criteria for Standard 1: Integrity of the forcing variable*

Criterion	To completely satisfy the standard, findings from the RDD study:	To partially satisfy the standard, findings from the RDD study:
A. The institutional integrity of the forcing variable must be established by an adequate description of the scoring and intervention assignment process.	Must satisfy this criterion.	Must satisfy any two of the three criteria (A, B, or C).
B. The statistical integrity of the forcing variable must be demonstrated by using statistical tests found in the literature (for example, McCrary, 2008) to establish the smoothness of the density of the forcing variable right around the cutoff.	Must satisfy this criterion.	
C. The graphical integrity of the forcing variable must be demonstrated by using a graphical analysis, such as a histogram or other type of density plot, to establish the smoothness of the density of the forcing variable right around the cutoff.	Must satisfy this criterion.	

Standard 1 for the review of RDDs includes the following three criteria:

- **Criterion A.** ***The institutional integrity of the forcing variable must be established by an adequate description of the scoring and intervention assignment process.*** This description will generally include information about the forcing variable used; the cutoff value selected; who selected the cutoff (for example, researchers, school personnel, or 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 and little incentive to change "true" obtained scores to allow or deny specific units access to the intervention. The study will not satisfy this standard if there is a clear opportunity to manipulate scores AND a clear incentive.
- **Criterion B.** ***The statistical integrity of the forcing variable must be demonstrated by using statistical tests found in the literature (for example, McCrary, 2008) 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 intervention group, in which case, there may be an unusual mass of units near the cutoff. The statistical test must fail to reject the null hypothesis of continuity in the density of the forcing variable at the 5 percent significance level.
- **Criterion C.** ***The graphical integrity of the forcing variable must be demonstrated by using a graphical analysis, such as a histogram or other type of density plot, to establish the smoothness of the density of the forcing variable right around the cutoff.*** There must not be strong evidence of a discontinuity at the

cutoff that is obviously larger than discontinuities in the density at other points along the forcing variable distribution, although some small discontinuities may arise when the forcing variable is discrete.

Standard 2. Sample attrition and baseline equivalence

An RDD must have acceptable levels of overall and differential attrition rates or else establish the equivalence of the groups being compared at the cutoff value of the forcing variable. The acceptable levels and default and allowable attrition boundaries are the same as used for an RCT and are described in Chapter III, Compositional change, and appendix C in the technical appendices. As with other group designs, review teams must choose between the cautious or optimistic attrition boundary to review RDD studies and must document their reasoning. The samples used to calculate attrition in an RDD study **must include all participants who were eligible** to be assigned to the intervention or comparison group using the forcing variable, and not only a subset of those participants known to the researcher. Therefore, attrition in an RDD study cannot be assessed unless all participants who were eligible to be assigned to conditions are known and their assigned conditions are known.

> *Example.* When age is used to assign students to a prekindergarten program, the assignment mechanism applies to all students in a defined geographical region, such as a state or district, and at a specified time, such as when a law was passed or in the fall of a certain school year. A study conducted with students enrolled in the state's schools several years after assignment would not meet this requirement because the intervention could have affected whether students remained in the state.

In an RDD study, attrition can be assessed within exogenous subgroups, meaning a subgroup identified using a variable that is not related to intervention participation. For example, attrition could be assessed separately within each site. Or, attrition could be calculated only using data within a bandwidth.

The levels of overall and differential attrition and the way that the levels are calculated determines whether an RDD study satisfies the attrition standard completely or partially as shown in table 16 and described below. To completely satisfy the attrition standard and to be eligible for the research rating of *Meets WWC Standards Without Reservations*, findings from RDD studies must have a low combination of overall and differential attrition using an approach that has the potential to adjust for the forcing variable most accurately (discussed below in the description of Criterion A). To partially satisfy the attrition standard and to be eligible for the research rating of *Meets WWC Standards With Reservations*, findings from RDD studies must either have a low combination of overall and differential attrition using an approach which may not provide as accurate an adjustment for the forcing variable, or must demonstrate baseline equivalence on key covariates, as described below.

Table 16. *Regression discontinuity design criteria for Standard 2: Sample attrition and baseline equivalence*

Criterion	To completely satisfy the standard, findings from the RDD study:	To partially satisfy the standard, findings from the RDD study:
A. The reported combination of overall and differential attrition rates is low using an approach among those that have the potential to most accurately adjust for the forcing variable.	Must satisfy this criterion.	Do not need to satisfy this criterion.

Continued on next page

Table 16. *Regression discontinuity design criteria for Standard 2: Sample attrition and baseline equivalence (continued)*

Criterion	To completely satisfy the standard, findings from the RDD study:	To partially satisfy the standard, findings from the RDD study:
B. The reported combination of overall and differential attrition rates is low when calculated using an approach among those that may not provide as accurate an adjustment for the forcing variable.	Do not need to satisfy this criterion.	Must satisfy one of the two criteria (B or C).
C. The study demonstrates baseline equivalence on key covariates.		

Standard 2 for the review of RDDs includes the following three criteria:

- **Criterion A.** ***The reported combination of overall and differential attrition rates is low using an approach among those that have the potential to most accurately adjust for the forcing variable:***
    - Study authors must report the predicted mean attrition rate at the cutoff estimated using data from below the cutoff and the predicted mean attrition rate at the cutoff estimated using data from above the cutoff. Both numbers must be estimated using a statistical model that controls for the forcing variable using the same approach that was used to estimate the impact on the outcome. Specifically, the impact on attrition must be estimated either (A) using the same bandwidth and/or functional form as was used to estimate the impact on the outcome or (B) using the same algorithm for selecting the bandwidth and/or functional form as was used to estimate the impact on the outcome. For example, if study authors used the Imbens & Kalyanaraman (2012) approach for selecting the bandwidth to estimate impact, they would use the same approach for selecting the bandwidth to estimate mean attrition rates above and below the cutoff. For applying this standard, the overall attrition rate will be defined as the average of the predicted mean attrition rates on either side of the cutoff, and the differential attrition rate will be defined as the difference in the predicted mean attrition rates on either side of the cutoff.
    - Study authors must calculate overall and differential attrition for the sample inside the bandwidth used for the impact analysis, with or without adjusting for the forcing variable. For example, if study authors used the bandwidth of 30 points around the cutoff to estimate impact, they would use the same bandwidth of 30 points around the cutoff to estimate overall and differential attrition. Although authors do not need to adjust for the forcing variable using this approach, other than by applying the bandwidth, the value of the forcing variable must be known for all subjects so that the bandwidth can be applied.
- **Criterion B.** ***The reported combination of overall and differential attrition rates must be low using one of the following approaches, which may not provide as accurate an adjustment for the forcing variable compared to one of the two approaches outlined under Criterion A:***
    - Study authors can calculate overall and differential attrition for the entire research sample, adjusting for the forcing variable.

- Study authors can calculate overall and differential attrition for the entire research sample without adjusting for the forcing variable.

If authors calculate overall and differential attrition both ways—that is, both with and without adjusting for the forcing variable—the WWC will review both and assign the highest possible rating to this part of the study design. However, only one of the two approaches may be used to calculate overall AND differential attrition. For example, if the overall attrition rate is calculated using the approach that does not adjust for the forcing variable, the differential attrition rate must be calculated using the approach that does not adjust for the forcing variable.

- **Criterion C.** ***The study demonstrates baseline equivalence on key covariates.*** The WWC will only assess baseline equivalence for the RDD findings that did not satisfy either Criterion A or Criterion B. These are the findings with high or unknown attrition. As with RCTs, the requirements for satisfying the baseline equivalence criterion of this standard depend on the attrition boundary used to assess attrition (see Chapter III, Determining optimistic or cautious attrition boundary when assessing attrition and figure 5).
  - *If the cautious attrition boundary is used to assess attrition*, satisfying Criterion C requires demonstrating that groups are similar, on key baseline covariates specified in the *Study Review Protocol,* at the cutoff, and applying the appropriate statistical adjustments. For an RDD, estimating baseline group differences involves calculating an impact at the cutoff on the covariate of interest, and study authors must either (1) use the same bandwidth and/or functional form as was used to estimate the impact on the outcome or (2) use the same algorithm for selecting the bandwidth and/or functional form as was used to estimate the impact on the outcome. Study authors may exclude sample members from this analysis for reasons that are clearly exogenous to intervention participation. For example, authors may calculate baseline equivalence using only data within the bandwidth that was used to estimate the impact on the outcome. The burden of proof falls on the authors to demonstrate that any sample exclusions were made for exogenous reasons.

    The baseline equivalence standard for other group designs applies to the results from this analysis (see Chapter III, Baseline equivalence standard). Specifically, if the impact for any covariate is greater than 0.25 standard deviation in absolute value, based on the variation of that characteristic in the pooled sample, this criterion is not satisfied. If the impact for a covariate is between 0.05 standard deviation and 0.25 standard deviation, the statistical model used to estimate the average intervention effect on the outcome must include a statistical adjustment for that covariate to satisfy this criterion. Differences of less than or equal to 0.05 standard deviation require no statistical adjustment.

    For dichotomous covariates, authors must provide the predicted mean covariate value—that is, the predicted probability—at the cutoff estimated using data from below the cutoff and the predicted probability at the cutoff estimated using data from above the cutoff.

    Both predicted probabilities must be calculated using the same statistical model that is used to estimate the impact on the covariate at the cutoff. These predicted probabilities are needed so that WWC reviewers can transform the impact estimate into standard deviation units.

This analysis must be conducted using only subjects with nonmissing values of the key outcome variable used in the study.

- *If the optimistic attrition boundary is used to assess attrition*, study authors must use an appropriate statistical adjustment for key covariates specified in the *Study Review Protocol*. Study authors do not need to demonstrate that groups differ by no more than 0.25 standard deviation on key covariates.

Standard 3. Continuity of the relationship between the outcome and the forcing variable

A well-executed RDD must demonstrate that, in the absence of the intervention, there is a smooth relationship between the outcome and the forcing variable at the cutoff score. This condition is known as continuity of the relationship between the outcome and the forcing variable at the cutoff score. This condition is necessary to ensure that any observed discontinuity in the outcomes of intervention and comparison group participants at the cutoff can be attributed to the intervention.

The continuity of the relationship between the outcome and the forcing variable cannot be checked directly. Therefore, the WWC uses indirect criteria to determine whether a study satisfies the continuity requirements, as shown in table 17 and described below.

Table 17. *Regression discontinuity design criteria for Standard 3: Continuity of the relationship between the outcome and the forcing variable*

Criterion	To completely satisfy the standard, findings from the DD study:	To partially satisfy the standard, findings from the RDD study:
A. There is no evidence, using graphical analyses, of a discontinuity in the outcome-forcing variable relationship at values of the forcing variable other than the cutoff value, unless a satisfactory explanation of such a discontinuity is provided.	Must satisfy this criterion.	Must satisfy one of the two criteria (A or B).
B. There is no evidence, using statistical tests, of a discontinuity in the outcome-forcing variable relationship at values of the forcing variable other than the cutoff value, unless a satisfactory explanation of such a discontinuity is provided.	Must satisfy this criterion.	

Standard 3 for RDDs includes the following two criteria:

- **Criterion A.** ***There is no evidence, using graphical analyses, of a discontinuity in the outcome–forcing variable relationship at values of the forcing variable other than the cutoff value, unless a satisfactory explanation of such a discontinuity is provided.*** An example of a "satisfactory explanation" is that the discontinuity corresponds to some other known intervention that was also administered using the same forcing variable but with a different cutoff value. Another example could be a known structural property of the assignment variable (for example, if the assignment variable is a construct involving the aggregation of both continuous and discrete components). The graphical analysis–such as a scatter plot of the outcome and forcing variable using either the raw data or

averaged/aggregated data within bins/intervals—must not show a discontinuity at any forcing variable value within the bandwidth (or, for the full sample if no bandwidth is used) that is larger than two times the standard error of the impact estimated at the cutoff value, unless a satisfactory explanation of that discontinuity is provided. (The standard error at the cutoff value is used because authors may not report the standard error at the point of the observed discontinuity.)

- **Criterion B.** ***There is no evidence, using statistical tests, of a discontinuity in the outcome-forcing variable relationship at values of the forcing variable other than the cutoff value, unless a satisfactory explanation of such a discontinuity is provided.*** The statistical tests must use the same algorithm for selecting the bandwidth and/or functional form as was used to estimate the impact on the outcome and be conducted for at least four values of the forcing variable below the cutoff and four values above the cutoff; these values can be either within or outside the bandwidth. At least 95 percent of the estimated impacts on the outcome at other values of the forcing variable must be statistically insignificant at the 5 percent significance level. For example, if impacts are estimated for 20 values of the forcing variable, then at least 19 of them must be statistically insignificant.[14]

Standard 4. Functional form and bandwidth

Statistical modeling plays a central role in estimating impacts in RDD studies. A reviewer needs to examine two components of the statistical modeling in an RDD:

- **The relationship (or functional form) between the outcome variable and the forcing variable.**
- **Bandwidth** around the cutoff (the range of forcing variable values used to select the analytic sample).

The WWC has established six criteria under standard 4 to examine functional form and bandwidth in RDDs as shown in table 18 and described below. To completely satisfy standard 4 and be eligible for the research rating of *Meets WWC Standards Without Reservations,* findings from RDD studies must satisfy criteria A, B, D, E, and F. To partially satisfy standard 4 and be eligible for the research rating of *Meets WWC Standards With Reservations,* findings from RDD studies must satisfy criteria A, either B or C, and E.

Table 18. *Regression discontinuity design criteria for Standard 4: Functional form and bandwidth*

Criterion	To completely satisfy the standard, findings from the RDD study:	To partially satisfy the standard, findings from the RDD study:
A. The local average treatment effect for an outcome is estimated using a statistical model that controls for the forcing variable.	Must satisfy this criterion.	Must satisfy this criterion.

Continued on next page

[14] If impacts are estimated for fewer than 20 values of the forcing variable, then all of them must be statistically insignificant at the 5 percent significance level.

Table 18. *Regression discontinuity design criteria for Standard 4: Functional form and bandwidth (continued)*

Criterion	To completely satisfy the standard, findings from the RDD study:	To partially satisfy the standard, findings from the RDD study:
B. The study uses a local regression, either linear or quadratic, or related nonparametric approach in which impacts are estimated within a justified bandwidth, meaning a bandwidth selected using a systematic procedure that is described and supported in the methodological literature, such as cross-validation.	Must satisfy this criterion.	Must satisfy one of the two criteria (B or C).
C. If the study does not use a local regression or related nonparametric approach or uses such an approach but not within a justified bandwidth, then it may estimate impacts using a "best fit" regression using either the full sample or the sample within a bandwidth; the bandwidth does not need to be justified.	Does not need to satisfy this criterion.	
D. The study provides evidence that the findings are robust to varying bandwidth or functional form choices.	Must satisfy this criterion.	Does not need to satisfy this criterion.
E. The study presents a graphical analysis displaying the relationship between the outcome and forcing variable, including a scatter plot—using either the raw data or averaged/aggregated data within bins/intervals—and a fitted curve.	Must satisfy this criterion.	Must satisfy this criterion.
F. The relationship between the forcing variable and the outcome is not constrained to be the same on both sides of the cutoff.	Must satisfy this criterion.	Do not need to satisfy this criterion.

Standard 4 for the review of RDDs includes the following six criteria:

- **Criterion A.** ***The local average treatment effect for an outcome is estimated using a statistical model that controls for the forcing variable.*** For both bias and variance considerations, it is never acceptable to estimate an impact by comparing the mean outcomes of intervention 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.** ***The study uses a local regression, either linear or quadratic, or related nonparametric approach in which impacts are estimated within a justified bandwidth, meaning a bandwidth selected using a systematic procedure that is described and supported in the methodological literature, such as cross-validation.*** For example, a bandwidth selection procedure described in an article published in a peer-reviewed journal that describes the procedure and demonstrates its effectiveness would be a justified bandwidth. An article published in an applied journal where the procedure happens to be used does not count as justification. A study that does not use a justified bandwidth does not completely satisfy this standard but could partially satisfy this standard if Criterion C is satisfied.

- **Criterion C.** ***If the study does not use a local regression or related nonparametric approach or uses such an approach but not within a justified bandwidth, then it may estimate impacts using a "best fit" regression using either the full sample or the sample within a bandwidth; the bandwidth does not need to be justified.*** For an impact estimate to satisfy this criterion, the functional form of the relationship between the outcome and forcing variable must be shown to be a better fit to the data than at least two other functional forms. Any measure of goodness of fit from the methodological literature can be used, such as the Akaike Information Criterion or adjusted R^2.
- **Criterion D.** ***The study provides evidence that the findings are robust to varying bandwidth or functional form choices.*** At least one of five types of evidence is sufficient to satisfy this criterion[15]:
    - In the case that Criterion B applies, the sign and significance of impact estimates must be the same for a total of at least two different justified bandwidths. For example, this criterion would be satisfied if the sign and significance of an impact are the same using a bandwidth selected by cross-validation[16] and a bandwidth selected by the method described in Imbens and Kalyanaraman (2012). Two impact estimates are considered to have the same significance if they are both statistically significant at the 5 percent significance level, or if neither of them is statistically significant at the 5 percent significance level. Two impact estimates are considered to have the same sign if they are both positive, both negative, or if one is positive and one is negative, but neither are statistically significant at the 5 percent significance level.
    - In the case that Criterion B applies, the sign and significance of impact estimates must be the same for at least one justified bandwidth and at least two additional bandwidths that are not justified.
    - In the case that Criterion C applies, the sign and significance of impact estimates must be the same using a total of at least two different goodness-of-fit measures to select functional form. For example, this criterion would be satisfied if the impact corresponding to the functional form selected using the Akaike Information Criterion is the same sign and significance as an impact corresponding to the functional form selected using the regression R^2. Note that both measures may select the same functional form.
    - In the case that Criterion C applies, the sign and significance of impact estimates must be the same for at least three different functional forms, including the "best fit" regression.
    - If the study meets both Criteria B and C, then the sign and significance of impact estimates must be the same for the impact estimated within a justified bandwidth and the impact estimated using a "best fit" regression.
- **Criterion E.** ***The study presents a graphical analysis displaying the relationship between the outcome and forcing variable, including a scatter plot—using either the raw data or averaged/aggregated data within bins/intervals—and a fitted curve.*** The display cannot be obviously inconsistent with the choice of bandwidth and the functional form specification for the analysis. Specifically, if the study uses a particular functional form for the outcome-forcing variable relationship, then the study must show graphically that this functional form fits the scatter plot reasonably well, and if the study uses a local linear regression, then the

[15] If a study presents more than one type of evidence, and one type shows findings are robust while another type does not, then this criterion is still satisfied. That is, studies are not penalized for conducting more sensitivity analyses

[16] An implementation of cross-validation for RDD analysis is described by Imbens and Lemieux (2008).

scatter plot must show that the outcome-forcing variable relationship is indeed reasonably linear within the chosen bandwidth.

- **Criterion F.** ***The relationship between the forcing variable and the outcome is not constrained to be the same on both sides of the cutoff.*** To satisfy this criterion, study authors must allow the relationship between the outcome and the forcing variable to be different on either side of the cutoff. This is because it is reasonable to suspect that the relationship may be different on either side of the cutoff when an intervention has an impact.

Standard 5. Fuzzy RDD

In a sharp RDD, all intervention group members receive intervention services, and no comparison group members receive services. In a fuzzy RDD, some intervention group members do not receive intervention services, or some comparison group members do receive intervention services, but there is still a substantial discontinuity in the probability of receiving services at the cutoff. In a fuzzy RDD analysis, the impact of service receipt is calculated as a ratio. The numerator of the ratio is the RDD impact on an outcome of interest. The denominator is the RDD impact on the probability of receiving services. This analysis is typically conducted using two-stage least-squares regression analysis.[17] Fuzzy RDD analysis is analogous to a complier average causal effect (CACE) or local average treatment effect analysis. Consequently, many aspects of this standard are analogous to WWC standards for CACE analysis in the context of RCTs described in Chapter V.

Reviewing a fuzzy RDD study requires evaluating the eligible study findings according to all standards discussed so far: integrity of the forcing variable (Standard 1), sample attrition (Standard 2), continuity of the relationship (Standard 3), and functional form and bandwidth (Standard 4). If a study finding satisfies all four of these RDD standards, then a reviewer needs to evaluate it using the fuzzy RDD standard (Standard 5).

The internal validity of a fuzzy RDD estimate depends primarily on the following three conditions:

1. The first condition is the exclusion restriction, which requires that the only channel through which assignment to the intervention or comparison groups can influence outcomes is by affecting take-up of the intervention being studied (Angrist et al., 1996). When this condition does not hold, group differences in outcomes would be attributed to the effects of taking up the intervention when they may instead be attributable to other factors differing between the intervention and comparison groups. The exclusion restriction cannot be completely verified, as it is impossible to determine whether the effects of assignment on outcomes are mediated through unobserved channels. However, it is possible to identify clear violations of the exclusion restriction—in particular, situations in which groups face different circumstances beyond their differing take-up of the intervention of interest.
2. The second condition is that the discontinuity in the probability of receiving services at the cutoff needs to be large enough to limit the influence of finite sample bias. The fuzzy RDD scenario can be interpreted as an instrumental variables model in which falling above or below the cutoff is an instrument for receiving intervention services (the participation indicator). Instrumental variables estimators will be subject to finite-

[17] The WWC also allows fuzzy RDD effects to be computed using a Wald estimator. The procedure for doing this is defined in the CACE section of appendix F.

sample bias if there is not a substantial difference in service receipt on either side of the cutoff, that is, if the instrument is "weak" (Stock & Yogo, 2005). Fuzzy RDD impacts need not be estimated using two-stage least-squares regression methods. For example, they can be estimated using Wald estimators. However, study authors must run the first-stage regression of the participation indicator on the forcing variable and the indicator for being above or below the cutoff and provide either the *F* statistic or the *t* statistic from this regression.

3. The third condition for the internal validity of a fuzzy RDD estimate is that two relationships need to be modeled appropriately: the relationship between the forcing variable and the outcome of interest, and the relationship between the forcing variable and receipt of services. Ideally, the fuzzy RDD impact would be estimated using a justified bandwidth and functional form, where justification is focused on the overall fuzzy RDD impact, not just the numerator or denominator separately. Several methods have been discussed in the literature for selecting a justified bandwidth that targets the ratio, which satisfy the WWC requirements (such as Calonico et al., 2014; Imbens & Kalyanaraman, 2012). However, in practice, study authors often use the bandwidth for the numerator of the fuzzy RDD, which is consistent with advice from Imbens and Kalyanaraman (2012).[18]

The criteria shown in table 19 operationalize the three conditions above, determining whether a study finding completely or partially satisfies the WWC's standard for fuzzy RDDs. To completely satisfy Standard 5 and be eligible for the research rating of *Meets WWC Standards Without Reservations,* findings from a fuzzy RDD must satisfy Criteria A through G. To partially satisfy standard 5 and be eligible for the research rating of *Meets WWC Standards With Reservations,* findings from a fuzzy RDD must satisfy Criteria A through F, and Criterion H.

All fuzzy RDD criteria are waived for studies that calculate impact estimates using a reduced-form model (where the outcome is modeled as a function of the forcing variable, an indicator for being above or below the cutoff, and possibly other covariates, but the participation indicator is not included in the model). This model is analogous to an intent-to-treat analysis in the context of RCTs. If a reduced-form model is used to estimate the fuzzy RDD impact, the study finding must meet the other four RDD standards to be eligible to receive the rating *Meets WWC Standards Without Reservations*.

Table 19. *Regression discontinuity design criteria for evaluating fuzzy regression discontinuity designs*

Criterion	To completely satisfy the standard, findings from the RDD study:	To partially satisfy the standard, findings from the RDD study:
A. The participation indicator must be a binary indicator.	Must satisfy this criterion.	Must satisfy this criterion.
B. The estimation model must have exactly one participation indicator.	Must satisfy this criterion.	Must satisfy this criterion.

Continued on next page

[18] Imbens and Kalyanaraman (2012, p. 14) wrote, "In practice, this often leads to bandwidth choices similar to those based on the optimal bandwidth for estimation of only the numerator of the RD estimate. One may therefore simply wish to use the basic algorithm ignoring the fact that the regression discontinuity design is fuzzy."

Table 19. *Regression discontinuity design criteria for evaluating fuzzy regression discontinuity designs (continued)*

Criterion	To completely satisfy the standard, findings from the RDD study:	To partially satisfy the standard, findings from the RDD study:
C. The indicator for being above or below the cutoff must be a binary indicator for the groups to which subjects are assigned.	Must satisfy this criterion.	Must satisfy this criterion.
D. The same covariates must be included in (a) the analysis that estimates the impact on participation and (b) the analysis that estimates the impact on outcomes.	Must satisfy this criterion.	Must satisfy this criterion.
E. The fuzzy RDD estimate must have no clear violations of the exclusion restriction.	Must satisfy this criterion.	Must satisfy this criterion.
F. The study must provide evidence that the forcing variable is a strong predictor of participation in the intervention.	Must satisfy this criterion.	Must satisfy this criterion.
G. The study must use a local regression or related nonparametric approach with a justified bandwidth.	Must satisfy this criterion.	Do not need to satisfy this criterion.
H. If Criterion G is not met, the fuzzy RDD impact can be estimated using a bandwidth that is only justified for the numerator, even if it is larger than a bandwidth justified for the denominator OR the denominator is estimated using "best fit" functional form.	Do not need to satisfy this criterion.	Must satisfy this criterion.

Standard 5 for the review of fuzzy RDDs includes the following eight criteria:

- **Criterion A.** ***The participation indicator must be a binary indicator.*** For example, the participation indicator could be a binary indicator for receiving any positive dosage of the intervention.
- **Criterion B.** ***The estimation model must have exactly one participation indicator.***
- **Criterion C.** ***The indicator for being above or below the cutoff must be a binary indicator for the groups to which subjects are assigned.***
- **Criterion D.** ***The same covariates must be included in (a) the analysis that estimates the impact on participation and (b) the analysis that estimates the impact on outcomes.*** In the case of two-stage least squares estimation, this means that the same covariates must be used in the first and second stages.
- **Criterion E.** ***The fuzzy RDD estimate must have no clear violations of the exclusion restriction.*** Defining participation inconsistently between the assigned intervention and assigned comparison groups would constitute a clear violation of the exclusion restriction. Therefore, the study must report a definition of take-up that is the same across assigned groups. Another violation of the exclusion restriction is the scenario in which assignment to the intervention group changes the behavior of subjects even if they do not take up the intervention itself. In this case, the intervention assignment might have effects on outcomes though channels other than the take-up rate. There must be no clear evidence that assignment to the intervention influenced the outcomes of subjects through channels other than take-up of the intervention.
- **Criterion F.** ***The study must provide evidence that the forcing variable is a strong predictor of participation in the intervention.*** In a regression of program participation on an intervention indicator and

other covariates, the coefficient on the intervention indicator must report a minimum *F* statistic of 16 or a minimum *t* statistic of 4 (Stock & Yogo, 2005).[19] For fuzzy RDD studies with more than one indicator for being above or below the cutoff, see the WWC Group Design Standards for RCTs that report CACE estimates for the minimum required first-stage *F* statistic.

- **Criterion G.** ***The study must use a local regression or related nonparametric with a justified bandwidth.*** Bandwidths must be selected using a systematic procedure that is described and supported in the methodological literature. Ideally, this method would be justified for the fuzzy RDD impact estimate, not just the numerator of the fuzzy RDD estimate. However, two other approaches are acceptable. First, it is acceptable to use separate bandwidths for the numerator and denominator, if both are selected using a justified approach, such as the Imbens and Kalyanaraman (2012) algorithm applied separately to the numerator and denominator. Second, it is acceptable to use the bandwidth selected for the numerator if that bandwidth is smaller than or equal to a justified bandwidth selected for the denominator.
- **Criterion H.** ***If Criterion G is not met, the fuzzy RDD impact can be estimated using a bandwidth that is only justified for the numerator, even if it is larger than a bandwidth justified for the denominator OR the denominator is estimated using "best fit" functional form.*** That is, the functional form of the relationship between program receipt and the forcing variable must be shown to be a better fit to the data than at least two other functional forms. Any measure of goodness of fit from the methodological literature can be used, such as the Akaike Information Criterion or adjusted R^2.

Applying RDD standards to studies that involve aggregate or pooled impacts

Some RDD studies may report pooled or aggregate impacts for some combinations of forcing variables, cutoffs, and samples.

- **Pooled impacts** are based on data from each combination of forcing variable, cutoff, and sample that are standardized and grouped into a single dataset for which a single impact is calculated.

  Example. Consider a study conducted at five sites. Pooled impacts would occur if study authors standardized and combined all data from those five sites into a single dataset, and then estimated a single intervention impact estimate using that dataset.

- **Aggregated impacts** are a weighted average[20] of impacts calculated separately for every combination of forcing variable, cutoff, and sample.

---

[19] The *F* statistic must be for the instrument only—not the *F* statistic for the entire first stage regression. If the unit of assignment does not equal the unit of analysis, then the *F* statistic or *t* statistic must account for clustering using an appropriate method (such as boot-strapping, hierarchical linear modeling [HLM], or the method proposed by Lee and Card, 2008). Also, in a working paper, Fier et al. (2016) suggested that in the fuzzy RDD context, the minimum first-stage *F* statistic that ensures asymptotic validity of a 5 percent two-sided test is much higher than would be required in a simple instrumental variables setting; specifically, they suggest 135. Until a published paper provides an *F* statistic cutoff that is appropriate for fuzzy RDD studies that use a justified bandwidth, the *F* statistic of 16 will be used as the interim criterion for assessing instrument strength.

[20] The WWC does not require a specific approach to weighting, and authors could choose to use unit weighting. The selection of the weighting approach by study authors will not affect a study's research rating.

Example. Consider a study conducted at five sites. Aggregate impacts would occur if study authors had five separate datasets for each of the five sites, estimated an intervention impact estimate separately in each of those five datasets, and then combined those five impact estimates using a weighted average.

The study's overall research rating will be the highest rated impact—including pooled and aggregate impacts—presented in the study. Study authors may improve the rating of a pooled or aggregate impact by excluding combinations of forcing variables, cutoffs, and samples that do not meet WWC standards for reasons that are clearly exogenous to intervention participation. For example, in a multisite study, a site that fails the institutional check for manipulation could be excluded from the aggregate impact, resulting in a higher rating for the aggregate impact. However, endogenous exclusions—those potentially influenced by the intervention—will not improve the rating of an aggregate impact because standards will be applied as if those exclusions were not made. For example, excluding sites that have a high differential attrition rate from an aggregate impact will not improve the rating of that impact because for the purpose of applying the attrition standard, the WWC will include those sites. It falls on study authors to demonstrate that any exclusions from the aggregate impact were made for exogenous reasons.

For pooled or aggregate impacts that are based on multiple forcing variables, cutoffs, or samples, additional guidance for applying the standards is provided next.

Standard 1: Integrity of the forcing variable

- **Criterion A.** ***If the institutional integrity of the forcing variable is not satisfied for any combination of forcing variable, cutoff, and sample that is included in a pooled or aggregate impact, then this criterion is not satisfied for that pooled or aggregate impact.*** However, it is permissible to exclude from a pooled or aggregate impact cases that do not satisfy this criterion. For example, if a pooled or aggregate impact is estimated using data from five sites, and the institutional integrity of the forcing variable is not satisfied in one of those five sites, then the pooled or aggregate impact does not satisfy this criterion. However, a pooled or aggregate impact estimated using data from only the four sites for which the institutional integrity of the forcing variable is satisfied would satisfy this criterion.
- **Criterion B.** ***For an aggregate or a pooled impact, this criterion is satisfied if it is satisfied for every unique combination of forcing variable, cutoff, and sample that contributes to the pooled or aggregate impact.*** In the case of a pooled impact, applying an appropriate statistical test to the pooled data also can satisfy this criterion. It is permissible to exclude from a pooled or aggregate impact cases that do not satisfy this criterion.
- **Criterion C.** ***For an aggregate or a pooled impact, this criterion is satisfied if it is satisfied for every unique combination of forcing variable, cutoff, and sample that contributes to the pooled or aggregate impact.*** In the case of a pooled impact, providing a single figure based on the pooled data also can satisfy this criterion. It is permissible to exclude from a pooled or aggregate impact cases that do not satisfy this criterion.

Standard 2: Sample attrition and baseline equivalence

In the case of a pooled impact, Criterion A and Criterion B described in the section on individual RDD findings can be applied directly if the authors calculate and report overall and differential attrition using the pooled

sample. Any sample excluded from calculating the pooled or aggregate impact for reasons of endogeneity—that is, because the sample was potentially influenced by the intervention—cannot be excluded from the attrition calculation.

In the case of an aggregate impact, Criterion A and Criterion B can be applied to the overall and differential attrition rates calculated as weighted averages of the overall and differential rates calculated for each unique combination of forcing variable, cutoff, and sample that contribute to the aggregate impact. Authors must calculate overall and differential attrition for each of those unique combinations in a way that is consistent with the standard described in the section on individual RDD findings, and the weights used in aggregation must be the same weights used to calculate the weighted impact being reviewed. The attrition standard described in the section on individual RDD findings is then applied to the combination of overall and differential attrition based on the weighted average.

In the case of a pooled impact, Criterion C can be applied as described in the section on individual RDD findings without modification. In the case of an aggregate impact, baseline equivalence can be established by applying the same aggregation approach to the impacts on baseline covariates as is used to aggregate impacts on outcomes.

Standard 3: Continuity of the relationship between the outcome and the forcing variable

- **Criterion A.** ***In the case of a pooled impact, this criterion can be applied as described in the section on individual RDD findings without modification.*** In the case of an aggregate impact, the requirements for this criterion must be applied cumulatively across all combinations of forcing variables, cutoffs, and samples. Specifically, there must not be evidence of a discontinuity larger than twice the standard error of the impact at any noncutoff value within the bandwidth of any forcing variable for any sample. This means that a graphical analysis must be presented for every combination of forcing variable, cutoff, and sample. In cases where impacts from disjointed—that is, nonoverlapping—samples are being aggregated, it is acceptable to exclude from the aggregate impact any impacts from samples that do not satisfy this criterion; such an exclusion is considered exogenous.
- **Criterion B.** ***In the case of a pooled impact, this criterion can be applied as described in the section on individual RDD findings without modification.*** In the case of an aggregate impact, the requirements for this criterion must be applied cumulatively across all combinations of forcing variables, cutoffs, and samples. That is, at least 95 percent of estimated impacts at values of the forcing variables other than the cutoffs, across all samples, must be statistically insignificant. In cases where impacts from nonoverlapping samples are being aggregated, it is acceptable to exclude from the aggregate impact any impacts from samples that do not satisfy this criterion; such an exclusion is considered exogenous.

Standard 4: Functional form and bandwidth

- ***In the case of a pooled impact,*** this standard can be applied as described in the section on individual RDD findings without modification.
- ***In the case of an aggregate impact,*** Criteria A, B, C, E, and F of this standard must be applied to every impact included in the aggregate**.** Any impacts excluded from the aggregate because they do not satisfy one

of those criteria will be treated as attrition. The aggregate impact will receive the lowest rating among all impacts.

- ***Criterion D can be applied only to the aggregate impact.*** That is, it is not necessary to show robustness of every impact included in the aggregate (although doing so is acceptable).

Standard 5: Fuzzy regression discontinuity design

- ***In the case of a pooled impact,*** this standard can be applied as described in the section on individual RDD findings without modification.
- ***In the case of an aggregate impact***, this standard must be applied to every impact included in the aggregate. Any impacts excluded from the aggregate will be treated as attrition, with two exceptions: impacts may be excluded if they do not meet Criterion E or Criterion F. The aggregate impact will receive the lowest rating among all impacts.

Reviewing findings from cluster-assignment RDDs according to WWC standards

Special considerations apply when reviewing cluster RDDs, as detailed in the following guidance. Findings from cluster-assignment RDDs are eligible for the research rating *Meets WWC Standards Without Reservations* if the findings completely satisfy all RDD standards using individual-level or cluster-level data. Findings from cluster-assignment RDDs are eligible for the research rating *Meets WWC Standards With Reservations* if the findings partially satisfy sufficient RDD standards using individual-level or cluster-level data.

Standards 1 (integrity of the forcing variable), 3 (continuity of the relationship), 4 (functional form and bandwidth), and 5 (fuzzy RDD)

Standards 1, 3, 4, and 5 for findings from cluster-assignment RDDs are assessed using the same criteria as described in the previous sections on reviewing RDDs according to WWC standards, using individual-level or cluster-level data. For example, if schools are assigned to conditions and the study estimates the impact of the intervention by examining student standardized test data averaged to the school level, then criteria B and C of Standard 1 (integrity of the forcing variable) could be assessed using school-level data or student-level data (the assessment of Criterion A does not rely on study data).

Standard 2 (sample attrition)

Findings from cluster RDDs can satisfy the attrition standard by meeting the three requirements.

1. Meet the requirements for completely or partially satisfying Standard 2 described in Chapter IV, Sample attrition and baseline equivalence for individual RDD findings.
2. Demonstrate low risk of bias from individuals who entered clusters after assignment (joiners), described in Chapter III, Compositional change in cluster RCTs (step 3b).
3. Meet the requirements for completely satisfying Standard 2 using individual-level data within nonattriting clusters, applying an acceptable reference sample as outlined in Chapter III, Compositional change in cluster RCTs (step 3c).

The attrition boundaries for cluster RDDs are the same as for cluster RCTs, as described in appendix C.

CHAPTER V. REVIEWING COMPLIER AVERAGE CAUSAL EFFECT ESTIMATES AND FINDINGS USING OTHER ADVANCED ANALYTICAL APPROACHES

This section describes the What Works Clearinghouse's (WWC's) procedures and standards for reviewing a research design that is less common: randomized controlled trials (RCTs) that estimate complier average causal effects (CACEs). Studies that use CACEs must meet the eligibility requirements presented in Chapter II, Screening Studies for Eligibility, have at least one finding that meets WWC outcome standards, and be free of confounds. The consequences for failing to meet these requirements are the same as they are for more common designs such as RCTs, quasi-experimental designs (QEDs), and regression discontinuity designs (RDDs). However, some review procedures for these studies are different from the procedures for other group designs. The following sections outline these procedures. After the sections on CACEs is the section on advanced analytical approaches, such as how the WWC reviews studies with missing data.

Key terms

Endogenous independent variable: The variable whose impact on outcomes is the impact of interest. In this context, the endogenous independent variable is a binary indicator for taking up the intervention. It is endogenous because its variation could be affected by subjects' decisions. A particularly uninterested member of the intervention group might elect not to participate, and the unobserved factors underlying the decision might also be correlated with outcomes, inducing a correlation between take-up and outcomes that is not reflective of a causal effect of the intervention itself.

Structural equation: An equation that models the outcome as a function of the endogenous independent variable and possibly other covariates. In this context, estimation of the structural equation produces an estimate of the CACE–the impact of intervention take-up on outcomes.

Instrumental variables: Variables that strongly influence take-up of the intervention but are assumed to be uncorrelated with other factors influencing the outcome variable. By definition, instrumental variables are excluded from the structural equation. In this context, the instrumental variables are binary indicators for the groups to which subjects were randomly assigned.

First-stage equation: An equation that models the endogenous independent variable as a function of the instrumental variables and possibly other covariates. In this context, the first-stage equation is modeling the extent to which take-up is influenced by randomly assigned group status. Assigned group status should influence take-up because sample members assigned to the intervention group are supposed to receive the intervention and those assigned to the comparison group are not.

Procedures and standards for CACEs

The key feature of RCTs is random assignment of participants to the intervention and comparison conditions. However, participants do not always comply with their assigned conditions. In the intervention group–the group in which participants are eligible for the intervention–some participants might choose not to receive intervention services. In the comparison group–the group in which participants are ineligible for the intervention–some participants might nevertheless receive the intervention. This phenomenon is referred to as noncompliance.

In the presence of noncompliance, study authors typically estimate either one or both of two impacts:

- The effect of being assigned to the intervention, known as the intent-to-treat (ITT) effect. ITT represents the mean difference in outcomes between the assigned intervention group and the assigned comparison group regardless of compliance.
- The effect of receiving the intervention, which is often estimated by CACE.[21] The CACE represents the average effect of taking up the intervention among compliers; that is, those who would take up the intervention if assigned to the intervention condition as well as those who would not take up the intervention if assigned to the comparison condition.

The advantage of the CACE is that it seeks to estimate the effect of the intervention on those who received it. An ITT seeks to estimate the effect of an intervention on those who were initially assigned to conditions, regardless of intervention receipt. An ITT evaluation of an effective intervention could still find a small effect in two scenarios in which not all participants are compliers. In the first scenario, participants assigned to the intervention group frequently decline to receive the intervention. An ITT would combine the effect of the intervention for those who received it and the effect for those who were assigned to receive it but did not. In the second scenario, participants in the comparison group frequently receive the intervention even as they were not expected to do so. An ITT would compare outcomes for two groups in which many participants received the intervention, finding a minimal difference. In both scenarios, a CACE would untangle the differences in participation from the effect of receipt of the intervention. If compliance were perfect, and everyone assigned to the intervention group received the intervention and everyone assigned to the comparison group did not, the CACE estimate would equal the ITT estimate.

The CACE cannot be estimated with a subgroup analysis because compliers cannot be fully distinguished from other sample members. In particular, among sample members assigned to the intervention group, compliers cannot be distinguished from always-takers—those who would always take up the intervention, regardless of their randomly assigned status—because both groups take up the intervention. Among sample members assigned to the comparison group, compliers cannot be distinguished from never-takers—those who would never take up the intervention, regardless of their randomly assigned status—because neither group takes up the intervention.

Instead, the CACE is typically estimated with an instrumental variables estimator, which uses only the variation in take-up that is induced by the random assignment process to estimate the impacts of taking up the intervention on outcomes. The instrumental variables estimator is implemented via two-stage least squares, which estimates the effect of assignment to the intervention group on receipt of the intervention, and then estimates the effect of predicted receipt of the intervention on the outcome of interest. As discussed later, conventional statistical tests based on instrumental variables estimators perform well only if sample members' random assignment to the intervention group has a strong association with receipt of the intervention.

[21] In some disciplines, the CACE is also referred to as the local average treatment effect. Influential papers by Imbens and Angrist (1994) and Angrist et al. (1996) provide a formal discussion of how the CACE can be identified and estimated.

This section is intended to specify the scenarios under which CACE estimates from RCTs are eligible for review and subsequently eligible to be rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*.

WWC eligibility criteria for CACE estimates

To be eligible for a WWC review, a CACE estimate from an RCT must be based on statistical methods that meet all four conditions listed next. Studies that do not meet these criteria will not be evaluated.

- ***The endogenous independent variable must be a binary indicator for taking up any portion of the intervention.*** The WWC does not yet have standards for evaluating studies that estimate the relationship between an outcome and a continuous measure of intervention dosage, so the endogenous independent variable must be binary. Moreover, because it is possible that any positive dosage of the intervention could affect outcomes, the endogenous independent variable must distinguish sample members who took up any portion of the intervention from those who did not.
- ***Each structural equation estimated by the study, regardless of the number of instruments, must have exactly one endogenous independent variable, a binary indicator for participation.*** With multiple endogenous independent variables, criteria for evaluating instrument strength (see Stock & Yogo, 2005) would require matrix algebraic quantities that are rarely reported in education evaluations.[22] The instrumental variables must be binary indicators for the intervention and comparison groups to which subjects are randomly assigned. If random assignment forms two assignment groups—one assigned intervention group and one assigned comparison group—then there will be one instrumental variable, a binary indicator that distinguishes the groups.

  In some cases, a CACE estimate may use multiple instrumental variables that induce variation in a single endogenous independent variable. For example, if random assignment is conducted separately in several sites, then a study could interact the intervention assignment indicator with site indicators, and then use both the intervention assignment indicator and the interaction terms as instruments. The site indicators would serve as covariates in both the first-stage and structural equations. The use of these multiple instruments allows the first-stage equation to model variation across sites in the extent to which assignment to the intervention group influences take-up.[23] Another example in which multiple instrumental variables may be warranted is when there are three or more groups—for instance, a group with highest assigned priority for receiving the intervention, a group with lower assigned priority, and an assigned comparison

[22] With multiple endogenous independent variables, evaluating instrument strength would require calculating the Cragg-Donald statistic, which is the minimum eigenvalue from the matrix analog of the first-stage *F* statistic (Cragg & Donald, 1993; Sanderson & Windmeijer, 2016; Stock & Yogo, 2005). Many applied researchers would find it challenging to calculate this statistic unless they had access to specific software that performs this calculation (for instance, the "ivregress" command in Stata). Moreover, if a study did not report this statistic, then the WWC would not be able to calculate it without the individual-level data used for the evaluation.

[23] A multisite CACE estimate does not have to use site-specific intervention assignment indicators; a single intervention assignment indicator can serve as the sole instrumental variable, in which case the study is choosing not to model differences across sites in the effects of intervention assignment on take-up.

group that cannot receive the intervention—to which each subject could be randomly assigned. In this scenario, the instrumental variables are binary indicators for all but one of the assignment groups.[24]

- ***The sets of baseline covariates—independent variables other than the endogenous independent variable and instrumental variables—must be identical in the structural equation and first-stage equation.*** If baseline covariates are included in the analysis, then the structural equation and first-stage equation must contain identical sets of baseline covariates, or else the study will violate either an eligibility criterion specified above or technical conditions needed for model estimation. In particular, if a baseline covariate from the first-stage equation is not included in the structural equation, then it is effectively serving as an instrumental variable that is not among the types of eligible instruments. If a baseline covariate from the structural equation is not included in the first-stage equation, then the model will lack enough sources of variation to estimate all of the coefficients in the structural equation—a scenario known as underidentification.
- ***The study must estimate the CACE using two-stage least-squares regression or a method that produces the same estimate as two-stage least-squares regression.*** In two-stage least-squares regression, the estimated impact of take-up on outcomes is equivalent to that produced by the following two stages. First, the first-stage equation is estimated with ordinary least squares and predicted values of take-up are obtained from these estimates. Second, the endogenous take-up variable is replaced by its predicted values in the structural equation, which is then estimated by ordinary least squares. From this second stage, the estimated coefficient on the predicted take-up variable is equivalent to the two-stage least-squares regression estimate of the CACE. The standard error of the coefficient must be adjusted to account for the first-stage prediction, as discussed next.

Process for rating CACE estimates

The WWC evaluates a CACE estimate from an RCT based on different criteria depending on whether the RCT has low or high attrition, as follows:

- **A CACE estimate from an RCT with low attrition** is rated *Meets WWC Standards Without Reservations* if it satisfies two conditions: no clear violations of the exclusion restriction and sufficient instrument strength.[25] It is rated *Does Not Meet WWC Standards* if either condition is not satisfied.
- **A CACE estimate from an RCT with high attrition** is rated *Meets WWC Standards With Reservations* if it satisfies three conditions: no clear violations of the exclusion restriction, sufficient instrument strength, and

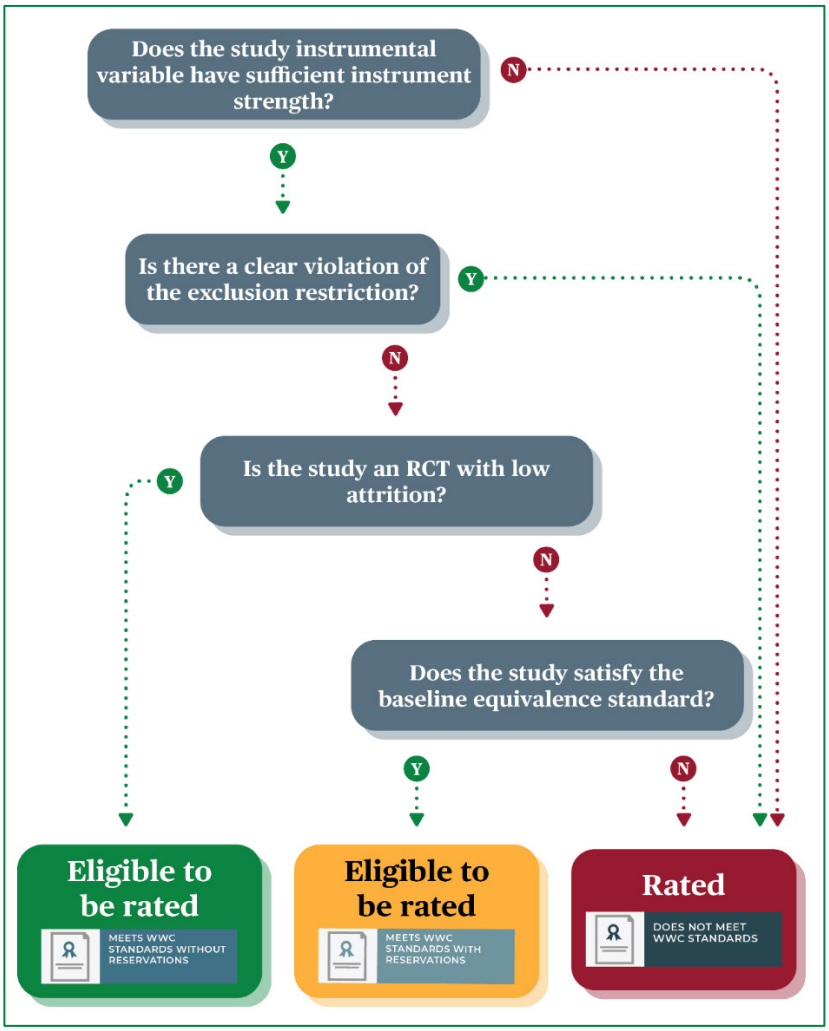
[24] In these examples, there is still only a single take-up variable, and thus, the study still estimates a single average impact of take-up on outcomes.

[25] Another assumption required for the internal validity of CACE estimates is called monotonicity (Angrist et al., 1996). Under this assumption, anyone who would take up the intervention if assigned to the comparison condition would also do so if assigned to the intervention condition. In other words, it is assumed that there are no individuals who would take up the intervention if assigned to the comparison condition but would not take up the intervention if assigned to the intervention condition; no participants are "defiers" who will always seek to enter the opposite condition from the one to which they are randomized. This assumption is not directly verifiable. However, it seems at least as plausible as other unverifiable assumptions that are needed for ITT impacts to attain causal validity, such as the assumption that each subject's outcome is unaffected by the intervention status of other subjects. Therefore, these standards assume that monotonicity is satisfied.

an acceptable adjustment for baseline differences. It is rated *Does Not Meet WWC Standards* if any of the conditions are not satisfied.

The review process for CACE estimates is outlined in figure 11. The following sections provide details on the procedures for assigning ratings to CACE estimates from RCTs with high and low attrition. Chapter III, Compositional change describes the method for determining whether an RCT has low or high attrition when rating CACE estimates. However, the formula for computing attrition in a CACE design is different from the formula for computing attrition in an RCT. See appendix G in the technical appendices for the appropriate attrition formula.

Figure 11. *Review process for studies that report findings from complier average causal effect analyses*



Note: To receive a research rating of *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*, the study finding also must satisfy the eligibility requirements in Chapter II, Study eligibility requirements and the finding must satisfy the WWC's outcome standards and be free of confounding factors.

Procedures for rating CACE estimates with low attrition

A CACE estimate with low attrition is rated *Meets WWC Standards Without Reservations* if it satisfies two criteria: no clear violations of the exclusion restriction and sufficient instrument strength. If either criterion is not met, then the CACE estimate is rated *Does Not Meet WWC Standards*. These criteria are described in detail below, and the conceptual background for these criteria is described in appendix G in the technical appendices.

Criterion 1: No clear violations of the exclusion restriction

For a CACE estimate to have no clear violations of the exclusion restriction (the assumption that all group assignment does not affect outcomes for individuals who would either always or never take up the intervention regardless of assignment), a necessary condition is that the study must report a definition of take-up that is the same across assigned groups. Moreover, the WWC's lead methodologist for a review has the discretion to determine that a study fails to satisfy the exclusion restriction as a result of a situation in which assignment to the intervention can materially influence the behavior of subjects even if they do not take up the intervention.

> *Example*. The exclusion restriction would be violated if subjects assigned to the intervention group received offers to convince them to enroll in the comparison group instead. Other violations of the exclusion restriction may be similar to confounding factors. One example is that all classes in the intervention group meet in the morning and all classes in the comparison group meet in the afternoon.

See appendix G in the technical appendices for additional discussion of violations of the exclusion restriction.

Criterion 2: Sufficient instrument strength

Depending on the number of instruments, a CACE estimate must report a first-stage *F* statistic—the *F* statistic for the joint significance of the instrument(s) in the first-stage equation—at least as large as the minimum required level shown in table 20. The minimum required levels are based on Stock and Yogo's (2005) derivations on the minimum first-stage *F* statistic needed to ensure that the actual type I error rate is unlikely to exceed .10 for a *t* test whose assumed type I error rate is .05.[26] When there is one instrument, study authors may report a *t* statistic instead. In this case, the *F* statistic is equal to the square of the *t* statistic.

Why does WWC worry about weak instruments?

An instrument is weak if it does not effectively predict its corresponding endogenous independent variable. In the WWC's case, this would mean that random assignment to treatment would not predict participation in the intervention; in other words, non-compliance is common. If noncompliance is very common, then our estimates of the CACE will be biased, trying to draw too much meaning from too few compliers. Furthermore, high noncompliance means that unadjusted confidence intervals from CACE estimates are incorrect, which could result in findings being assigned an inappropriate level of statistical significance

[26] Specifically, the minimum required first-stage *F* statistic is the critical value for rejecting the null hypothesis that the instruments are weak enough to yield type I error rates exceeding .10. See Stock and Yogo (2005) for details. Although it is common for researchers to use a rule of thumb that the *F* statistic must exceed 10, table 20 imposes a stronger requirement. Stock and Yogo's (2005) analyses are a refinement and improvement to the Staiger-Stock (Staiger & Stock, 1997) rule of thumb, which states that instruments with a first-stage *F* value less than 10 should be deemed weak.

When baseline covariates are included in the two-stage least-squares regression, the first-stage *F* statistic assesses the joint significance of the instruments in the first-stage equation while controlling for the baseline covariates. In such cases, the *F* statistic should only reflect the significance of the instruments, and not the significance of the baseline covariates. If the unit of assignment differs from the unit of analysis, then the study must report first-stage *F* statistics after adjusting for clustering.

If a CACE estimate does not have an associated first-stage *F* statistic reported in the study, then the WWC will send an author query. If the study authors do not provide this statistic after being queried, then the WWC will try to calculate the first-stage *F* statistic using the formula listed in appendix G in the technical appendices, if there is only one instrumental variable and no clustering. If none of these options enables the first-stage *F* statistic to be identified, then the study does not demonstrate sufficient instrument strength and is rated *Does Not Meet WWC Standards*.

Table 20*. First-stage* F *statistic thresholds for satisfying the criterion of sufficient instrument strength*

Number of instruments	Minimum required first-stage *F* statistic	Number of instruments	Minimum required first-stage *F* statistic
1	16.38	16	52.77
2	19.93	17	55.15
3	22.30	18	57.53
4	24.58	19	59.92
5	26.87	20	62.30
6	29.18	21	64.69
7	31.50	22	67.07
8	33.84	23	69.46
9	36.19	24	71.85
10	38.54	25	74.24
11	40.90	26	76.62
12	43.27	27	79.01
13	45.64	28	81.40
14	48.01	29	83.79
15	50.39	30	86.17

Source: Stock and Yogo (2005).

Procedures for rating CACE estimates with high attrition

A CACE estimate from an RCT with high attrition is rated *Meets WWC Standards With Reservations* if it satisfies three criteria: no clear violations of the exclusion restriction, sufficient instrument strength, and the baseline equivalence standard. If either criterion is not satisfied, then the CACE estimate is rated *Does Not Meet WWC Standards*.

The first two criteria are identical to those discussed in the previous section for rating CACE estimates from RCTs with low attrition. The remainder of this section describes the third criterion: the baseline statistical adjustment requirement.

The baseline equivalence standard for CACE estimates in RCTs with high attrition follows the same baseline equivalence standard described in Chapter III, Compositional change.[27]

Procedures and standards for repeated-measures analyses

This section provides guidance on two types of group designs in which subjects are observed in multiple time periods, sometimes referred to as repeated-measures analyses:

- Analyses of simple gain scores.
- Analyses in which the dependent variable includes data from multiple time points.

When might the WWC use repeated-measures analyses?

While many WWC-reviewed analyses compare students who received an intervention against students who did not, using data collected around the end of the intervention period, repeated-measures analyses use data collected at multiple time periods to learn more about the trajectory of students' progress. One setting in which a repeated-measures analysis might make sense would be if the intervention is of indefinite duration, such as assignment to a charter school. Another example would be if a study examines the long-term effects of a short-term intervention, such as a behavioral intervention that seeks to instill good habits that are maintained after the intervention is complete.

The additional considerations for these analyses described next do not apply to analyses in which preintervention measures of the outcome are instead included as covariates in the analytical model, and they do not apply in single-case designs. Regardless of the approach used to analyze the repeated measures, the baseline adjustment requirement must still be satisfied on the baseline measures specified in the *Handbook*. The WWC considers analyzing simple gain scores, difference-in-differences adjustments, and fixed effects at the level of the unit of assignment as acceptable statistical adjustments, but only when there is evidence that the baseline and outcome measures are measured on the same scale (for example, *z* scores), based on the requirements described in Chapter III, Compositional change.[28]

Analyses of simple gain scores

Simple gain scores can be calculated by subtracting a pretest from the posttest. Some study authors use the resulting difference as the dependent variable in an impact analysis. The analyses of simple gain scores are eligible

[27] Because attrition is the key source of bias that can lead to baseline differences in RCTs, assumptions about attrition behavior shape what types of assumptions about baseline differences are reasonable. Baseline differences emerge when intervention group members who leave the study are different from comparison group members who leave the study, resulting in a baseline imbalance between groups among those who remain in the study. Stated differently, baseline differences emerge when assignment to the intervention is associated with the composition of people who stay or leave. The approach to calculating attrition, explained in Chapter III, Compositional change, was built on the notion that assignment to the intervention is unlikely to have opposite effects on attrition rates for different subpopulations. By similar logic, assignment to the intervention is unlikely to have opposite effects on the types of sample members who leave the study in different subpopulations. For this reason, the WWC finds it reasonable and realistic to assume that baseline differences have the same sign for always-takers, compliers, and never-takers.

[28] The repeated-measures analyses discussed in this section—simple gain scores and analyses in which the dependent variable includes data from multiple time points—would rarely use regression adjustment or analysis of covariance to adjust for a preintervention measure of the outcome. However, a repeated measures analysis may use these adjustment approaches to account for other preintervention measures that might be specified in the *Study Review Protocol*.

to meet WWC group design standards. However, to be reported by the WWC, effect sizes from gain score analyses must be based on standard deviations of the outcome measure collected at the follow-up time point without adjustment for the baseline measure. The WWC will convert gain score standard deviations to unadjusted standard deviations if the study authors reported the baseline-outcome correlation (see equation E.19 in appendix E in the technical appendices). The WWC will request the unadjusted standard deviations of the posttest if they cannot be computed but are required to compute the effect size. If the WWC cannot calculate an effect size for an outcome based on acceptable standard deviations, then the finding cannot be used to assign an effectiveness rating to the intervention in that outcome domain, although the finding can still meet WWC standards.

Analyses in which the dependent variable includes data from multiple time points

In these repeated-measures analyses, the analysis includes multiple observations for units assigned to the intervention and comparison conditions, and the dependent variable includes data from all time points. For example, students are observed in two or more periods (at least one preintervention and one postintervention) and the analysis includes multiple observations for each student, one at each point in time. These analyses include, but are not limited to, the following:

- ***Difference-in-differences analyses***. Units are assigned to conditions and observed once before and once after the intervention is delivered. The difference in outcomes between conditions is obtained for observations before and after the start of the intervention period, and the difference in these differences is the effect of the intervention. This requires including an indicator for the time period associated with the intervention in the model, along with an interaction term for assignment to the intervention group in the postintervention period.
- ***Comparative interrupted time series***. Units are assigned to conditions and observed at several points in time. As a generalization of difference-in-differences, an indicator for the time period associated with the intervention must be included in the model along with an interaction term representing assignment to the intervention group in the postintervention period. These analyses may include additional terms that adjust for the correlation between time and outcomes, and may include linear trends, nonlinear trends, or fixed effects corresponding to each observation time point.
- ***Growth curve models***. Units are assigned to conditions and observed at several points in time. Indicators for each intervention period must be included in the model, as well as interaction terms for assignment to the intervention group and the time period associated with the intervention.

To be eligible for review as a group design study, a repeated-measures analysis must measure the effect of the intervention by comparing exclusive intervention and comparison groups, meaning that a participant can belong to only a single group at each point in time. Analyses in which the same participant is analyzed as a member of both the intervention and comparison groups at different times are not eligible for review unless separate intervention and comparison groups are used for group contrasts at each time period after baseline. In other words, if all intervention observations are evaluated at the same time and all comparison observations are evaluated at a different time, time is a confounding factor and the analysis does not meet standards. Additionally, to be eligible to meet WWC standards, the analysis must adequately account for the time periods associated with the intervention and preintervention conditions. A study that fails to account for differences across time periods risks producing biased intervention effects if a change in policy or environment, unrelated to the intervention, that may be correlated with the outcome occurs in some time periods and not others.

Example. Consider a study of an intervention that is provided to students in group A during period 1 and to students in group B during period 2. Students in groups A and B receive the comparison condition when not receiving the intervention. This study design is often referred to as a "switching replications" or "crossover" design. If the study authors examined the impact of the intervention separately in each period by comparing students who received the intervention in that period with the distinct group of students who did not, then this study would be eligible for review as a group design study, and its findings would have the opportunity to meet WWC standards. However, if the study authors examined the impact of the intervention separately for each group of students by estimating differences between comparison period outcomes and intervention period outcomes, then this study would fail to meet standards in both samples because the time period would be confounded with intervention status in each group of students, thereby making it a series of two within-group designs. When the analysis of such a study does not provide an impact estimate comparing exclusive intervention and comparison groups from each time period, the WWC will consider the study ineligible for review.

Figure 12 illustrates this example. Analytic samples that would meet standards are outlined in yellow; these include comparing group A with group B in either time period, as well as a repeated-measures analysis that would have both group A and group B included twice, once in the intervention group and once in the comparison group. Analytic samples that would not meet standards are outlined in red; these would entail comparing two time periods within a single group of students, in which all students receive the same condition at the same time.

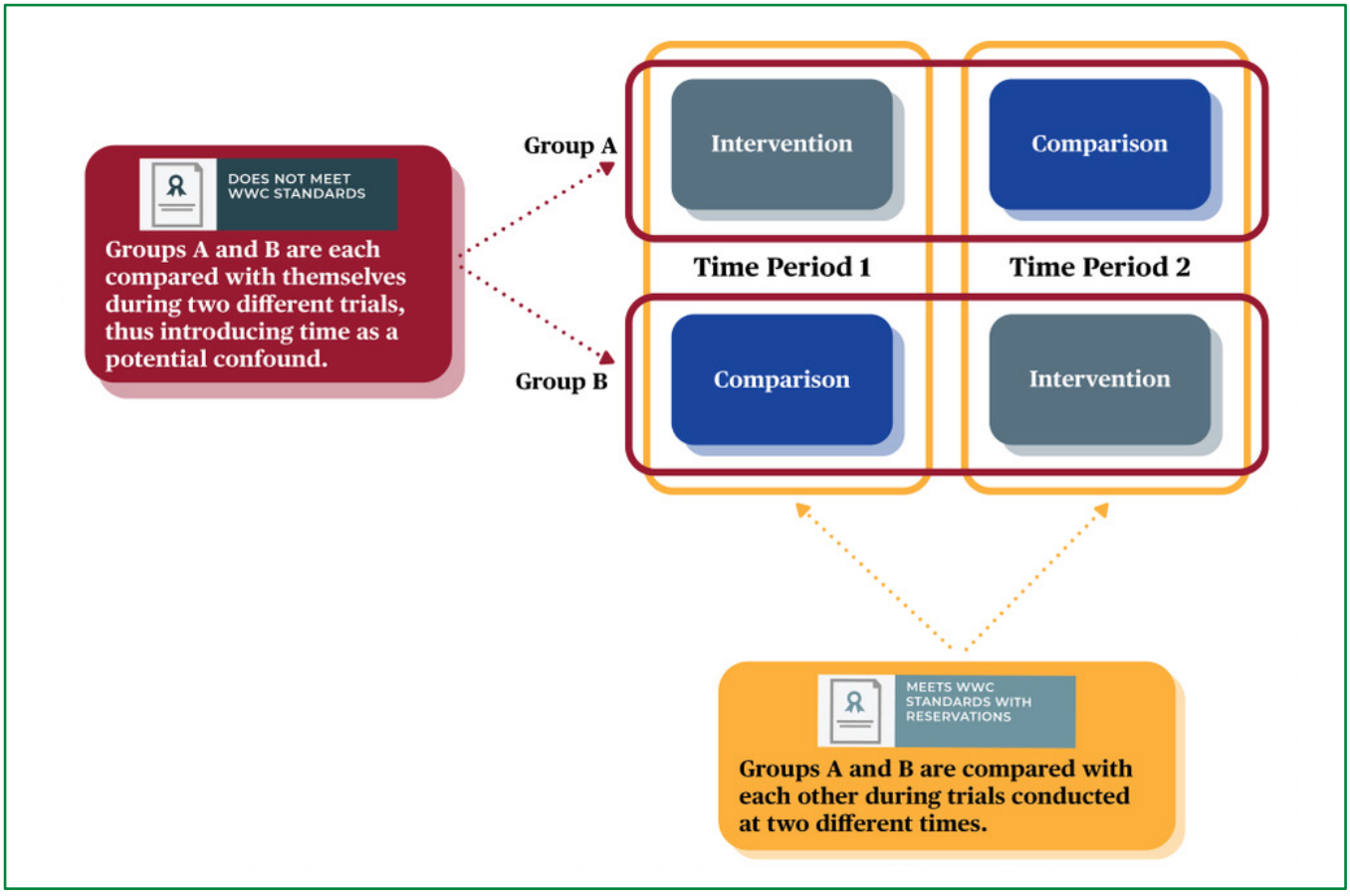
For repeated measures analyses with randomization to condition, the WWC reviews impact findings separately at each point in time included in these analyses to assess sample attrition from the time of assignment to condition. Each impact estimate, or an average of impacts across time periods, is eligible to be rated *Meets WWC Standards Without Reservations* if all groups were formed in an RCT with a low risk of bias due to compositional change and otherwise is eligible to be rated *Meets WWC Standards With Reservations*. When study authors observe more than one preintervention period, the WWC requires that the authors of QEDs, RCTs with a high risk of bias due to compositional change, or RCTs of unknown bias due to compositional change satisfy the baseline equivalence standard in the preintervention period closest to the start of the intervention. However, if the study is a QED, then the average impact is eligible to be rated *Meets WWC Standards With Reservations* only if the baseline equivalence standard is satisfied separately for the two time points. Although the groups may have been equivalent at the start of period 1, exposure to the intervention for subjects in group A during period 1 might lead to differences in the groups at the start of period 2. Alternatively, if subjects were randomly assigned to conditions in period 1, then the WWC will review the period 1 finding as an RCT, while the WWC will review the period 2 impact estimate as a QED that must satisfy the baseline equivalence standard because subjects in group A were exposed to the intervention.

Quasi-experimental designs, RCTs with a high risk of bias due to compositional change,[29] or compromised RCTs that pool impact estimates across time periods are ineligible for review by the WWC because they would require authors to conduct additional analyses beyond simply providing descriptive information about the study

[29] As one exception, low cluster-level attrition RCTs that demonstrate representativeness do not need to satisfy the baseline equivalence standard, despite having a high risk of bias due to compositional change, as noted in step 3d.

samples. However, if authors report separate impact estimates for each postintervention time point, the WWC will review those findings.

Figure 12. *Eligible and ineligible samples for repeated-measures analyses*



Note: The contrasts outlined in yellow are eligible to meet standards, while the contrasts outlined in red reflect a confound with time and therefore are not eligible to meet standards.

Requirements for difference-in-differences analyses

In a difference-in-differences analysis, which includes just two time periods—that is, preintervention and postintervention, the WWC requires including indicators for the intervention condition, the time period associated with the intervention, and an interaction between these two indicators. This requirement is typically satisfied by an ordinary least squares analysis that includes the intervention, time period, and interaction indicators as independent variables. The outcome variable must be measured on the same scale in both time periods. In such an analysis, the coefficient on the interaction term provides the difference-in-differences estimate of the impact of the intervention. The *p* value of this estimate is used to assess the statistical significance of the impact. Another acceptable analytic approach is a repeated-measures analysis of variance with one between-groups factor distinguishing the intervention and comparison groups and at least one within-groups factor distinguishing time period, as long as it contains an interaction of these two factors.

A study that instead reports only the coefficient on the intervention indicator and does not include an interaction between the intervention indicator and time period provides a biased estimate of the effect of the intervention because it measures the average difference in the outcome between the intervention and comparison groups across both the preintervention and postintervention periods. Such an analysis does not provide a credible estimate of the effectiveness of the intervention. If the study authors do not provide the WWC with findings from a credible analysis, then the study finding will be rated *Does Not Meet WWC Standards*.

Requirements for comparative interrupted time series and growth-curve analyses

Analyses with more than two time periods, including most comparative interrupted time series and growth curve analyses, must account for the preintervention and postintervention periods, including an interaction with the intervention indicator, but they can also account for additional time periods. For example, a finding from a comparative interrupted series study that uses fixed effects for each time period and for each unit of assignment would be eligible to receive the research rating *Meets WWC Standards With Reservations* if the study also estimated impacts separately for each postintervention time point and satisfied the baseline equivalence standard using the baseline period closest to the start of the intervention.

Procedures and standards for analyses with endogenous covariates

A covariate is considered endogenous if it may be affected by the intervention, making it impossible to tell whether the effect of the intervention on the outcome of interest is happening directly or through the intervention's influence on the endogenous covariate. When one or more potentially endogenous covariates are included in the analysis, the WWC can either use alternative model specifications reported in the study that do not include these endogenous covariates or request unadjusted means—or adjusted means based on only the nonendogenous covariates—and unadjusted standard deviations from the study authors. However, if the potentially endogenous measure is used to satisfy the baseline equivalence standard described in Chapter III, Baseline equivalence standard, then the WWC will note in its reporting that the study measures the effect of the portion of the intervention that occurred after the measure was assessed and until the time of the follow-up assessment. Even though the baseline measure may have been influenced by the intervention, it can be used to satisfy the baseline equivalence standard. It is not necessary to include the same reporting note for a baseline measure assessed shortly after the start of the intervention that was included as a covariate in the analysis, but it is not used to satisfy the baseline equivalence standard.

Procedures and standards for analyses with imputations for missing data

Despite the best efforts of researchers, sometimes it is not possible to collect data for all subjects in a study sample. To address missing data for baseline or outcome measures, study authors might use a variety of analytical approaches. Study authors might focus on the analytic sample of subjects for which all data were collected, or the study authors may impute values for the missing data so that more subjects can be included in the analysis. The review process for a study with missing data depends on the research design, the method used to address the missing data, and whether the study has missing baseline data, outcome data, or both.

The steps in the review process for studies with missing data are outlined in figure 13 and described below. Steps 1 and 2 must be performed for any study with missing data, while step 3 relates to studies with imputed outcome data in the analytic sample.

Step 1. Did study authors use an acceptable approach to address all missing data in the analytic sample?

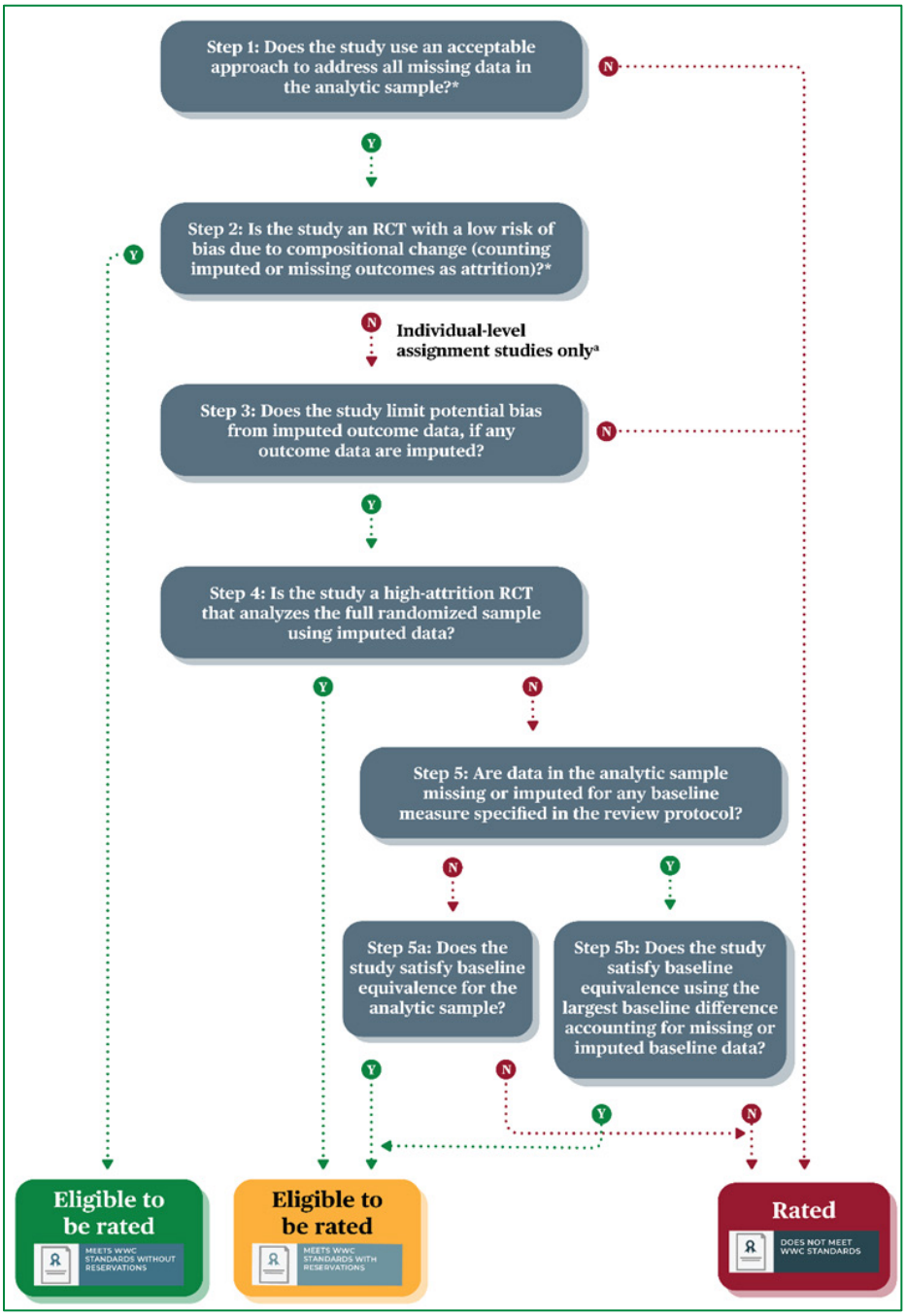
The first step in the review process for studies with missing data is to determine whether an acceptable approach was used to address the missing data. Acceptable methods include complete-case analysis, regression imputation, dummy imputation, maximum likelihood, and nonresponse weighting. To be eligible to be rated *Meets WWC Standards With Reservations* or *Meets WWC Standards Without Reservations*, an analysis must use one of these methods to address the missing data (see table 21). This requirement applies to all data used in the analysis, whether for an outcome measure or a baseline measure. More specifically, the requirement applies both to baseline measures specified in the *Handbook* as required for assessing baseline equivalence and those not specified.

When an analysis uses one or more of these methods and satisfies all other requirements to receive a research rating of *Meets WWC Standards With Reservations* or *Meets WWC Standards Without Reservations*, the WWC will report findings, including effect sizes, according to the general approach to WWC reporting outlined in Chapter VII. However, the WWC will not report statistical significance for methods that do not provide accurate standard error estimates. For some imputation methods, the WWC will report statistical significance provided certain requirements are met, as described in the last column in table 21.

To obtain appropriate estimates of statistical significance in cluster-level assignment studies that analyze individual-level data, approaches to address missing outcome data must account for the nonindependence of observations within clusters. This can be done using standard approaches (for example, hierarchical linear modeling) in complete case analyses. However, for the WWC to confirm statistical significance in a study with cluster-level assignment that uses regression imputation, maximum likelihood, or nonresponse weights to address missing outcome data, and analyzes individual-level data, the study must provide evidence that the approach appropriately adjusts the standard errors for clustering by citing a peer-reviewed journal article or textbook that describes the procedure and demonstrates its effectiveness. In analyses using these three approaches that do not include an acceptable adjustment for clustering, the WWC will not apply its adjustment for clustering because it may not be accurate for analyses using imputation methods. The WWC does not currently have a recommended method of calculating standard errors when using imputation methods in cluster-level assignment studies, and the burden for demonstrating that the approach is appropriate rests with the study authors.

Finally, if a study uses an approach not listed in table 21 that is supported by a citation to a peer-reviewed journal article or textbook that describes the procedure and demonstrates that it can produce unbiased estimates under an assumption that the missing data are unrelated to unmeasured factors, the WWC may consider it an acceptable approach after review by experts. If so, the WWC will release guidance that updates the list of acceptable approaches. Table H.1 in appendix H in the technical appendices provides an expanded version of the following table, including detailing the relevant additional requirements for some approaches such as multiple imputation.

Figure 13. *Research ratings for randomized controlled trials and quasi-experimental designs with missing outcome or baseline data*



Note: To receive a rating of *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*, the study must satisfy the other requirements presented in Chapter II, Study eligibility requirements, including that the study must examine at least one eligible outcome measure that meets review standards and be free of confounding factors.

a. Steps 3-5 apply to individual-level assignment studies only. Steps 1 and 2 apply to cluster-level assignment studies in that such studies must use an acceptable missing data approach (step 1) and count imputed outcome data as attrition in assessing cluster-level attrition, individual-level attrition, and representativeness (step 2). Steps 3-5 do not apply to cluster-level assignment studies, however. Instead, the cluster-level baseline equivalence standard (step 4b in Chapter III) applies for cluster-level assignment studies that must satisfy the baseline equivalence standard and have missing or imputed data in the analytic sample for the outcome measure or a required baseline measure.

Table 21. *Acceptable approaches for addressing missing baseline or outcome data*

Method	Research design	Type of missing data that can be addressed	Additional requirements to meet WWC standards
Complete case analysis	All	Baseline and outcome	N/A
Regression imputation	All	Baseline and outcome	The imputation model must: 1. be conducted separately by condition or include an indicator variable for condition; 2. include all covariates used for adjustment in the impact model; and 3. include the outcome when imputing missing baseline data.
Dummy imputation	Uncompromised RCTs only[a]	Baseline only	N/A
Maximum likelihood	All	Baseline and outcome	Use standard statistical package or include relevant citations
Nonresponse weights	All	Outcome only[b]	The model to predict missingness must: 1. estimate probabilities of missingness separately by condition or include an indicator variable for condition; and 2. include all baseline measures specified in the *Handbook* as required for baseline equivalence.

N/A is not applicable. RCT is randomized controlled trial.

Note: See appendix H for additional details for analyses with missing data.

a. However, for quasi-experimental designs and compromised randomized controlled trials, dummy imputation can still be applied to baseline measures not specified in the *Handbook* as required to adjust for baseline differences.

b. With nonresponse weights, participants without observed outcome data will not be included in impact estimation models, but participants with observed outcome data will be weighted so that they resemble the full sample with and without outcome data.

Step 2. Is the study an RCT with a low risk of bias due to compositional change (counting imputed or missing outcomes as attrition)?

The second step in the review process for studies with missing data is to determine whether the study finding is from an RCT with a low risk of bias due to compositional change as defined in Chapter III, Compositional change. Individual-level RCTs satisfy this requirement by demonstrating low attrition. Cluster RCTs satisfy this requirement by demonstrating low cluster-level attrition, demonstrating low individual-level attrition, and excluding high-risk joiners (if any) from the analytic sample.

When calculating overall and differential attrition rates, sample members with imputed or missing outcome data are counted as attrition. Maximum likelihood methods can include sample members with missing data in the analysis without imputation; these sample members are also counted as attrition. Sample members with

imputed or missing data can introduce bias if the patterns of missingness depend on unmeasured factors (sometimes known as "nonignorable missing data"). For cluster RCTs, sample members with missing or imputed outcome data count as attrition for cluster-level attrition (step 3a in chapter III), individual-level attrition (step 3c), and lack of representativeness (step 3d).

When the risk of bias due to compositional change is low, the WWC will ignore the potential bias from imputed or missing data because the amount of missing or imputed data is unlikely to lead to bias that exceeds the WWC's tolerable level of potential bias. An RCT with a low risk of bias due to compositional change is eligible to be rated *Meets WWC Standards Without Reservations* as long as the study used an acceptable method to address missing data.

Step 3. Did the study authors limit potential bias from missing and imputed outcome data, if any outcome data are missing or imputed?

Imputed outcome data can affect the rating of a QED, an RCT with a high risk of bias due to compositional change, or a compromised RCT. To be eligible for a research rating of *Meets WWC Standards With Reservations*, these types of studies must satisfy an additional requirement designed to limit potential bias from using imputed outcome data instead of actual outcome data. This requirement does not apply if the study imputed only baseline, not outcome, data.

The following steps (missing data steps 3–5) only apply to individual-level assignment studies, as the derivation of equations for assessing bias assume individual-level assignment. For reviewing cluster-level assignment studies, a reviewer should consult the earlier guidance in Chapter III for such studies, modifying as required by missing data step 1 (needs an acceptable missing data approach) and missing data step 2 (sample members with missing or imputed outcome data count as attrition). Missing data steps 3–5 do not apply to cluster-level assignment studies, however. For cluster-level assignment studies, if the sample of individuals used to assess baseline equivalence differs from the sample of individuals used in analyses of outcomes due to missing or imputed data, then a reviewer should use the standards for assessing cluster-level baseline equivalence, as opposed to individual-level baseline equivalence.

What unmeasured factors could a missing outcome variable depend on?

The WWC standards for imputed outcome data assume that the missing data depend only on observed predictors. In reality, not all factors that predict the values of the missing data are observed, or even observable. Many academic and behavioral outcomes are affected by recent events in the student's home, such as whether family members are fighting or whether the student is receiving enough to eat; these would be very difficult to measure reliably and in a timely fashion. Even variables that are influential and measurable, such as students' parents' educational attainment, are frequently not collected. However, it is unlikely that there are unobserved factors that would affect the realizations of the missing outcome data that would not affect the observed outcome data, so

The imputation methods the WWC considers acceptable are based on an assumption that the missing data depend on measured factors, not unmeasured factors. If that assumption does not hold, then impact estimates may be biased. Therefore, authors of group design studies besides RCTs with a low risk of bias due to compositional change that use acceptable approaches to impute outcome data should demonstrate that they limit the potential bias from using imputation methods to less than 0.05 standard deviation as described in this

step. See appendix H in the technical appendices for more information on how the WWC assesses whether study authors used acceptable imputation methods to limit the potential bias to the estimate of the effect of the intervention.

An analysis of a sample with imputed outcome data can produce biased estimates of the effect of the intervention if the subjects with observed data differ from the subjects with missing data, and some of the differences are unmeasured. In this case, if outcomes could be obtained for all sample members, then the average for subjects in the intervention or comparison condition with observed outcome data would differ from the average for subjects whose outcome data were not observed. Comparing the differences in these means for the intervention and comparison groups, if known, would indicate the magnitude of possible bias, but because the missing outcomes are not observed, the WWC instead assesses the degree of bias using baseline data.

The WWC estimates the potential bias from missing outcome data due to unmeasured factors by comparing means of the baseline measure specified in the review protocol as required for assessing baseline equivalence, separately for the intervention and comparison groups, for two samples: the complete analytic sample and the analytic sample restricted to cases with observed outcome data. A smaller difference in these two means within one or both conditions lowers the likelihood that the missing data are related to factors that could lead to bias in the impact estimate.

To translate the intervention and comparison group differences in baseline means into an estimate of bias in the outcome effect size, the WWC uses the pooled standard deviation of the baseline measure and the correlation between the baseline and outcome measure. The missing data section of the technical appendices provides the formulas the WWC uses to estimate the potential bias (equations H.7–H.9 in appendix H). The technical appendices describe the approach used when the *Handbook* specifies that baseline adjustments are required on multiple baseline measures. The formulas used to assess the bias also differ if the baseline measure is not observed for all subjects in the analytic sample (equations H.16–H.18).

- **When the baseline measure is observed for all subjects in the analytic sample**, the WWC requires the following data from the study authors: (a) the means and standard deviations of the baseline measure for the analytic sample, separately for the intervention and comparison groups; (b) the means of the baseline measure for the subjects in the analytic sample with observed outcome data, separately for the intervention and comparison groups; and (c) the correlation between the baseline and the outcome measures. The correlation can be estimated on a sample other than the analytic sample, such as the complete case sample, or from data from outside the study if a content expert judges the settings to be similar. However, the correlation must not be estimated using imputed data.
- **When the baseline measure is imputed or missing for some subjects in the analytic sample**, in addition to (c), the following data are required: (d) the means of the baseline measure for the subjects in the analytic sample with observed baseline data, separately for the intervention and comparison group; (e) the means of the baseline measure for the subjects in the analytic sample with observed baseline and outcome data, separately for the intervention and comparison groups; (f) the standard deviations of the baseline measure for either the sample of subjects in the analytic sample with observed baseline data or the sample

with observed baseline and outcome data; and (g) the number of subjects with observed baseline data in the analytic sample by condition.

The WWC has two special considerations for applying the requirement in step 3 when an analysis uses nonresponse weights or complete case analysis:

- **When the analysis uses nonresponse weights to address missing outcome data**, in addition to (b) and (c), the WWC requires (h) the means of the baseline measure for the subjects in the sample used to estimate the weights, including cases with missing outcome data, separately for the intervention and comparison groups.
- **When the analytic sample is restricted to only observations with nonmissing outcome data**, otherwise known as a complete case analysis or listwise deletion, a study does not need to satisfy this requirement. The exclusion of complete case analyses from this requirement is not intended to imply that complete case analyses are believed to be a stronger approach for addressing missing data. Rather, the WWC's approach recognizes that the attrition standard and baseline equivalence standard can limit bias in complete case analyses because the missing data affect the analytic sample.

If these data are not reported in the study, then the WWC will request them from the study authors.

Step 4: Is the study a high-attrition RCT that analyzes the full randomized sample using imputed data?

The fourth step in the review process for missing outcome data addresses a second way that imputed outcome data can affect the rating of a study. When study authors analyze an individual-level, high-attrition RCT by imputing outcome data so that they analyze the full sample that was randomized to conditions, the study does not need to satisfy the baseline equivalence standard to be eligible to receive the rating *Meets WWC Standards With Reservations*.

In general, the WWC requires that individual-level, high-attrition RCTs assessed with the cautious boundary satisfy the baseline equivalence standard because of a risk of bias from compositional differences between the remaining intervention and comparison group members. However, some high-attrition RCTs impute all missing outcome data and analyze the original randomized sample. These high-attrition RCTs and all high-attrition RCTs assessed using the optimistic boundary are required to use an acceptable adjustment strategy for baseline differences to satisfy the baseline equivalence standard. Imputing missing outcome data and analyzing the full randomized sample preserves the integrity of the originally randomized groups. Although compositional differences are not considered to present a threat of bias, these studies, like other high-attrition RCTs, are eligible to be rated no higher than *Meets WWC Standards With Reservations*. These studies are not eligible for the highest rating because of the risk of bias from imputing a larger amount of missing outcome data compared with a low-attrition RCT.

All QEDs, high-attrition RCTs that do not analyze the original randomized sample, high-attrition RCTs assessed using the cautious attrition boundary, and compromised RCTs must demonstrate baseline equivalence (see step 5 in figure 13).

Step 5. Are data in the analytic sample missing or imputed for any baseline measure specified in the Handbook*?*

QEDs, high-attrition RCTs that do not impute data to analyze the full randomized sample, and compromised RCTs must satisfy the baseline equivalence standard to be eligible to be rated *Meets WWC Standards With Reservations*. A reviewer should proceed to step 5a if the study has no missing or imputed baseline data for all measures specified in the *Handbook* as required for assessing baseline equivalence. Otherwise, a reviewer should proceed to step 5b if the study had some missing or imputed baseline data for any measure specified in the *Handbook* as required for assessing baseline equivalence.

RCTs that do not impute data to analyze the full randomized sample and that have high attrition assessed against the optimistic attrition boundary may be rated *Meets WWC Standards With Reservations* if they use an acceptable adjustment strategy for baseline group differences, even if baseline equivalence cannot be assessed on the full analytic sample.

Step 5a. Does the study satisfy the baseline equivalence standard for the analytic sample?

If all of the missing or imputed baseline data in the analytic sample are for baseline measures not required for satisfying baseline equivalence in the outcome domain, or if no baseline data are missing or imputed, then baseline equivalence can be assessed using the usual approach described in Chapter III, Baseline equivalence standard. A study that satisfies the baseline equivalence standard using actual data for the analytic sample is eligible to be rated *Meets WWC Standards With Reservations*. Additionally, RCTs that do not impute data to analyze the full randomized sample and that have high levels of compositional change when attrition is assessed against the optimistic attrition boundary may be rated *Meets WWC Standards With Reservations* if they use an adequate baseline adjustment, even if baseline equivalence cannot be assessed on the full analytic sample.

An analysis that uses nonresponse weights to address missing outcome data must satisfy baseline equivalence using observed data for the analytic sample using weighted means.

Step 5b. Does the study satisfy baseline equivalence using the largest baseline difference accounting for missing or imputed baseline data?

If some data are missing or imputed for a baseline measure that is specified in the *Handbook* as required for satisfying baseline equivalence in the outcome domain, then the WWC uses a different process to assess baseline equivalence. In this case, the WWC estimates how large the baseline difference might be under different assumptions about how the missing data are related to measured or unmeasured factors. The largest of these estimates in absolute value is used as the baseline difference for the study.

Just as in studies with complete baseline data, a study with missing or imputed data for a required baseline measure is eligible to be rated *Meets WWC Standards With Reservations* if the largest estimated baseline difference does not exceed 0.25 standard deviation when the analysis includes an acceptable adjustment for the baseline measure, or 0.05 standard deviation otherwise. A study that satisfies this alternative baseline equivalence standard is eligible to be rated *Meets WWC Standards With Reservations*.

The WWC's approach to estimating the baseline difference in studies with missing or imputed baseline data is similar to the approach used to estimate bias from using imputed outcome data, described above. Instead of comparing means of the baseline measure, the WWC compares means of the outcome measure, separately for the intervention and comparison groups, for two samples: the entire analytic sample and the analytic sample restricted to cases with observed baseline data. A larger absolute difference in these means within a group indicates that the data may be missing in a way that is related to unmeasured sample characteristics, and the measured impact of the intervention may be biased.

To translate the intervention and comparison group differences in outcome means into an estimate of a baseline effect size, the WWC uses the pooled standard deviation of the outcome measure and the correlation between the baseline and outcome measure. Appendix H in the technical appendices provides the formulas the WWC uses to estimate the baseline effect size (equations A, B, C, and D). When the *Handbook* specifies that baseline equivalence must be assessed on multiple baseline measures, the formulas in appendix H must be applied to each required baseline measure. The formulas used to estimate the baseline difference vary based on two factors: whether the outcome measure is observed for all subjects in the analytic sample and whether the outcome data are missing or imputed.

- **When the outcome measure is observed for all subjects in the analytic sample**, the WWC requires the following data from the authors: (a) the means and standard deviations of the outcome measure for the analytic sample, separately for the intervention and comparison groups; (b) the means of the outcome measure for the subjects in the analytic sample with observed baseline data, separately for the intervention and comparison groups; (c) the correlation between the baseline and the outcome measures; and (d) an estimate of the baseline difference based on study data. As noted in step 3 of the section on imputed outcome data, the correlation can be estimated on a sample other than the analytic sample but must not be estimated using imputed data. If the authors did not impute the baseline data, then the WWC will use baseline means and standard deviations to measure the baseline difference for the portion of the analytic sample with observed baseline data. However, if the study did impute baseline data, then the WWC will include the imputed data when calculating the means but will use standard deviations based only on the observed data.
- **When the outcome measure is imputed for some subjects in the analytic sample**, in addition to (c) and (d), the following data are required: (e) the means of the outcome measure for the subjects in the analytic sample with observed outcome data, separately for the intervention and comparison groups; (f) the means of the outcome measure for the subjects in the analytic sample with observed baseline and outcome data, separately for the intervention and comparison groups; (g) the standard deviations of the outcome measure for either the sample of subjects in the analytic sample with observed outcome data or the sample with observed baseline and outcome data; and (h) the number of subjects with observed outcome data in the analytic sample by condition.

If these data are not reported in the study, then the WWC will request them from the authors.

The two special considerations for applying the requirement in step 5b when an analysis uses nonresponse weights or complete case analysis are as follows:

- An analysis that uses nonresponse weights to address missing outcome data must satisfy baseline equivalence using observed data for the analytic sample using weighted means.
- Because no baseline data are missing or imputed, a complete case analysis that excludes cases with missing baseline data must satisfy the baseline equivalence standard using the observed data for the analytic sample, as described previously in Chapter III, Baseline equivalence standard, rather than using the formulas in the technical appendices. In other words, the complete case analysis must satisfy baseline equivalence using step 5a and not step 5b.

CHAPTER VI. REVIEWING FINDINGS FROM SINGLE-CASE DESIGN STUDIES

Single-case designs (SCDs) are experimental designs with the potential to demonstrate causal effects that generally include a small number of participants. SCD studies must broadly adhere to some of the same guidelines in terms of their outcomes and general eligibility requirements. However, when SCD studies are focused on causal effect estimates, they differ from group designs in how they generate causal effect estimates. As a result, SCD studies require a different review process with different specific standards from group designs. These standards will guide WWC reviewers in identifying and evaluating evidence from SCDs. If a study is eligible for review as an SCD, it is reviewed using the criteria described next to determine whether it receives a research rating of *Meets WWC Standards Without Reservations*, *Meets WWC Standards With Reservations*, or *Does Not Meet WWC Standards*. SCD studies may also contain more than one experiment, and each experiment should receive its own rating. See the section corresponding to each design sub-type for guidance.

Additional eligibility requirements for SCDs

The eligibility criteria for a WWC review of SCDs are as described in Chapter II, Screening Studies for Eligibility. That is, the study must be made publicly available, released within the 20 years preceding the review, use eligible populations, examine eligible interventions, and have eligible outcomes. In addition, studies that are eligible for review as SCDs are identified by the following features:

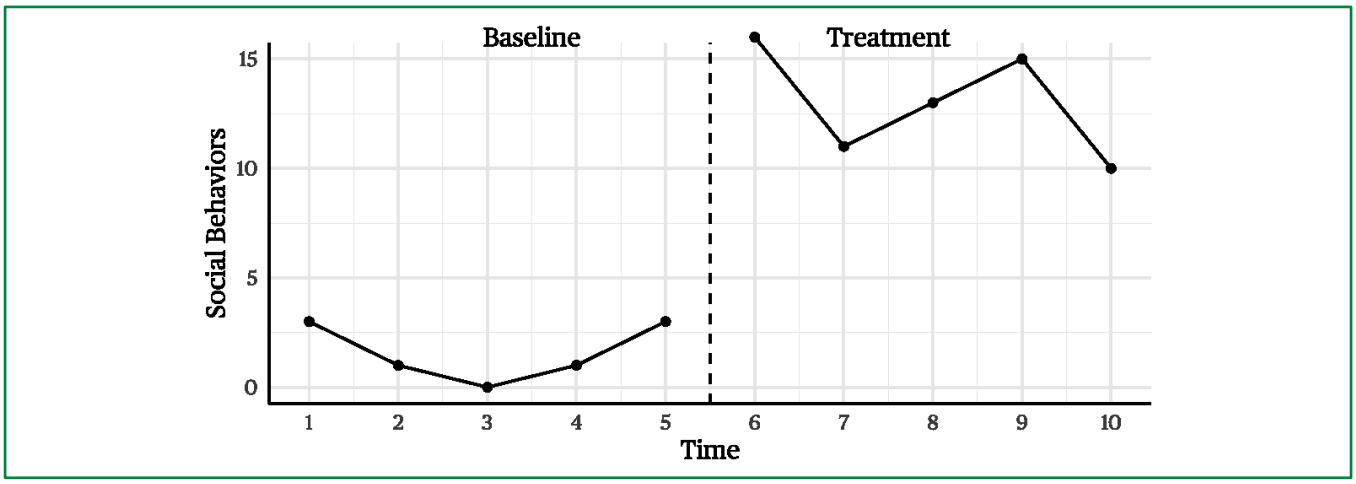
1. An individual case is the unit of intervention administration and data analysis. A case is most commonly a single participant. It also may be a cluster of participants, such as a classroom or school.
2. Within the design, the case can provide its own control for purposes of comparison. For example, the case's series of repeated outcome measurements prior to the intervention is compared with the series of repeated outcome measurements during and after receiving the intervention.
3. The outcome variable is measured *repeatedly* within and across different conditions. These different conditions are frequently structured as phases, such as the first baseline phase, first intervention phase, second baseline phase, and second intervention phase.

Figure 14 displays the simplest form of an SCD, a single individual (case) with one baseline phase and one treatment phase. This simple design is sometimes referred to as an AB design. In SCDs, a phase typically refers to a set of data points from the same condition, observed across time without the interruption of data points from a different condition. When phases are referred to using a string of capital letters, each letter represents a phase from a different condition. For instance, an ABC design would be a design with three phases. The A phase would typically be the baseline phase, the B phase would be an intervention phase, and the C phase would be another intervention phase, either in the form of a modified intervention, an alternative intervention condition, or a maintenance phase. An ABCABC design would be a six-phase design with three conditions that would begin like the ABC design described above, but in the fourth phase it would return to the original baseline A phase, then transition back to another B phase, and finally transition once again to another C phase. Some designs deviate from this typical structure and rapidly alternate interventions within the same experimental phase. See the alternating treatment design section for more information about these designs.

In the example in figure 14, the effect of the intervention can be conceptualized as the change in the outcome between the baseline phase (the A phase) and the treatment phase (the B phase). However, most SCD experts consider this simple form of the SCD to have weak internal validity because the effect of the intervention could be due to some other change that co-occurred with the intervention, such as developmental changes for the participant or changes that took place in the classroom that were unrelated to the intervention. Ensuring that SCDs guard against these threats is an important component of the WWC's SCD standards.

WWC standards apply to a wide range of SCDs, including reversal/withdrawal designs, multiple baseline designs, alternating and simultaneous intervention designs, changing criterion designs, and variations of these core designs like multiple probe designs. These designs, along with standards for combinations for these designs, are described in greater detail later.

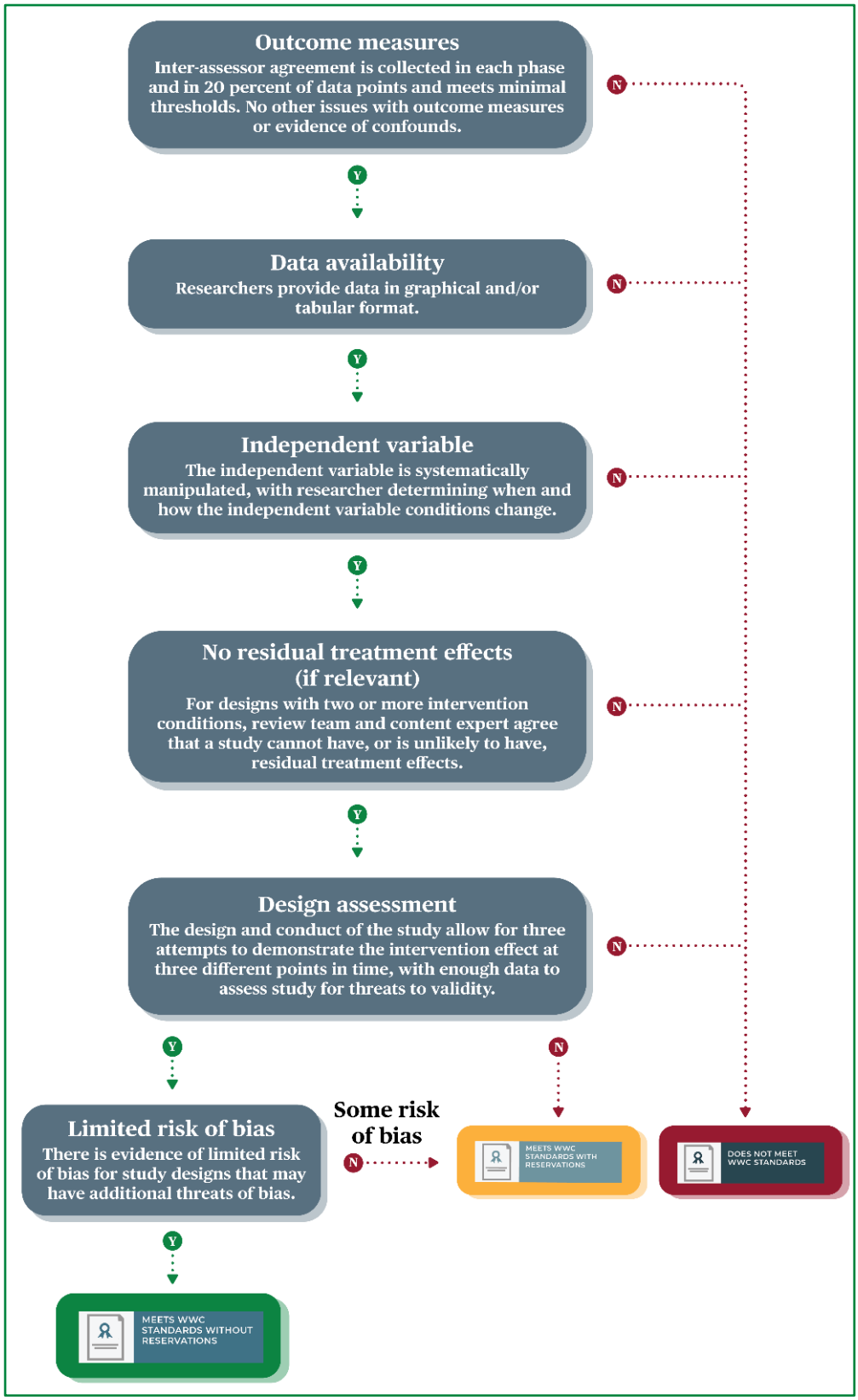
Figure 14. *Basic single-case design*



Reviewing findings from SCDs according to WWC standards

The process for reviewing SCD studies that are found eligible for review is presented in figure 15. After a study is found eligible for a WWC review as an SCD, the next step is the same as in other designs and includes reviewing the study's outcome measures and checking for confounding factors. If none of the outcome measures are consistent with the WWC's standards or if the study contains a confounding factor, the study will receive a research rating of *Does Not Meet WWC Standards* and the review will stop.

Figure 15. *Single-case design review process for eligible study findings*



Outcome measure standards

The WWC's outcome measure standards for SCDs are similar to those for randomized controlled trials (RCTs) and quasi-experimental designs (QEDs) described in Chapter III, Outcome measure standards, including (1) face validity, (2) reliability, (3) not overaligned with the intervention, and (4) consistent data collection procedures. Differences in these standards for SCDs are described next.

Standard 1: Face validity

The requirements for face validity are the same as those for group designs. To show evidence of face validity, an outcome measure must appear to measure what it claims to measure. To demonstrate face validity, a measure must have a clear definition of what it measures, such as a skill, an event, a condition, or an object, and assess that skill or event. For instance, a measure described as a test of reading comprehension that only assesses reading fluency does not demonstrate face validity.

Standard 2: Reliability

For a measure to demonstrate reliability, study authors must present evidence that the outcome has acceptably low levels of measurement error. In group designs, study authors typically report measures of internal consistency, temporal stability, or test-retest reliability. In SCDs, outcomes are most frequently direct observations of behavior. For these direct observation outcomes, the most applicable form of reliability is interassessor agreement, also known as inter-rater reliability or interobserver agreement. Although more than 20 statistical measures can represent interassessor agreement (for example, see Berk, 1979; Suen & Ary, 1989), commonly used measures include the percentage or proportional agreement and Cohen's kappa coefficient, which adjusts for the expected rate of chance agreement (Hartmann et al., 2004). Minimum acceptable values of interassessor agreement are at least .80 if measured by percentage agreement, and at least .60 if measured by Cohen's kappa (Hartmann et al., 2004).

To meet the WWC's interassessor agreement requirements for direct observation outcomes, the following criteria must be met:

1. The outcome variable must be measured systematically over time by more than one assessor for each case.
2. The study authors must collect interassessor agreement in each phase.
3. The study authors must collect interassessor agreement data for at least 20 percent of the data points.
4. The interassessor agreement must meet the minimum acceptable values for each outcome across all phases and cases (however, the interassessor agreement values are not required to meet minimum acceptable values separately for each case or phase). The raw data from the secondary assessor that were gathered for the purposes of interassessor agreement do not need to be reported. It is enough to report summary measures of interassessor agreement.

If a study contains measures that are not direct observations of behavior, such as a test of an academic outcome, then the reliability standards for these measures will follow the guidelines in Chapter III, Outcome measure standards.

Standard 3: Not overaligned

Overalignment occurs when an outcome measure contains content or materials provided to the cases in one condition but not another. This rule does not apply when material covered by an outcome measure must be explicitly taught, or when an outcome measure is broadly educationally relevant. Content experts can provide advice on whether an outcome has broad educational relevance. These two caveats to the overalignment standard are particularly important to SCDs, which frequently focus on narrow, specific outcomes that may require explicit teaching, or on daily-living outcomes with educational relevance. The functional skills domain from the *Study Review Protocol* contains examples such as dressing, preparing and eating food, or hygiene, where the researcher might teach the participant a checklist or a set of steps that need to be repeated and the outcome might be some measure of success at repeating the checklist or steps that were taught to the participant.

Standard 4: Consistent data collection procedures

Data must be collected in the same manner for the intervention and comparison conditions. If no information is provided, the WWC assumes that data were collected consistently. In the context of SCDs, the reviewer should ensure that the data collection procedures were similar across conditions for a given case. Reviewers should look for details indicating that data were collected in different modes, with different timing, or by using different personnel in the different conditions.

In terms of timing, the major concern is whether the data collection takes place at a different time of day between conditions. For instance, if all baseline data points are collected in the morning but the intervention data points all are collected in the afternoon, this would represent inconsistent data collection procedures. However, in many SCDs, the introduction of the intervention is staggered in a time-lagged fashion across participants in the design. Staggered introduction of an intervention that is an intentional element of the design does not represent an issue with inconsistent data collection procedures.

Additional consideration: Independence of outcome measure

The consideration for independence is unchanged for SCD designs. That is, in some outcome domains as specified in the *Study Review Protocol*, the WWC will consider whether the measure is independent of the intervention.

Confounding factors

A confounding factor occurs when a component of the research design is perfectly aligned with either the intervention or comparison condition, across all cases or phases of the experiment.

A factor that is aligned with a particular case—as opposed to condition—is not considered a confounding factor because any factor that is completely aligned with a single case will be present in all conditions of the study. The interventionist may be a confounding factor often observed in SCDs. Teachers, parents, or peers—collectively labeled *interventionists*—can administer the intervention to study cases. However, when all study cases experience a different interventionist across baseline and intervention phases of the study, the study has a potential confounding factor.

As it can sometimes be difficult to determine whether something is a confounding factor, the examples that follow describe situations for which the interventionist is and is not a confounding factor. Cases might have a different interventionist across the baseline and intervention phases, noted by underline in the examples below.

- *Example of a confounding factor*. One teacher teaches all cases in the baseline condition, and a different teacher teaches all cases in the intervention condition.

	Baseline	Intervention
Case 1	Teacher 1	Teacher 2
Case 2	Teacher 1	Teacher 2
Case 3	Teacher 1	Teacher 2

- *Example of a confounding factor*. One teacher teaches all cases in the baseline condition, and that same teacher and another teacher (or trainer) teaches all cases in the intervention condition.

	Baseline	Intervention
Case 1	Teacher 1	Teacher 1 + Teacher 2
Case 2	Teacher 1	Teacher 1 + Teacher 2
Case 3	Teacher 1	Teacher 1 + Teacher 2

There are similar-appearing circumstances that are <u>not</u> confounding factors.

- *A nonexample of a confounding factor*: One teacher teaches all cases in both phases.

	Baseline	Intervention
Case 1	Teacher 1	Teacher 1
Case 2	Teacher 1	Teacher 1
Case 3	Teacher 1	Teacher 1

- *Nonexample of a confounding factor*: Multiple teachers teach different cases; teachers do or do not teach different phases.

	Baseline	Intervention
Case 1	Teacher 1	Teacher 1
Case 2	Teacher 2	Teacher 2
Case 3	Teacher 3	Teacher 3

Additional confounding factors include contextual or procedural changes between conditions that might affect outcome measurement and participant responding and are not a component of the intervention of interest. A nonexhaustive list of examples of confounding factors include shifts in experimental session length, changes in the quantity or quality of reinforcement provided to the participant, changes in the environment or setting (for example, one condition takes place in the classroom while another phase takes place on the playground), or changes in the social context that affect opportunities to respond.

Review process for eligible findings from SCDs

To be considered potential evidence of an intervention's effectiveness by the WWC, an SCD must meet four standards: data availability, researcher-manipulated independent variable, no residual treatment effects, and design assessment. To receive a research rating of *Meets WWC Standards Without Reservations*, findings based on some designs must meet requirements for limiting sources of bias. These standards are summarized in figure 15 and detailed in the following sections.

Data availability

SCD study authors need to provide raw data in graphical or tabular format for their findings to meet WWC standards. Graphical or tabular data must present the raw data that corresponds to the individual observation sessions. Summary data, such as the within-phase mean for each phase, are not sufficient to meet this requirement.

Data sharing in the form of plots is standard practice in SCD research. Sharing data allows other researchers to perform their own reanalysis in the form of visual analysis, and access to the raw data allows the WWC to assess whether the study meets WWC standards of internal validity for SCDs, as well as allow for effect size estimation when appropriate. If the data are not available in graphical or tabular format, then the study will receive a research rating of *Does Not Meet WWC Standards.*[30]

Researcher-manipulated independent variable

The researcher must determine the time at which an individual case transitions between phases or conditions. Researcher control over the timing of the intervention is a crucial element of why these designs can be considered experimental designs and are potentially eligible to receive the WWC's highest rating.

To meet this requirement, there must be evidence that the independent variable was systematically manipulated by the researcher. Although the researcher may operate in consultation with other individuals involved in the conduct of the study, such as parents, teachers, or school administrators, the final choice of when the independent variable conditions change must rest with the researcher. Randomization designs or masked visual analysis designs are considered to have satisfied this requirement, so long as the researcher made the final choices around the procedures or decision rules for phase transitions.

If the study does not discuss who manipulated the independent variable, but there is no evidence that it was someone other than the researcher, then reviewers should assume that this standard was met. If there is evidence that someone other than the researcher manipulated the independent variable, then the finding will receive a research rating of *Does Not Meet WWC Standards*.

[30] When data are available only in graphical format, the WWC will extract tabular data from the plots. The WWC intends to explore ways of making the extracted tabular data available for researchers interested in performing confirmatory analyses or syntheses of findings from studies reviewed by the WWC.

Residual treatment effects

Residual treatment effects are a potential confound in designs with more than one intervention. Alternating treatment designs and other SCDs with an intervening third condition are potentially subject to residual treatment effects. When there are two or more interventions in the intervention phase of an alternating treatment design, the reviewer must examine the study to ensure that there is limited risk of residual treatment effects.

Residual treatment effect

Residual treatment effect refers to the form of carryover where the effects of one intervention spill over into observation sessions for a separate intervention or experimental conditions being observed within the same design.

When a review team identifies an eligible alternating treatment design experiment that uses two or more interventions, the review team should ask a content expert to assess whether residual treatment effects are likely given the specific interventions or experimental conditions, the timing of the and length of observation sessions, the order of the interventions or experimental conditions, and the outcomes in the experiment. The review team can rely on previous approval of similar conditions and outcomes from a content expert, but the plausibility of residual effects should not be solely informed by the data reported in a given study. The review team will then assign the study for review and pass along the content expert's determination to the reviewers. Reviewers should raise any additional concerns they have about residual treatment effects as part of their reviews. Reviewers should focus on the plausibility of residual treatment effects based on theoretical and contextual considerations given the research design and intervention characteristics but should not raise concerns based on data reported in the study.

If a content expert and reviewer both agree that residual treatment effects likely exist, then the finding is rated *Does Not Meet WWC Standards* because the measures of effectiveness cannot be attributed solely to the intervention. If the content expert and reviewer disagree, then review team leadership should revisit the issue. If the content expert and reviewer both agree that residual treatment effects are unlikely, then the reviewer should complete the review assuming there are no residual treatment effects.

Reversal/withdrawal, multiple baseline, and multiple probe designs generally have longer phases and a longer time between data points than alternating treatment designs. More time will pass between the noncontiguous phases that will be compared (for example, between the first B and second A in an ABCAB reversal/withdrawal design); this feature may make residual treatment effects less important even if they are present. If the reviewer and content expert agree that residual treatment effects are unlikely, or are unlikely to be meaningful, then the reviewer should work with the review team leadership and content experts to identify how best to proceed with the review, focusing only on the intervention of interest and the relevant comparison condition when assigning a research rating–that is, ignoring any third or fourth interventions. If a study finding is judged to have a reasonable likelihood of residual treatment effects, then the finding is rated *Does Not Meet WWC Standards*.

Research design requirements

The primary goal of the SCD research design requirements is to ensure that the study was designed in a way that allows for at least three demonstrations of an intervention effect at three different points in time with reasonable certainty that the observed data are sufficient to capture important information about the pattern of responding.

The pattern of responding includes information such as the within-phase mean, the within-phase variability, and any increasing or decreasing trend that might be present.

The three demonstrations criterion is based on professional convention (Horner et al., 2012). In practice, this means that there must be at least three phase changes between the two conditions being compared within a review, which occur at three different points in time. For reversal/withdrawal designs, this will be at least three phase changes within a case. Three phase changes requires that a case has at least four total phases. For a multiple baseline or multiple probe design, this would be at least three tiers with phase changes at three different times.

This standard includes design-specific conventions regarding the number of phases and data points per phase required to meet the three demonstrations criterion. Specific variations of SCDs have additional requirements. These requirements are intended to ensure that the study is designed in a way to support at least three opportunities to demonstrate an intervention effect at three different points in time.

When there are a sufficient number of opportunities to demonstrate an intervention effect, but a limited number of data points are available, a study will receive a lower rating. However, in some cases there are a small number of data points but it is still possible to be reasonably certain that the observed data points capture all the necessary information about the phase.

Any phase with three or more data points and no within-phase variability represents enough data that reviewers can be reasonably certain that additional data points would not provide additional information. Figure 16 provides some example baseline phases with both zero and low variability. Although data points at the scale minimum or scale maximum are not the only contexts in which there might be zero within-phase variability, these are likely the most common scenarios where zero within-phase variability will be observed.

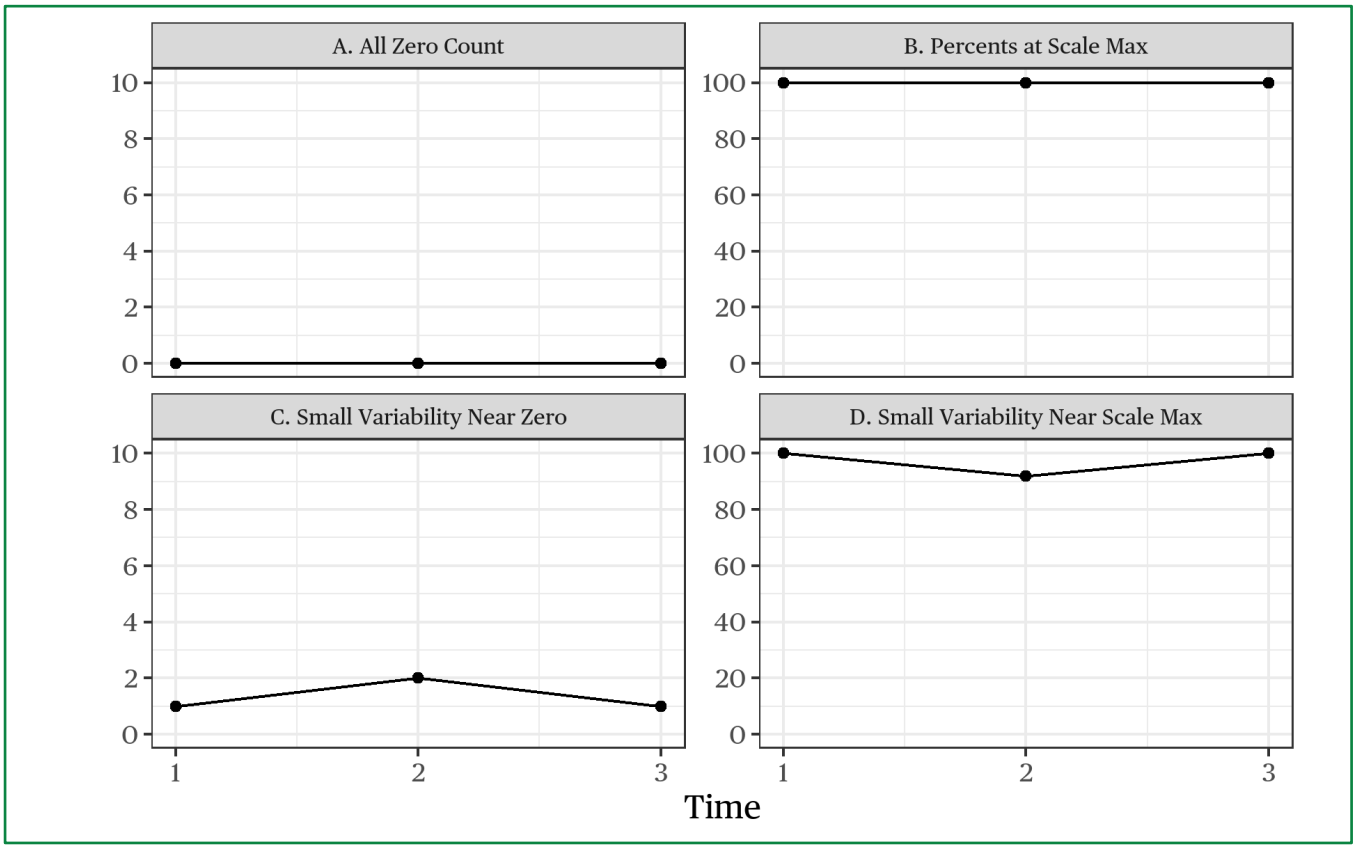
Panels A and B show plots with three data points at the scale minimum and scale maximum, respectively. There is no variability within each phase, and reviewers can be reasonably certain that if the researcher had gathered another data point, it would contain no additional information about the average level or variability within the phase.

Panels C and D show examples where there is relatively little variability near the scale minimum or scale maximum. The data points within a phase are similar but do have some variability. There can be less confidence about what the exact value of an additional data point might be and how much variability there is within each phase.

Throughout the rest of the WWC standards for SCDs, there are occasions where phases with three or more data points will allow a finding to be rated *Meets WWC Standards Without Reservations* despite otherwise not having enough data points to meet the described requirements. These occasions are described in the sections regarding treatment reversal/withdrawal designs and multiple baseline/multiple probe designs, and then in a later section on therapeutic baseline trends.

When the data are available as a table in the study or from the study authors, reviewers should examine the table to observe whether short phases of three or more data points have exactly the same value for all data points, and therefore meet this requirement. When the raw data are not available as a table or from the study authors, a visual inspection that confirms unambiguous zero variance is sufficient to meet this requirement. In other words, if the reviewer believes the authors intended to convey that all the data points within phase had the exact same value, this is sufficient to meet the requirement.

Figure 16. *Zero- and low-variability baseline examples*



Reversal/withdrawal designs (AB^K)

Phases. Findings from treatment reversal/withdrawal designs must have a minimum of two phases per condition to be eligible to be rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*. In the simplest design that compares a baseline condition with an intervention condition, this will require four phases. Any case with fewer than four phases and at least two phases per condition will be eligible to receive a research rating of *Does Not Meet WWC Standards*.

Data points per phase. For treatment reversal/withdrawal design findings to be eligible to be rated *Meets WWC Standards Without Reservations*, the first baseline phase must have at least six data points,[31] and at least two

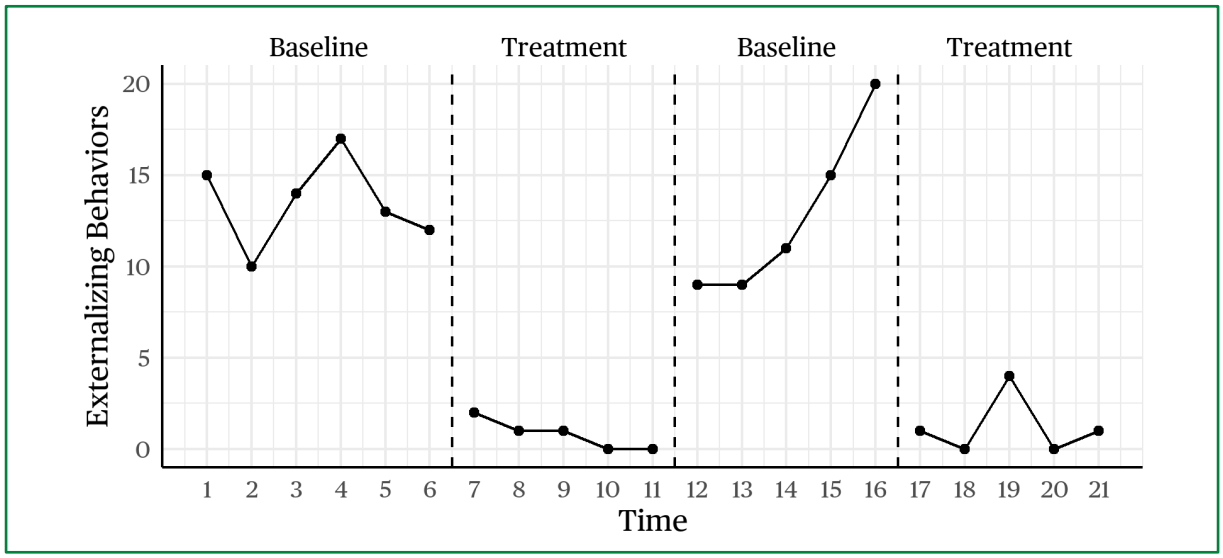
[31] A small group of applied and methodological SCD experts agreed that requiring more data in the baseline phase to ensure that reviewers could make accurate judgments was a reasonable requirement.

phases per condition must have five or more data points per phase. In the simplest design that compares a baseline and intervention condition, this corresponds to an initial baseline phase with six or more observations, a treatment phase with five or more observations, a return-to-baseline or withdrawal phase with five observations, and a second treatment phase with five or more observations. Additionally, any phase with three or more data points and zero within-phase variability also will count toward the required phases for a finding to be eligible to be rated *Meets WWC Standards Without Reservations*, including initial baseline phase.

For treatment reversal/withdrawal design findings to be eligible to be rated *Meets WWC Standards With Reservations*, two phases per condition must have three or more data points per phase. Findings that do not meet either set of requirements will receive a research rating of *Does Not Meet WWC Standards*.

Figure 17 provides a simple reversal/withdrawal design example eligible to receive a research rating of *Meets WWC Standards Without Reservations* due to meeting the number of phases and number of data points per phase requirements for the reversal/withdrawal design standards. Each participant- or unit-outcome combination in a reversal/withdrawal design is a single experiment, and therefore it should receive separate ratings.

Figure 17. *Reversal/withdrawal design example*



Changing criterion designs. The reversal/withdrawal design standards can be applied to changing criterion designs, with a small modification. A changing criterion design is similar in structure to the reversal/withdrawal design, except that each phase change after the initial baseline to treatment phase represents a modified treatment phase with an incremental goal, or criterion, for the behavior of interest. Each baseline or intervention change or criterion change should be considered a phase change. As such, there should be at least three different criterion changes to establish three attempts to demonstrate an intervention effect. In some studies that use a changing criterion design, the researcher may reverse or change the criterion back to a prior level to further establish that the change in criterion was responsible for the outcomes observed on the dependent variable. This should be considered a phase change, as in the reversal/withdrawal design.

Multiple baseline/multiple probe designs

Phases. Multiple baseline designs must have a minimum of six phases split into two conditions for their findings to be rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations.* The simplest example of this design has three tiers stacked vertically. Each tier is made up of a baseline and treatment phase (in other words, an A to B comparison). Additionally, transitions from the baseline phase to the intervention phase must have at least three unique timings to ensure that there are three opportunities to demonstrate the intervention effect at three different points in time. Findings with fewer than six phases will be rated *Does Not Meet WWC Standards*.

Data points per phase. For findings from multiple baseline designs to be eligible to be rated *Meets WWC Standards Without Reservations*, the first baseline phase within each tier must have at least six data points. Additionally, all subsequent phases must have five or more data points per phase. Any phase with three or more data points and zero within-phase variability also will count toward the required phases for a finding to be eligible to be rated *Meets WWC Standards Without Reservations*, including the first baseline phase in each tier.

For multiple baseline design findings to be eligible to be rated *Meets WWC Standards With Reservations,* three phases per condition must have three or more data points per phase. Findings from multiple baseline designs that do not meet either set of requirements will be rated *Does Not Meet WWC Standards.*

Concurrence. The timing of the design's implementation requires a degree of concurrence across the design. The concurrence requirement encompasses three elements:

1. Tiers must be organized to allow for vertical comparison. This means that data points at time 1 for every tier must take place prior to all data points at time 2 for every tier, and all data points at time 2 for every tier must take place prior to all data points at time 3 for every tier, and so on (Slocum et al., 2022). Reviewers should assume this standard is met, unless authors provide evidence of nonconcurrence, such as describing the design as a nonconcurrent multiple baseline or graphing data in a way that suggests nonconcurrence.
2. All tiers must have data collected in the baseline phase prior to the introduction of intervention to any case.
3. Cases that have not yet received the intervention must have data at or after the time another case enters the intervention.
    - **If appropriate for the design, training phase data must be present**. Some interventions require that the participant be trained in the intervention. The requirement for training will be discussed by the authors if training is necessary. Studies that do not discuss training need not meet the training data requirements. If the effect of the intervention is expected to be immediate at the onset of training, then data for the training phases must be present for every tier and can be considered part of the intervention. If the intervention effect is not expected until after the completion of the training, then tiers still in the baseline phase must continue baseline measurement at or after the time point when a preceding tier has the first intervention probe after completing training. This process prevents an overlap in the training/intervention phase for any two tiers and allows for cases that have begun to receive the full effect of intervention to be compared vertically with those cases still in the baseline.

Findings from any multiple baseline or multiple probe design that fail to meet the concurrence requirement will be rated *Does Not Meet WWC Standards.*

Tiers in multiple baseline designs

A tier refers to a single row in a set of stacked rows in a multiple baseline design. In figure 18, tiers are cases. In that example, the comparisons between baseline and intervention cases can be made within and across cases. Tiers might also be outcomes or contexts within a single individual or case, and effect replication takes place purely within a single case.

Figure 18 provides an example of a multiple baseline design that is eligible to receive a research rating of *Meets WWC Standards Without Reservations* due to meeting the number of phases, number of data points per phase, and concurrence requirements for multiple baseline designs. Although the *y*-axis of the plot does not specify the exact date or time of the observations, the WWC generally assumes that authors have aligned their displays in a way that allows for vertical comparison, absent any evidence to the contrary. Figure 19 provides an example with evidence to the contrary. In multiple baseline and multiple probe designs, each stacked plot generally represents a single experiment. Each combination of stacked plot and outcome should receive a separate rating.

Figure 18. *Multiple baseline design example*

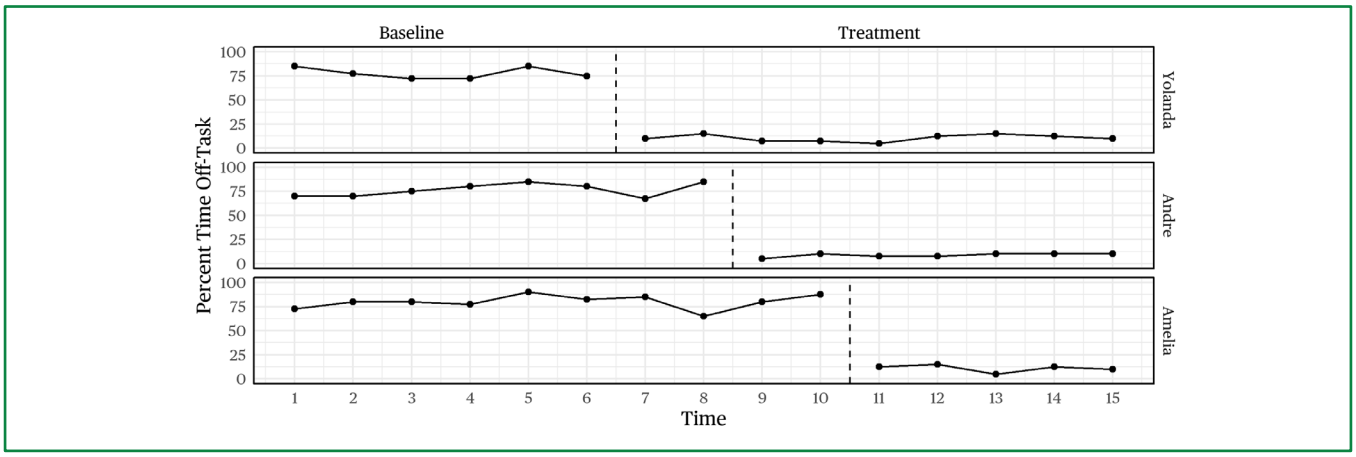


Figure 19 displays an example of a multiple baseline design that does not allow for vertical comparison and fails the requirement that all cases must have data in the baseline phase prior to the introduction of the intervention to any case. Although the first data points for Yolanda, Andre, and Amelia are all arranged in a stacked fashion, the actual timing of those data points are in five-day intervals from each other. Additionally, data points for Amelia do not begin until halfway through Yolanda's treatment phase. This example would be rated *Does Not Meet WWC Standards.*

Figure 19. *Example violations of first and second concurrence requirements*

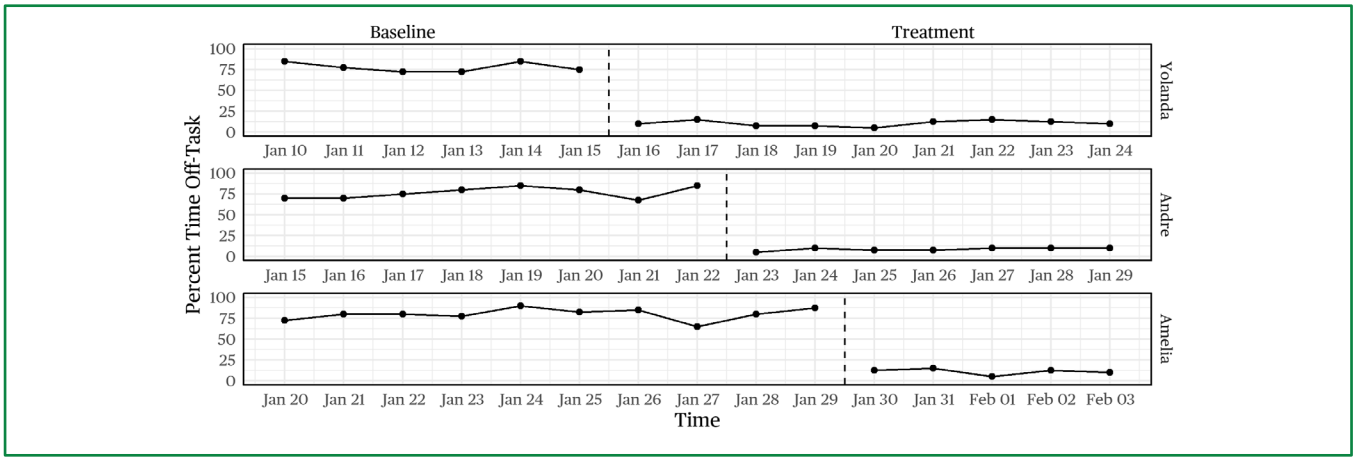


Figure 20 displays an example of multiple baseline design that would not meet the third concurrence requirement. In this example, Andre's baseline ends after time 6, prior to the onset of the intervention for Yolanda. No data allow for a vertical comparison to ensure that there is no change in Andre's responses prior to the onset of intervention. The final data points in the baseline phase prior to the onset of the intervention are important for judging any change in the trajectory of the outcome data points, including the WWC's baseline trend requirement. This example would be rated *Does Not Meet WWC Standards.*

Figure 20. *Example violation of the third concurrence requirement*

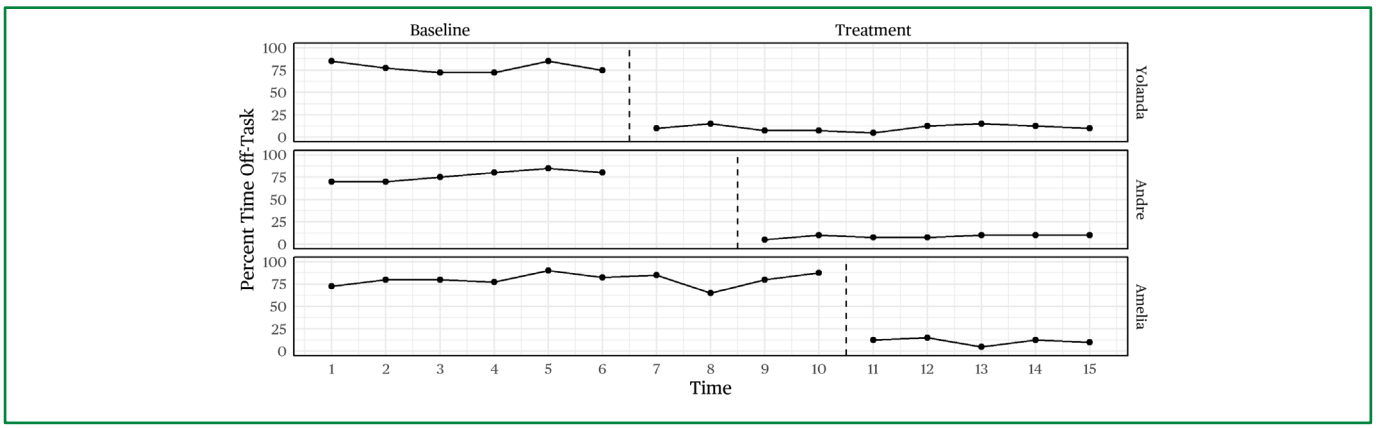
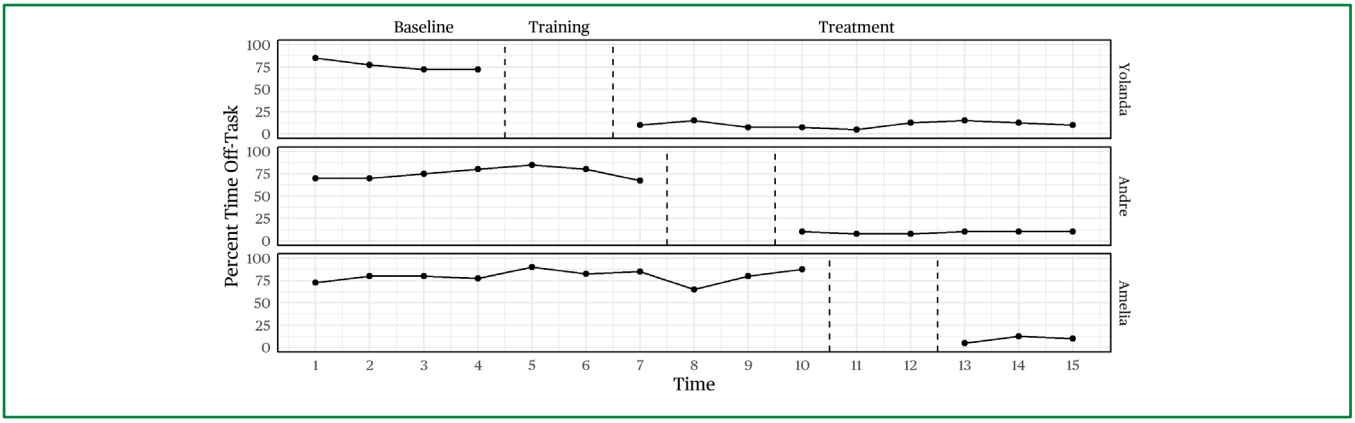


Figure 21 displays an example of a multiple baseline design with empty training phases. These empty training phases are not appropriate for interventions where the training is quick and impact is expected to be immediate. For those types of interventions, the training data would represent the beginning of the treatment phase from the WWC's perspective and therefore are important to include as a part of the impact estimate. In those cases, this design would be rated *Does Not Meet WWC Standards.* For interventions that require a longer training period to have an impact, empty training phases are acceptable. The requirements for concurrence should ignore the training phase and focus on overlap between the baseline and treatment phases. In cases where the longer training is appropriate, this design would be eligible to receive a rating of *Meet WWC Standards With Reservations.* However, in cases where the training phases last as long or longer than the longest treatment phase in a design, the reviewer

should consult a content expert to ensure that the training phases do not constitute an extra condition where data should be available.

Figure 21. *Example of empty training phases*



Multiple probe design requirements. These designs are a special case of multiple baseline designs. *Planned* missing data is a key element of the multiple probe design and is the major difference between a multiple baseline design and a multiple probe design. Multiple probe designs must meet **all** the multiple baseline design requirements **and** additional criteria because baseline data points are intentionally missing.[32] For multiple probe designs to meet WWC standards, the following must be true:

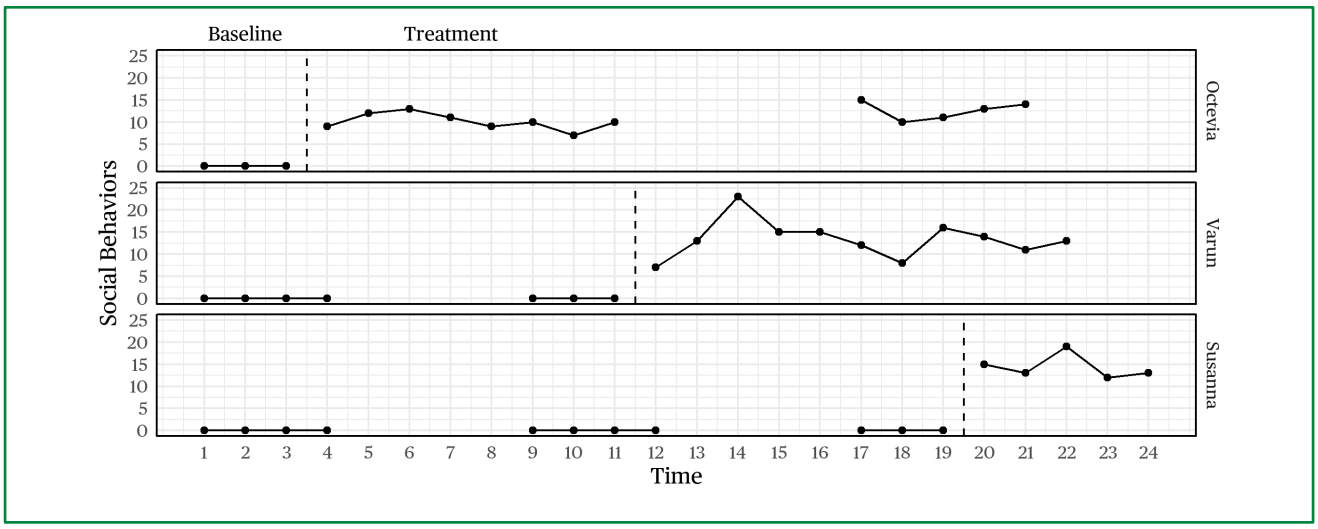
- **Initial preintervention data collection sessions must overlap**. For findings to receive a research rating of *Meets WWC Standards Without Reservations,* each tier must have three data points in the first three sessions. For findings to receive a research rating of *Meets WWC Standards With Reservations,* there must be at least one session within the first three sessions where probe points overlap vertically for all tiers in the design.
- **Probe points must be available just prior to introducing the independent variable.** Within the three sessions just prior to introducing the independent variable, the design must include three consecutive probe points for each case to be rated *Meets WWC Standards Without Reservations* and at least one probe point immediately preceding the onset of intervention for each case to be rated *Meets WWC Standards With Reservations*.
- Each case not receiving the intervention must have a probe point in a session where another case either first receives the intervention or reaches a prespecified intervention criterion described by the researchers.
    - For designs with a training phase, when impacts are expected only after complete delivery of training, the "first receives the intervention" language should be interpreted as the time point when a case has the first intervention probe after completing their training.

Findings from multiple probe designs that fail to meet any of these requirements **in addition to the general multiple baseline design requirements** will receive a research rating of *Does Not Meet WWC Standards*.

[32] Multiple baseline designs with unintentional missing data should not be reviewed under the multiple probe requirements. Reviewers should note any unplanned missing data in the study review guide.

Figure 22 displays an example of a multiple probe design that would be eligible to be rated *Meets WWC Standards Without Reservations*. The initial three data collection sessions have data for all cases. All cases have at least three data points directly before intervention begins for any other case. Each case still in the baseline has a data point when the other case first receives the intervention. The baseline phases are all phases with zero variability, and as a consequence three data points are enough data to be rated *Meets WWC Standards Without Reservations.*

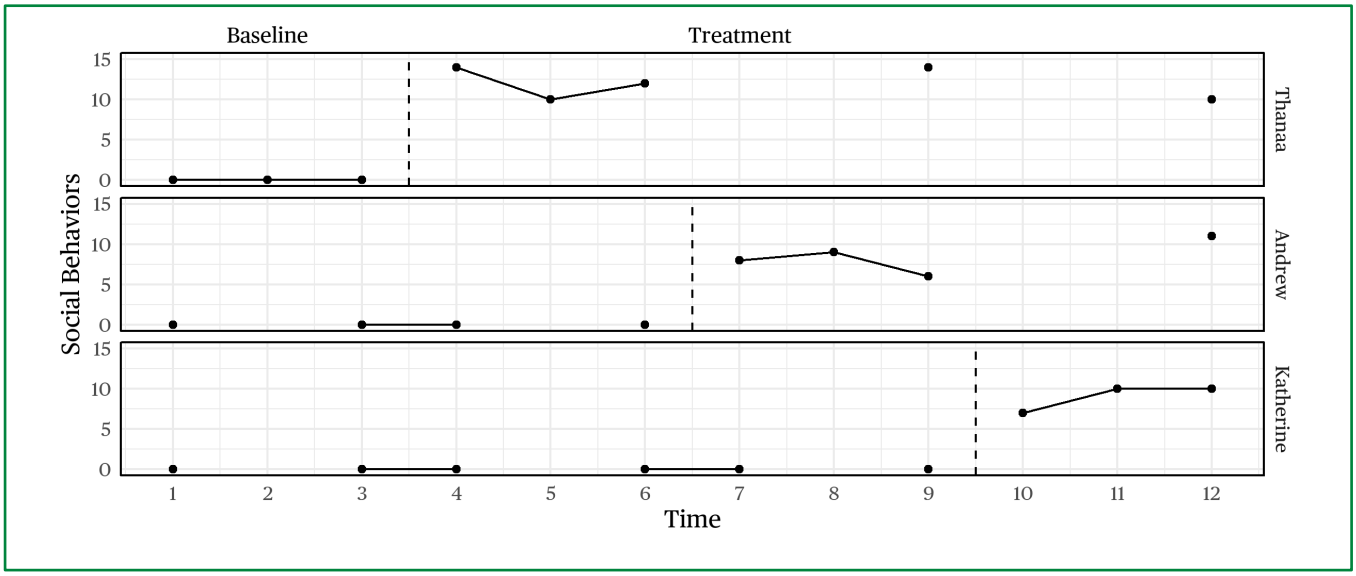
Figure 22. *Multiple probe design, example 1*



Note that data collection after the onset of the intervention in a multiple probe design may be intermittent or continuous. The WWC has no specific requirements for the intervention phase other than the requirements for a minimum number of data points per phase.

Figure 23 displays an example of a multiple probe design that would be potentially eligible for a research rating of *Meets WWC Standards With Reservations*. Each case has a single overlapping data point at time 1 and at time 3, but the study does not have the data point at time 2 for Andrew or Katherine. Each case has at least one data point in the three sessions prior to Thanaa receiving the intervention. Andrew and Katherine each have at least one data point in the three sessions prior to Andrew receiving the intervention. Katherine has at least one data point in the three sessions prior to receiving the intervention. Each case still in the baseline phase has a data point when another case enters the intervention.

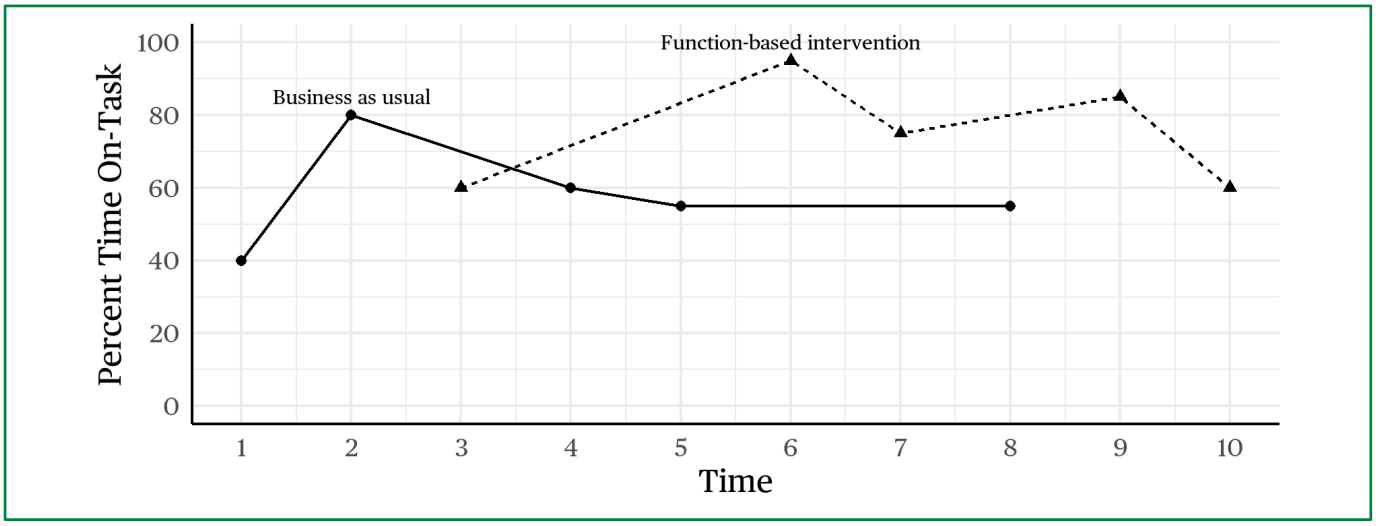
Figure 23. *Multiple probe design, example 2*



Alternating treatment designs

Figure 24 displays an alternating treatment design example that would be eligible to be rated *Meets WWC Standards Without Reservations* due to meeting the data points per condition and contiguous data points requirement for alternating treatment designs.

Figure 24. *Alternating treatment design example*



Some alternating treatment designs will contain both a baseline phase and a phase that rapidly alternates between two or more conditions, and other designs will contain only a set of rapidly alternating conditions as seen in Figure 24. In the context of an alternating treatment design, the three demonstrations can take place between a single baseline and the intervention of interest at three different points in time during the rapidly alternating phase; between a business as usual condition and an intervention condition within the rapidly

alternating phase; or some combination of the two. This reflects the fact that assessment of the intervention effect in most applied work compares conditions holistically across all data points within a design.

An important consideration exists when designs include multiple intervention comparisons—for example, A versus B, A versus C, C versus B. The WWC considers each comparison between conditions as a separate contrast. Accordingly, each contrast should be reviewed for eligibility and research rating separately. Although the design refers to "alternating treatments," the rapidly alternating phase can contain a business-as-usual condition. Contrasts containing a business-as-usual condition will most frequently be the findings of interest for the WWC.

Data points per condition. Findings from alternating treatment designs must have at least five data points per condition to be rated *Meets WWC Standards Without Reservations*. Designs must have at least four data points per condition for their findings to be rated *Meets WWC Standards With Reservations*. Any findings based on fewer data points will result in the research rating of *Does Not Meet WWC Standards*.

Contiguous data points. Within a phase involving the rapid alternation of treatments, there should be a maximum of two sequential data points of the same intervention condition without the interruption of another condition. Any comparison with more than two contiguous data points in the rapidly alternating phase without the interruption of another condition shall receive a research rating of *Does Not Meet WWC Standards.*

Some designs will continue to gather data on the intervention or condition deemed most successful after the completion of rapid alternation. More than two contiguous data points examining the most successful intervention after the rapid alternation ends will not be considered a violation of this requirement.

Additionally, designs that use an unrestricted randomization procedure to assign condition order are exempt from the contiguous data points requirement. The design will still need to allow for three demonstrations of the intervention effect at three different points in time in order to be eligible to be rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations.*

Other SCD designs

SCDs that use methodology not currently described in the *Handbook* can still meet the WWC research design requirements. Review team leadership must document and use published professional conventions for the research design under review. For an SCD design not currently described in the *Handbook* to be rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations,* it must contain three attempts to demonstrate an intervention effect at three different points in time.

Limited risk of bias

This section is relevant for designs where the primary comparison is for the pattern of responding between separate phases, such as baseline and intervention phases. Of the designs for which the WWC has explicit standards, this includes treatment reversal/withdrawal designs, changing criterion designs, multiple baseline designs, and multiple probe designs. The only design with explicit standards that is not subject to the assessment of limited risk of bias is the alternating treatment design; the contrast of interest is not strictly confined to between-phase comparisons. Any other designs not explicitly listed in the design standards but reviewed under a

set of published professional conventions will be subject to an assessment of limited risk of bias if the design's primary contrast of interest is a comparison between separate phases.

This section is relevant only to those designs that are potentially eligible to receive a research rating of *Meets WWC Standards Without Reservations* after being reviewed under the data availability, independent variable, residual treatment effects, and design assessment standards. Designs that limit the risk of bias are eligible to receive a research rating of *Meets WWC Standards Without Reservations.* Designs with a potential risk of bias are eligible to be rated *Meet WWC Standards With Reservations.* Presently, the potential risk of bias that the WWC assesses is related to therapeutic baseline trend and a lack of reversibility. Baseline trend is an important consideration in designs where the primary comparison is between baseline phases and treatment phases, such as reversal/withdrawal designs, multiple baseline designs, or multiple probe designs. The presence of a trend in the direction of the expected treatment effect in the initial baseline phase(s)—that is, a therapeutic trend—is of particular concern. If there is notable improvement in the outcome across the initial baseline phase or in the data points just prior to the onset of an intervention phase for a case, or just prior to the onset of an intervention for a preceding case in designs like the multiple baseline design, then there is some ambiguity around whether intervention effects can be attributed solely to the intervention and not some intervening factor such as individual maturation or other events within a classroom or intervention context.

Reversibility is an important consideration in reversal/withdrawal designs where several baseline or business-as-usual phases are alternated with treatment phases. In the most credible forms of these designs, both the intervention and outcome must allow for the outcome to return to initial baseline levels when the intervention is withdrawn. If some form of learning takes place during the intervention, then outcomes assessed in the return-to-baseline phases may not fully return to initial baseline levels, and the contrasts involving those return-to-baseline phases will be attenuated. Incomplete reversibility does not mean that a study cannot serve as evidence for an intervention's effectiveness; it simply may not be the highest quality evidence and therefore *Meets WWC Standards With Reservations* is the highest research rating that a finding with incomplete reversibility can receive.

The WWC has identified the use of a quantitative nonoverlap measure as an appropriate method to assist with judgments of baseline trend and reversibility.

Nonoverlap measures

Nonoverlap measures were created to help researchers describe the proportion of data in an intervention phase that demonstrates improvement over a baseline phase. Research has shown that nonoverlap measures are broadly consistent with visual analytic judgments (Parker et al., 2014). Their integration is intended to allow for a review process that is broadly consistent with visual analytic judgments. Assessing nonoverlap involves examining the distribution of data points across phases. Nonoverlap of data points provides evidence that a change has occurred. For example, if behavior counts range from 1 to 4 in a baseline phase and from 5 to 9 in an intervention phase, then there is 100 percent nonoverlap in data points across phases. This result may be taken as evidence that behavior changed from baseline to intervention in the context of the SCD research designs previously described.

While nonoverlap measures have traditionally been used to describe the effects of interventions, they also can be used to describe the similarity between two sets of data points. For instance, if there is no therapeutic trend in a

baseline phase, then the data points at the end of the phase should be similar to the data points at the beginning of the phase. If a behavior is reversible, the pattern of responding in any withdrawal or return to baseline phases in a treatment reversal/withdrawal design should be similar to the initial baseline phase. Higher levels of nonoverlap are associated with less consistency between the data points at the beginning of the phase and the data points at the end of the phase, implying the presence of a trend in the data or incomplete reversibility.

The WWC selected the nonoverlap of all pairs Parker and Vannest (2009) to help reviewers make judgments regarding baseline trend and reversibility because unlike many other nonoverlap indices, the magnitude of the index is not a function of the number of data points in a phase (Pustejovsky, 2019). The SCD standards contain benchmark values of the nonoverlap of all pairs for WWC reviewers to identify problematic instances of baseline trend and reversibility. These benchmarks represent the maximum acceptable values of the nonoverlap of pairs. Lower values represent a larger degree of overlap, where overlap is data points with values opposite the intended direction of the effect.

Individual design types that use the nonoverlap of all pairs contain general guidance on the use of the nonoverlap of all pairs. Details of the calculation can be found in appendix I in the technical appendices.

Minimal therapeutic baseline trend

For research designs in which the finding contrasts baseline phases and intervention phases, and that have findings eligible to receive a rating of *Meets WWC Standards Without Reservations,* reviewers should assess any initial baseline phases to ensure that there is minimal therapeutic trend. Designs with more than one baseline phase (such as a treatment reversal design) do not require assessment for baselines phases after the first.

Reviewers should assess baseline trends comparing the last three data points with all other data points within the initial baseline phase. A nonoverlap of all pairs of .85 or smaller will be considered evidence of minimal baseline trends.[33] Any baseline phase with at least three data points and zero within-phase variability will be assumed to have satisfied this requirement. Any finding that fails to meet this requirement is still eligible to receive a research rating of *Meets WWC Standards With Reservations*. For multiple baseline designs or multiple probe designs, all baselines within the design will be subject to this requirement, and failure will cause the entire design to be rated *Meets WWC Standards With Reservations*. Designs with more than the minimum number of cases might still be eligible to be rated *Meets WWC Standards Without Reservations* if an eligible subset meets the baseline trend requirement, as described in the section regarding designs with extra cases below.

Evidence of reversibility

For research designs with return-to-baseline or withdrawal phases and findings that are eligible to receive a rating of *Meets WWC Standards Without Reservations,* reviewers should assess any return-to-baseline or withdrawal phases compared with the initial baseline phase to ensure that minimal reversibility was achieved.

[33] The WWC arrived at the criterion of .85 in consultation with applied and methodological SCD experts. The .85 criterion is relatively arbitrary and novel. It is intended to be a low bar that only will reduce the rating of studies with egregious design issues. The WWC is committed to monitoring the consequence of the baseline trend and the following reversibility rule, and to making changes, additions, or deletions to this section in the future to continue to align the WWC standards with the practices of applied researchers.

Simple multiple baseline designs or multiple probe designs without embedded reversals are not subject to this requirement, nor are alternating treatment designs.

Reviewers should assess the reversibility of the outcomes using the nonoverlap of all pairs to compare the baseline and any return to baseline. A nonoverlap of all pairs of .85 or less will be taken as evidence of achieving at least minimal reversibility. Any finding that fails to meet this requirement is still eligible to receive a research rating of *Meets WWC Standards With Reservations*.

Designs with extra cases/phases or combination designs, or designs with extra conditions

Extra cases or phases. Reversal/withdrawal, multiple baseline, and multiple probe designs may have more than the minimum required number of phases, cases, or tiers required to meet standards. For example, a reversal/withdrawal design could have six phases (ABABAB), or a multiple baseline design could have four cases where each case has two phases. In general, as long as there are a sufficient number of phases, cases or tiers, and data points for a study finding to be rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations* to meet the minimum requirements for the design, additional phases, cases, or tiers with fewer than the required number of data points will not cause a study to receive a lower rating.

In addition, any finding should receive the highest rating that any subset of its design is eligible for, and those subsets need not be sequential to qualify. There are two important caveats. First, the subset must still contain three opportunities to demonstrate an intervention effect at three different points in time. Second, nonsequential phases within a case should not be compared with each other, and so may not constitute a demonstration of an intervention's effectiveness. Design subsets also must meet any other design-specific requirements for the design type contained in that subset.

Figure 25 displays an ABABAB reversal/withdrawal example with more than the minimum number of phases to meet standards. The first A phase contains six data points and the subsequent three phases contain five data points each, but the final two phases (the last AB pair) contain only two data points each. This study finding would still be eligible to receive a research rating of *Meets WWC Standards Without Reservations* because it contains enough information to potentially demonstrate the intervention effect at three different points in time within the first four phases.

Figure 25. *Treatment reversal design with extra phases that is rated* Meets WWC Standards Without Reservations

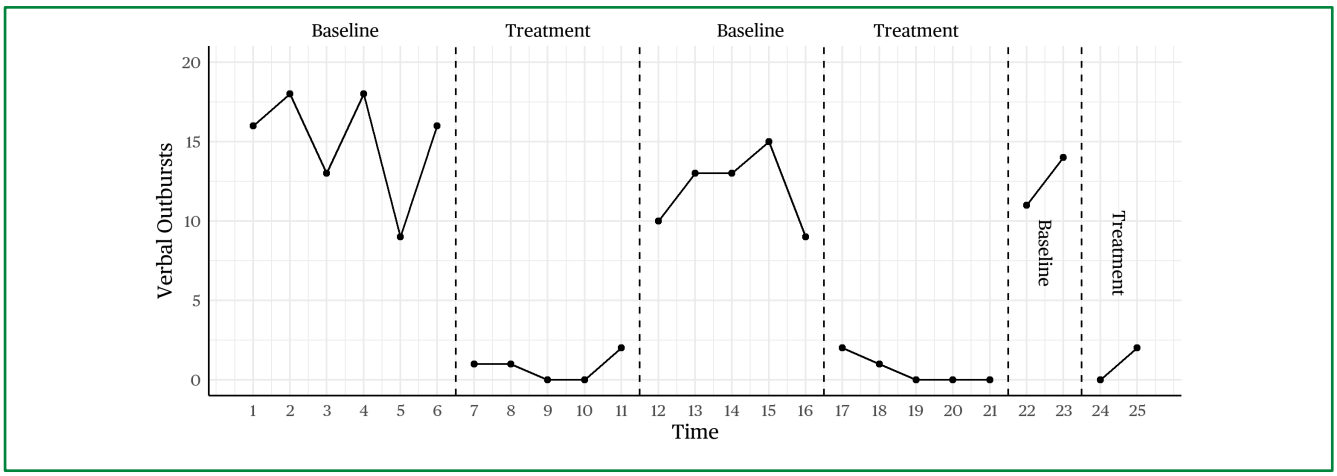


Figure 26 displays another example treatment reversal design with six phases. In contrast to the previous example, this finding would be rated *Does Not Meet WWC Standards.* Although it has four phases that meet the data point requirements for a rating of *Meets WWC Standards Without Reservations*, the two phases in the middle cannot be used as part of a phase transition. Nonsequential phases cannot serve as demonstrations of an intervention's effect. Therefore, only two potential demonstrations of the intervention effect have a sufficient number of data points: one between the first and second phases and one between the fifth and sixth phases.

Figure 26. *Treatment reversal design with extra phases that is rated* Does Not Meet WWC Standards

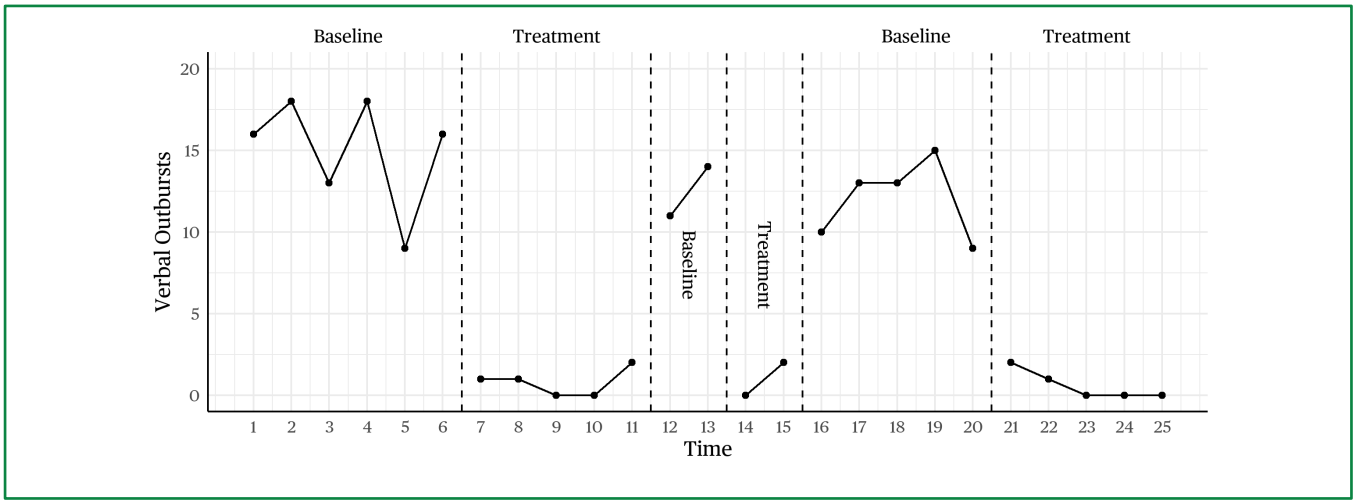


Figure 27 displays a multiple baseline design with four cases that all receive an intervention at staggered times. The first, second, and fourth case all have a sufficient number of data points to be rated *Meets WWC Standards With Reservations*. However, the third case dropped out after only two intervention data points were gathered. This finding would still be eligible to be rated *Meets WWC Standards With Reservations* because it contains enough information to demonstrate the intervention effect at three different points in time.

Combination designs. Findings from combination designs (such as a multiple baseline with embedded reversals) should receive the highest possible rating that any subset of their design is eligible for. Figure 28 displays a multiple baseline design with three cases. The second and third cases only contain the traditional baseline and treatment phase, but the first case also contains an embedded reversal/withdrawal design. The second pair of phases in the first case are brief. These two phases contain only three data points each and therefore are at best eligible to receive a research rating of *Meets WWC Standards With Reservations* under the reversal/withdrawal design requirements. However, given that the subset of initial baseline and treatment phases for all three cases, when considered as a multiple baseline design, would be eligible to receive a research rating of *Meets WWC Standards Without Reservations,* the finding from that combination design should be rated *Meets WWC Standards Without Reservations.*

Figure 27. *Multiple baseline design with four cases rated* Meets WWC Standards Without Reservations

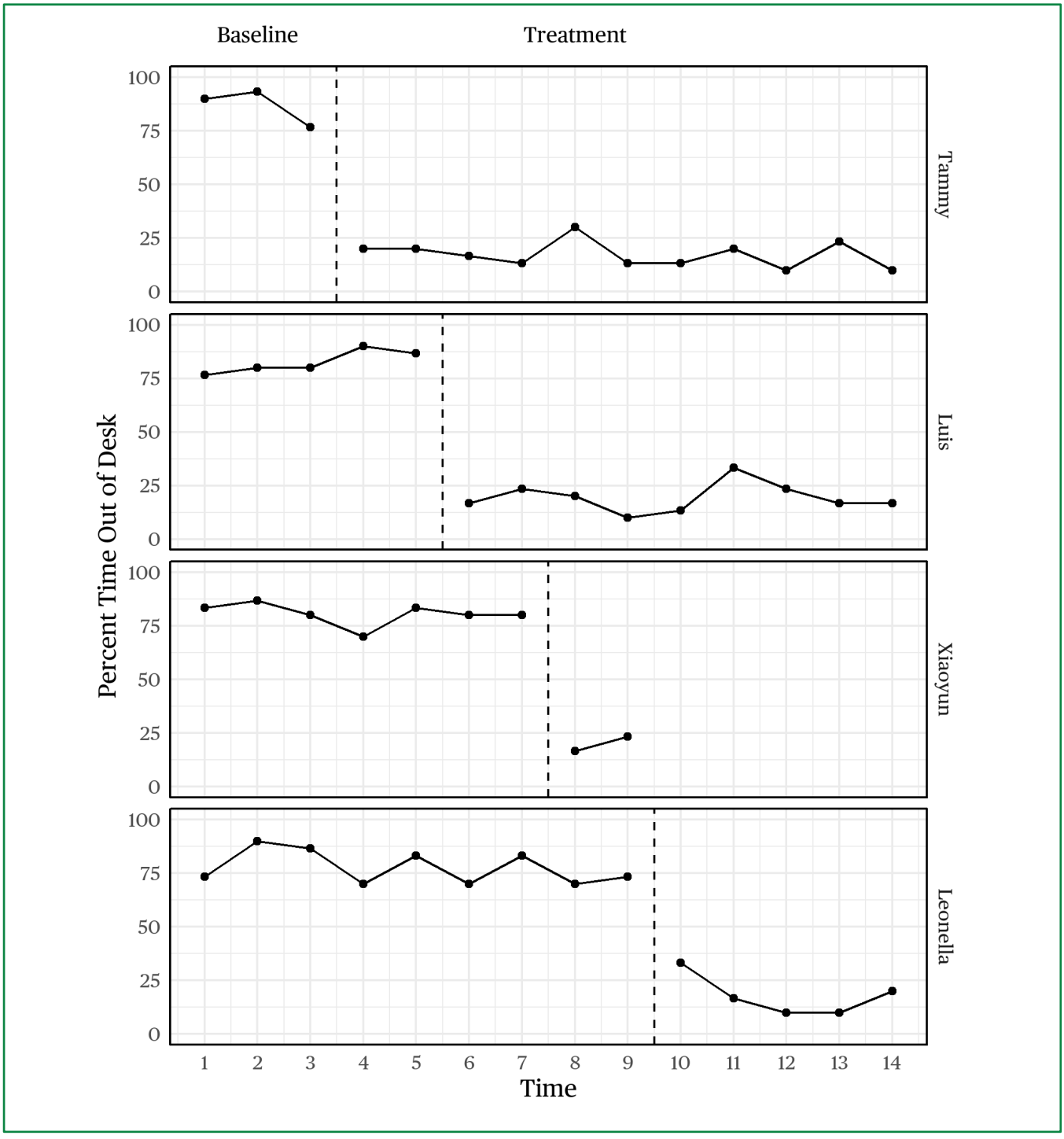
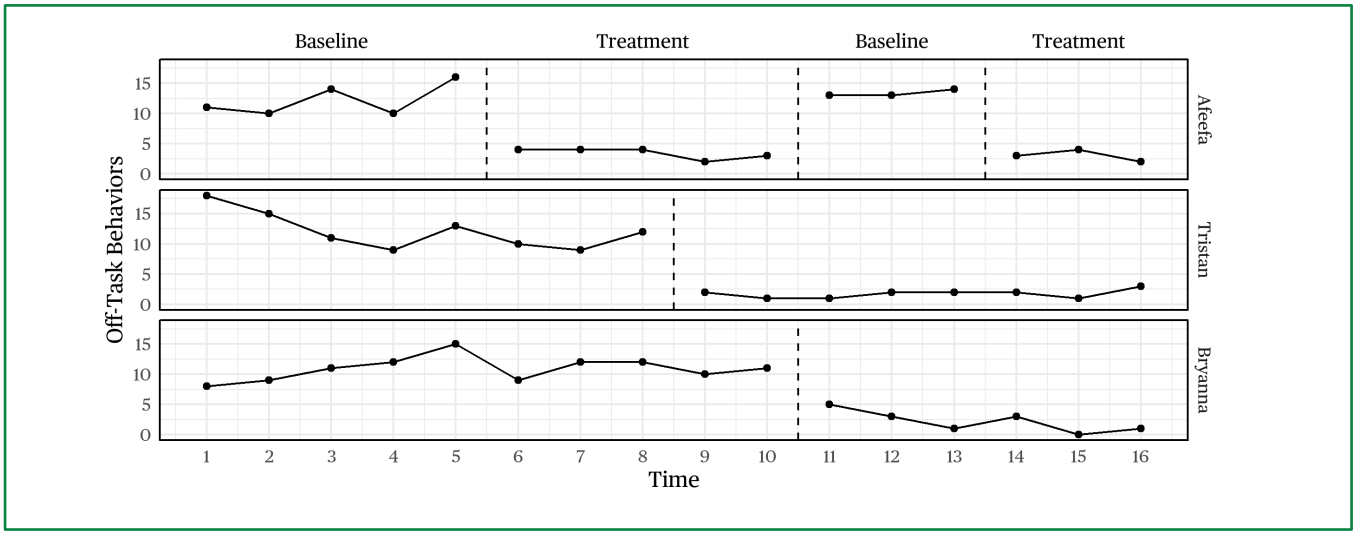


Figure 28. *Combination multiple baseline design with reversals that are rated* Meets WWC Standards Without Reservations



Extra conditions. Treatment reversal/withdrawal, multiple baseline designs, and alternating treatment designs may have more than one intervention condition. Unless otherwise specified by review team leadership, reviews should focus on contrasts comparing two conditions rather than reviewing three or more conditions at once.

CHAPTER VII. SYNTHESIS AND REPORTING OF RESULTS

The final phase in the What Works Clearinghouse (WWC) study review process is synthesis and reporting of results. The synthesis and reporting procedure described in this chapter applies to studies that use any research design eligible for WWC review. If differences exist, which is most often the case for single case design (SCD) studies, they are highlighted throughout the chapter. The WWC reports two sets of results: a *research rating* reflecting the quality of the research design for estimating the impact of the intervention and an *effectiveness rating* characterizing the evidence of the intervention's effects in a specific outcome domain. The two ratings are independent of one another. A study that receives a high research rating for research design and execution of design, such as *Meets WWC Standards Without Reservations*, may have a low rating for the effectiveness of the intervention, such as uncertain effects.

Synthesis and reporting of the study's results is based on the identification of eligible findings. The WWC defines a finding as the measured effect of the intervention relative to a specific comparison condition on an outcome for a sample at a certain point in time. The WWC determines the study's eligible findings and a domain to which they belong using the *Study Review Protocol*.

The use of supplemental findings in evidence synthesis products

Practice guides and intervention reports may include supplemental as well as main findings in their evidence syntheses and effectiveness ratings. A reviewer conducting an evidence synthesis should consult a topic area synthesis protocol for guidance.

Synthesis and reporting of the study's results is also based on the designation of findings as *main* or *supplemental* following the criteria below. While main findings must be reviewed as part of the study review, supplemental findings are only reviewed if specified in the *Study Review Protocol*, or if they are needed for purpose of the review (for example, a subgroup of participants, such as English language learners, is the focus of the review), or if these findings are needed to construct a main finding in an outcome domain.

Criteria for designating findings as main or supplemental

The WWC uses the criteria below for evaluating findings in all study reviews. Note that for synthesis products (for example, intervention reports, practice guides), the team conducting the synthesis may use supplemental findings as main findings. The determinations of main versus supplemental findings in the underlying individual study reviews will remain unchanged, however. For example, if a synthesis product is focused on a subgroup of students that would be considered a supplemental finding in individual study reviews, the team conducting the synthesis may choose to analyze subgroup findings as main for the purpose of the synthesis. Such decisions will be documented in a topic area synthesis protocol.[34]

Main findings are based on eligible study findings that have the following characteristics:

- They are measured for the full analytic sample from the study, or for subsamples that can be summed to equal the full analytic sample, if study authors do not report findings for the full analytic sample. In the latter

[34] The WWC does not apply the criteria for main and supplemental findings to SCDs. This is because in SCDs, the combination of outcomes and samples typically result in findings designated as main findings.

case, the subsample findings contribute to a main finding but are reported separately as supplemental findings. Review team leadership may allow findings for multiple student grade levels to count as multiple main findings without needing to first combine them as a single main finding.

- They are measured as a composite as opposed to subscales in the same domain.
- They rely on independent measures, if in an outcome domain specified by the *Study Review Protocol* as requiring independent measures.
- They are measured at the time period closest to the end of the intervention, unless they are in a domain for which the *Study Review Protocol* specifies that they are to be measured at a later time period.
- They are based on the more continuous version when both a continuous version and a dichotomized version of the same outcome measure are available. The *Study Review Protocol* may identify exceptions to this general rule.
- They have an effect size if an acceptable computational approach exists for the research design.

Supplemental findings include eligible study findings with the following characteristics:

- They include findings for eligible subgroups of study participants who do not represent the full analytic sample, as articulated in the *Study Review Protocol*. Findings for subgroups not identified in the *Study Review Protocol* may be reviewed as needed for the purpose of the review. For example, findings for some additional subgroups may be reviewed because they are cited as evidence for a grant competition.
- They include findings on subscale scores as opposed to composite scores when the composite was also administered to participants.
- They include findings relying on nonindependent measures when the *Study Review Protocol* excludes these measures from contributing to main findings in the corresponding outcome domain.
- They include findings at additional time periods not represented in the main findings, such as additional follow-up time periods or interim time periods.
- They are based on a dichotomized version when both a continuous version and a dichotomized version of the same outcome measure are available.
- They include findings for which effect sizes are not available, even though an acceptable computational approach is available for that research design.

When main findings are unavailable or unclear, review team leadership may use one of the approaches below.

- If the study has no main findings for the full analytic sample but supplemental findings meet WWC standards for nonoverlapping subsamples that sum to the full sample, such as grade or cohort, then whenever possible the WWC will pool these supplemental findings as the main finding, using the procedures described in appendix F in the technical appendices. The WWC also will report the corresponding subsample findings as supplemental findings.
- If there is not a single comparison group for the study but multiple possible comparison groups, then review team leadership will use the procedures described in appendix F in the technical appendices to create a pooled

comparison group. The WWC also will report the findings for each potential comparison group as supplemental findings.

- When study authors report a set of sensitivity analyses that focuses on the same or very similar samples but applies different analytic methods to obtain each finding, the WWC relies on the primary analysis identified by study authors in their reporting of findings. The WWC generally will *not* review the remaining sensitivity analyses, unless specifically needed for the purpose of the study review, but will note the existence of these additional analyses in the study review documentation.
- If study authors report findings from both intent-to-treat (ITT) and complier average causal effects (CACE) analyses, the WWC usually will rely on findings from the ITT analysis to identify main findings. The choice may be based on the type of research question that review team leadership judges is of greatest interest to decisionmakers. Appendix G in the technical appendices provides more detail.
- Findings with missing effect sizes cannot contribute to main findings if an effect size computational procedure exists for the research design. However, some research designs such as certain types of SCDs may not have an available WWC procedure for calculating the effect size (see appendix E in the technical appendices), even if the study provided its raw data. In those cases, the finding can still represent a main finding but cannot be used to assign an effectiveness rating to the intervention in the corresponding outcome domain.

If no main finding meets WWC standards in an outcome domain, the WWC may still report supplemental findings in that domain that meet WWC standards, either if required by the *Handbook* or for the purpose of the study review.

Determining the study's research rating based on the research ratings of findings

The WWC assigns one of three research ratings to each eligible main and supplemental finding: *Meets WWC Standards Without Reservations*, *Meets WWC Standards With Reservations*, or *Does Not Meet WWC Standards*. These ratings are based on how well the study met the design-specific criteria described in previous sections.

Findings from the same study can have different research ratings

For example, if one finding is for a low-attrition sample from an RCT, it will be eligible for the highest research rating of *Meets WWC Standards Without Reservations*. A different finding from the same study for a sample that experienced high attrition will be eligible for the research rating of *Meets WWC Standards With Reservations*.

After identifying the research ratings of each main and supplemental finding from the study, the WWC will assign a *study-level* research rating. As shown in table 22, if a study has at least one main finding that is rated *Meets WWC Standards Without Reservations*, the study-level research rating will be *Meets WWC Standards Without Reservations*. This rule applies even if the study has other findings that did not receive the highest research rating. If a study's highest rated main finding is rated *Meets WWC Standards With Reservations,* then the study's research rating will be *Meets WWC Standards With Reservations*. If a study only has supplemental findings that met WWC standards, the study-level research will be *Meets WWC Standards With Reservations*, even if its supplemental findings were rated higher. The lower research rating

reflects the WWC's concern about the study containing no main findings that met WWC standards. Finally, if no main or supplemental findings met WWC standards—that is, all reviewed findings received a research rating of *Does Not Meet WWC Standards*—then the study will be rated *Does Not Meet WWC Standards*.

Table 22. *Criteria for the study-level research rating based on research design and execution*

Study-level research rating	Criteria
Meets WWC Standards Without Reservations	At least one main finding is rated *Meets WWC Standards Without Reservations*.
Meets WWC Standards With Reservations	At least one main or supplemental finding is rated *Meets WWC Standards With Reservations*. OR At least one supplemental finding is rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*, but all main findings are rated *Does Not Meet WWC Standards* or the study does not have main findings.
Does Not Meet WWC Standards	All main and supplemental findings are rated *Does Not Meet WWC Standards*.

Determining an effectiveness rating based on the evidence of the intervention's effects

After the study is rated based on the rigor of design and execution of design, the WWC will rate the effectiveness of the intervention. The goal of this step in the WWC review is to signal whether the intervention did or did not affect a change in outcomes. The WWC only completes this step for studies rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*.

The WWC uses main findings to characterize the intervention's effectiveness by computing a domain-level composite of the main findings in each outcome domain. The WWC does not use supplemental findings to compute the domain-level composites for characterizing the intervention's effectiveness in individual studies. Recall that for synthesis products, the team leadership may use supplemental findings as main findings in the synthesis. For example, the effectiveness rating in a synthesis product may be based on specific subgroup findings or findings based on nonindependent measures, if specified in the topic area review protocol. The following sections describe how the WWC determines and reports evidence of the intervention's effectiveness. Because the WWC relies on effect sizes and their statistical significance to determine effectiveness ratings, a discussion about these measures is included at the conclusion of this section.

Effectiveness ratings in reviews of individual studies and intervention reports

The WWC uses the same approach to characterize the evidence of effectiveness from individual studies and intervention reports, which combine findings from all eligible studies of the intervention that meet WWC standards.

The WWC's effectiveness ratings include the tiers of evidence defined in the Elementary and Secondary Education Act as reauthorized by the Every Student Succeeds Act (ESSA) of 2015 and operationalized for discretionary grant programs in the Education Department General Administrative Regulations (34 CFR, Part 77). The WWC incorporates these evidence tier definitions into its effectiveness ratings to simplify the usability of ratings for education decisionmakers who often need to identify evidence that aligns with the U.S. Department of Education's definitions. Based on these criteria, the evidence from individual studies and intervention reports is

categorized as one of five effectiveness ratings: strong evidence (Tier 1), moderate evidence (Tier 2), promising evidence (Tier 3), uncertain effects, or negative effects. The characterization of intervention effectiveness depends on the sign of the average effect for each outcome domain composite and its statistical significance. For favorable impacts (Tiers 1–3), the effectiveness rating depends on the research ratings for the findings, whether the positive impact in a single study is overridden by a negative impact in the same domain, and whether the finding is based on a multisite sample—consisting of two or more states, districts, counties, cities, districts, schools, or campuses—and on a sample of at least 350 individuals.

The WWC characterizes evidence from individual studies and intervention reports by outcome domain. A WWC-constructed outcome domain composite may include a single main finding, in which case the WWC will consider that finding a domain-level finding. If an outcome domain includes several main findings, the WWC will use fixed-effects meta-analysis to combine every effect size into one meta-analytic average (Hedges & Vevea, 1998). For intervention reports, the WWC will create a composite for each outcome domain by calculating a meta-analytic average of every effect size from main study findings included in the report (and supplemental findings if allowed by a topic area synthesis protocol). How the WWC conducts a fixed-effects meta-analysis to combine effect sizes is described in detail in appendix F in the technical appendices. The WWC's criteria for characterizing findings from individual studies and intervention reports are summarized in table 23.

Table 23. *What Works Clearinghouse effectiveness ratings in individual studies and intervention reports by outcome domain*

Effectiveness rating and evidence tier	Criteria
Strong evidence 	• Summary: Positive effects, with no overriding negative effects, from well-designed, well-executed experimental research conducted in multiple sites and with a sufficiently large sample. • The fixed-effects meta-analysis of main study findings (or the single main finding) in the outcome domain is statistically significant and positive; AND • More than 50 percent of the meta-analytic weight of main study findings (or the single main finding) in the outcome domain is based on finding(s) that are rated *Meets WWC Standards Without Reservations*; AND • The analytic sample includes multiple sites (states, counties, cities, districts, schools, or campuses); AND • The analytic sample includes 350+ unique individuals; AND • If a study contributes to an intervention report, the intervention report did not find negative effects in the same outcome domain.

Continued on next page

Table 23. *What Works Clearinghouse effectiveness ratings in individual studies and intervention reports by outcome domain (continued)*

Effectiveness rating and evidence tier	Criteria
Moderate evidence 	• Summary: Positive effects, with no overriding negative effects, from well-designed and well-executed quasi-experimental research conducted in multiple sites and with a sufficiently large sample, OR, for intervention reports only, positive effects, with no overriding negative effects, from well-designed and well-executed experimental research conducted in multiple sites. • The fixed-effects meta-analysis of main study findings (or the single main finding) in the outcome domain is statistically significant and positive; AND • The analytic sample includes multiple sites (states, counties, cities, districts, schools, or campuses); AND • More than 50 percent of the meta-analytic weight of main study findings (or the single main finding) in the outcome domain is based on finding(s) that are rated *Meets WWC Standards With Reservations* and the analytic sample includes 350+ unique individuals; OR • **For intervention reports only**, more than 50 percent of the meta-analytic weight of main study findings (or the single main finding) in the outcome domain is based on finding(s) that are rated *Meets WWC Standards Without Reservations* and the analytic sample includes at least 20 unique individuals across multiple sites; AND • If a study contributes to an intervention report, the intervention report did not find negative effects in the same outcome domain.
Promising evidence 	• Summary: Positive effects, with no overriding negative effects, from well-designed and well-executed experimental or quasi-experimental research conducted in a single site or lacking a sufficiently large sample. • The fixed-effects meta-analysis of main study findings (or the single main finding) in the outcome domain is statistically significant and positive; AND • The analytic sample includes only one site (state, county, city, district, school, or campus) or is insufficiently large to meet Tier 1 or 2 requirements; AND • If a study contributes to an intervention report, the intervention report did not find negative effects in the same outcome domain.
Uncertain effects	• The fixed-effects meta-analysis of main study findings (or the single main finding) in the outcome domain is not statistically significant, or the statistical significance is unknown or cannot be calculated; OR • The study has no main finding but at least one supplemental finding meets WWC standards.
Negative effects	• The fixed-effects meta-analysis of main study findings (or single main finding) in the outcome domain is statistically significant and negative.

Note: The characterization of evidence is based on main findings only, with the exception under "uncertain effects" for supplemental findings that meet WWC standards or intervention reports utilizing supplemental findings.
Under *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0,* procedures, the WWC will allow Tier 2 rating for intervention reports based on samples of at least 20 individuals from studies that are rated *Meets WWC Standards Without Reservations*.

When an individual study or intervention report has only supplemental findings

As shown in table 23, a study or intervention report cannot be classified as Tier 1-3 based on supplemental findings alone, as these may not generalize to independent measures of effectiveness for the full analytic sample at the end of the intervention. In the cases where only supplemental findings meet WWC standards, the study or intervention will be classified as showing uncertain effects in the outcome domain.

When a study or intervention report has multiple outcome domains

A study or intervention report receives the highest rating of evidence among its outcome domains. For example, if the WWC determined that a study demonstrated strong evidence in one outcome domain and uncertain effects in another outcome domain, then the study or intervention report will be listed as showing strong evidence in at least one outcome domain.

Effectiveness ratings for practice guide recommendations

A practice guide is a publication based on research that presents recommendations to help educators address challenges in their classrooms and schools. In contrast with an intervention report, a practice guide focuses not on characterizing evidence for a single intervention but on identifying a set of intervention components that, when implemented appropriately, may improve student outcomes or other outcomes relevant for educators. Each guide includes practice recommendations based on a systematic review of studies by the WWC, on practitioner experience, and on the opinions of a panel of nationally recognized experts.

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

- The extent and quality of evidence meeting WWC standards,
- Effects on relevant outcomes,
- Relevance of the research to the scope of the practice guide,
- Whether the recommendations were directly tested in the research, and in some cases,
- Expert opinion informed by relevant research.

As in individual studies and intervention reports, the WWC rates evidence for each recommendation in practice guides by outcome domain. Favorable evidence is categorized as one of four effectiveness ratings aligned with ED evidence definitions (34 CFR, Part 77): strong evidence (Tier 1), moderate evidence (Tier 2), promising evidence (Tier 3), or demonstrates a rationale (Tier 4). Favorable evidence for each outcome domain and recommendation is determined by calculating a meta-analytic average of every effect size from main study findings included in the report, as described in detail in appendix F in the technical appendices. Table 24 outlines the criteria the WWC uses to determine the level of evidence supporting each practice guide recommendation.

Using effect size and statistical significance in effectiveness ratings

To determine an effectiveness rating based on the evidence of the intervention's effects, WWC relies on effect sizes and their statistical significance.

Effect size

The WWC develops effectiveness ratings by first determining the magnitude of an intervention effect or effect size. For randomized controlled trials (RCTs), quasi-experimental designs (QEDs), and regression discontinuity designs (RDDs), the WWC uses a standardized mean difference metric (Hedges' *g*) to compare and aggregate the magnitude of intervention effects within and across studies (effect sizes for SCD are described in Chapter VII, Design-comparable effect sizes for SCD studies). This effect size metric represents the mean difference between intervention and comparison observations on a standard deviation scale. Dividing by the standard deviation places study findings on a common scale across different outcome measures.

Table 24. *Effectiveness ratings for recommendations in practice guides*

Criteria	Strong evidence	Moderate evidence	Promising evidence	Demonstrates a rationale
				
Extent and quality of evidence	The relevant favorable main findings are rated *Meets WWC Standards Without Reservations* and include multiple sites (states, counties, cities, districts, schools, or campuses) and 350+ unique individuals.	The relevant favorable main findings meeting WWC standards include multiple sites (states, counties, cities, districts, schools, or campuses) and either 20 to 349 unique individuals in findings rated *Meets WWC Standards Without Reservations*, or 350+ unique individuals in findings that are rated *Meets WWC Standards With Reservations*.	The relevant favorable main findings meet WWC standards but do not include more than one site (state, county, city, district, school, or campus), or include an insufficient number of unique individuals to meet evidence tier definitions for strong or moderate evidence.	The research base does not include relevant main findings that meet WWC standards.

Continued on next page

Table 24. *Effectiveness ratings for recommendations in practice guides (continued)*

Criteria	Strong evidence	Moderate evidence	Promising evidence	Demonstrates a rationale
				
Effects on relevant outcomes	The fixed-effects meta-analytic synthesis of main findings shows positive effects for at least half of the relevant outcome domains and uncertain effects for any other outcome domains relevant to the recommendation.	The fixed-effects meta-analytic synthesis of main findings shows positive effects for at least half of the relevant outcome domains and uncertain effects for any other outcome domain relevant to the recommendation.	The fixed-effects meta-analytic synthesis of main findings shows positive effects for at least one relevant outcome domain and is not overridden by negative effects in at least as many outcome domains relevant to the recommendation.	Any fixed-effects meta-analytic synthesis of main findings does not show positive effects for any outcome domain relevant to the recommendation, or any positive effects are overridden by negative effects in at least as many outcome domains relevant to the recommendation.
Relevance to scope	The research covers the full range of populations and settings that are the focus of the recommendation.	The research overlaps with the populations and settings that are the focus of the recommendation.	The research overlaps with either the populations or settings that are the focus of the recommendation.	The recommendation reflects expert opinion based on defensible theory and/or reasonable extrapolations of research.
Direct versus indirect tests of the recommend-dation	The recommendation is a major component of the interventions evaluated in the research. The practice guide addresses all major components of the interventions.	The recommendation is a major component of the interventions evaluated in the research.	The recommendation is a component of the interventions evaluated in the research.	The recommendation reflects expert opinion based on defensible theory and/or reasonable extrapolations of research.

Note: A recommendation must satisfy all applicable criteria in the same column for the WWC to characterize the practice as supported by an evidence base at that level. If only one study provides a relevant finding, then the study's domain-level finding replaces the meta-analytic synthesis of findings from multiple studies. "Positive effect" refers to an average effect that is statistically significant and favorable for the corresponding outcome domain.

Importance of effect sizes. Effect sizes play a role in all aspects of the WWC's reporting of effectiveness ratings. If an effect size cannot be computed due to missing information, the WWC classifies the finding a *supplemental* rather than a *main* finding because the finding cannot be used to characterize the intervention's effectiveness. For instance, a study could report an unstandardized mean difference but not the standard deviations needed for computing the effect size. Designating findings with missing effect sizes as supplemental limits their contribution to the WWC's evidence base. The following sections describe this point in more detail (including some uncommon exceptions to the general rule).

Author Queries

When the study does not contain important information required to determine the study's eligibility for a review or its research rating, including information needed to compute the effect size, the WWC may send an author query to request missing information. The WWC will *not* ask authors to conduct new analyses. The WWC's procedure for sending an author query it described in appendix B in the technical appendices.

WWC reviewers should therefore be vigilant in ensuring they have extracted sufficient information from studies to enable computing effect sizes. Appendix E in the technical appendices details the strategies and information that the WWC uses to compute effect sizes. This appendix aims to account for variation in how study authors may report their results. Means and standard deviations may be unavailable in some cases, but studies may report other information such as *t* or *F* statistics that can be converted to effect sizes.

Guidelines for computing effect sizes. WWC reviewers will enter statistical information into an Online Study Review Guide that will apply the statistical formulas described in appendix E in the technical appendices. A reviewer should follow these four principles when entering information and choosing a computation method:

1. **Prefer covariate-adjusted over unadjusted mean differences:** The WWC often requires that study authors adjust for baseline differences, such as for QEDs or high-attrition RCTs. Effect sizes based on unadjusted means would be unacceptable in such cases. The WWC prefers covariate adjustment for other studies, such as RCTs with a low risk of bias due to compositional change where adjustments are not required, given the increased precision of estimating intervention effects. Whenever possible, a reviewer should therefore use covariate-adjusted mean differences to compute effect sizes. These adjusted mean differences can come from various types of models such as multiple regression, analysis of covariance, or analyses of gain scores.
2. **Prefer unadjusted, individual-level standard deviations of the outcome:** For consistency, the WWC computes effect sizes using the unadjusted individual-level standard deviations of the outcome, reported separately for the intervention and comparison groups. The WWC uses this information to compute the pooled within-group standard deviation used to standardize the mean difference. Appendix E describes approaches to convert other information to this target type of standard deviation, including extensions for the total standard deviation, analyses of *z* scores, standardized regression coefficients, group mean standard errors, cluster-level standard deviations, gain score standard deviations, covariate-adjusted standard deviations, and baseline standard deviations. In some cases, these alternative conversions require additional information (such as the baseline-outcome correlation when converting gain score standard deviations to

unadjusted standard deviations). The end goal for these conversions is the same: unadjusted, individual-level outcome standard deviations.

3. **The lack of reported information may constrain the computation choice:** Practical limitations in author reporting will often guide which approach a reviewer should use to compute effect sizes. If the reported information allows for multiple computation options, review team leadership has the discretion to choose the option that best aligns with the two principles noted earlier (for instance, prefer using the unadjusted standard deviations, rather than computing them indirectly through other conversions). An author query is needed if an effect size calculation procedure exists for the research design, but the study article did not report sufficient information to extract an effect size.

   Example. Consider an individual-level assignment study that used multiple regression to adjust for baseline differences, reporting the (a) unstandardized regression coefficient for the impact estimate, (b) coefficient *t* statistic, and (c) multiple correlation R^2 between the covariates and outcome. This information is insufficient to calculate the effect size using a standard formula (mean difference divided by standard deviation) because the study did not report the needed standard deviations. However, a reviewer could use the covariate-adjusted *t* statistic and R^2 value to compute the effect size using a formula described more fully in appendix E (equation E.12) in the technical appendices. This approach is allowable, although the WWC generally favors a more standard formula (equation E.1) if the study also reported the unadjusted standard deviations for the intervention and comparison groups. An author query would be needed if the study only reported the regression coefficient and its test statistic (by themselves, those two statistics are insufficient to calculate effect sizes for covariate-adjusted analyses).

4. **Prefer covariate-adjusted percentages for dichotomous outcomes:** Computing effect sizes for dichotomous outcomes requires special formulas that do not apply to continuous outcomes. Nevertheless, one principle applies to both types of outcomes: prefer covariate-adjusted means over unadjusted means. For dichotomous outcomes, this principle means using covariate-adjusted percentages (or proportions) from a logistic, probit, or linear probability regression model, as described more fully in appendix E in the technical appendices. The WWC will allow, but does not prefer, computing effect sizes based on logistic regression coefficients. This approach requires that the study also reports the standard error for the logistic regression coefficient. Unlike continuous outcomes, standard deviations are not used to compute effect sizes for dichotomous outcomes.

5. **For RDD studies**, the predicted means or probabilities for continuous and dichotomous outcomes must be calculated using the same statistical model that is used to estimate the impact on the outcome at the cutoff.

Design-comparable effect sizes for SCD studies. For SCD studies, the WWC will estimate a design-comparable effect size, if feasible. The design-comparable effect size is comparable to a Hedges' *g* for group designs and can be synthesized together with effect sizes from group designs.

The design-comparable effect size only can be estimated from treatment reversal/withdrawal designs and multiple baseline/multiple probe designs with three or more individuals. Other SCDs currently do not have an available procedure for computing the design-comparable effect size. However, the WWC will store the information from reviews of these studies, including tabular data on study findings, for possible use by the

research community in conducting independent visual analyses and development of new methods of design-comparable effect size estimation. If the WWC identifies additional methods of design-comparable effect size estimation appropriate for use by the WWC, the WWC will document these methods in a supplement to the *Procedures and Standards Handbook* or in a new version of the *Handbook*. Each topic area synthesis protocol will note which version of the *Handbook* and any supplement will govern the meta-analytic synthesis of findings from systematic reviews in that topic area.

Due to the model-dependent nature of the design-comparable effect size, the WWC has produced special guidance for the estimation of effect sizes from SCDs. When estimating effect sizes from multiple SCD studies for the purpose of synthesis, review teams should wait until all findings eligible for synthesis have been identified before estimating an effect size. Review teams should collaborate with a visual analysis expert to identify possible common time trends within a domain that may need to be accounted for in the design-comparable effect size model.

Occasionally, SCDs will contain designs from two different units of analysis. This can represent a challenge for synthesis in the case of treatment reversal/withdrawal designs, where individual cases are rated separately but three or more cases can be combined to estimate an effect size. For instance, a design might include six treatment reversal/withdrawal designs, three with student-level outcomes and three with classroom-level outcomes. Effect sizes should not mix units of analysis, so individual level-data would first be synthesized into effect sizes separately from classroom-level data. Then, if the effect sizes were in the same domain, the classroom-level effect sizes could be rescaled to be comparable to the individual-level effect sizes using formula E.27 in appendix E, and then aggregated with the individual-level data to estimate a domain-level effect size. Additionally, appendix E provides information on how the WWC approaches analyses of cluster-level data.

Another issue specific to treatment reversal/withdrawal designs is that estimating the design-comparable effect size will sometimes require combining data from designs with different ratings. For instance, a study may include two treatment reversal/withdrawal designs that are rated *Meets WWC Standards Without Reservations*, and one treatment reversal/withdrawal design that is rated *Meets WWC Standards With Reservations*. So long as the outcomes and interventions are the same, the WWC will combine these designs together to yield one effect size. The finding will receive the highest rating of any design used to estimate the effect size.

In a similar fashion, effect sizes should use as much data in a design as possible. For instance, imagine a multiple baseline design with three cases that contains an embedded reversal design in the first tier. When reviewing the multiple baseline design component alone, the multiple baseline design receives a rating of *Meets WWC Standards Without Reservations*. In contrast, the treatment reversal/withdrawal design receives a rating of *Meets WWC Standards With Reservations*, and the single treatment reversal/withdrawal cannot be used to estimate an effect size. When possible, the WWC prefers to include the additional information from the embedded treatment reversal design in the effect size for the multiple baseline design. As already discussed, this finding would receive a rating of *Meets WWC Standards Without Reservations*.

However, designs that include the modeling framework the WWC presently uses cannot accommodate within- or across-phase trends for treatment reversal/withdrawal designs. To estimate an effect size using all the data from a multiple baseline design with embedded reversals, the review teams should have identified that the change-in-

levels modeling specification is most appropriate for the relevant domain. If the review team, in collaboration with a visual analysis expert, has identified models with trends as most appropriate for findings in the relevant domain, an effect size should be estimated using as much of the data as the model can accommodate. In the example provided, this would be only the data in the multiple baseline subset, excluding the data from the treatment reversal/withdrawal subset of the overall design.

Appendix E contains more information about effect size estimation in SCDs.

Statistical significance

To adequately assess the effects of an intervention, it is important to know the statistical significance of the estimates of the effects in addition to the effect size.

Guidelines for computing statistical significance. The WWC computes the statistical significance for a finding by first computing the standard error for the effect size. This standard error represents the uncertainty in the effect size estimate, with larger values corresponding to less precise estimates. Statistically significant findings are those with large effect sizes relative to the standard error. Appendix E provides further detail on computing the standard error and statistical significance.

When to use study-reported or WWC-calculated values for effect size and statistical significance

Study authors may report their own effect size estimate and determination of statistical significance. The WWC generally prefers *WWC-computed effect sizes* and *effect sizes that have been adjusted for baseline differences* even if this adjustment is not required. The WWC generally prefers *study-reported statistical significance* unless the study authors failed to adjust for baseline differences or clustering in the data when required. For some studies, the WWC also may need to combine multiple findings into a composite finding at the outcome domain level and determine the effect size and statistical significance of the WWC-constructed composite.

The WWC generally prefers WWC-calculated effect sizes (relative to study-reported effect sizes) because their computation can be verified, ensuring comparability across studies. Nevertheless, a reviewer may select to use the study-reported effect size in two common cases:

1. The WWC cannot compute an effect size based on the reported information; or
2. The WWC can compute an effect size based on unadjusted means, but the study-reported effect size adjusts for baseline differences.

The common Cohen's *d* effect size metric does not apply a small-sample size adjustment, although a post-hoc adjustment can convert it to a Hedges' *g* metric (see equation E.3 in appendix E in the technical appendices). A reviewer should flag whether study authors had applied a small-sample size adjustment to the effect size estimate, enabling the Online Study Review Guide to apply post-hoc adjustments if needed.

While the WWC may use effect sizes computed using a single-group standard deviation, if the pooled standard deviation is unavailable, reviewers should not use the study-reported effect size in the following cases:

- The study-reported value was based on unadjusted means, but the study requires adjustment for baseline differences.
- The study-reported value did not use unadjusted, individual-level standard deviations of the outcome to standardize the mean differences.
    - For instance, study-reported effect sizes based on gain score standard deviations are not acceptable due to the lack of comparability with effect sizes based on the unadjusted posttest standard deviations.
- The study authors report insufficient information on the computation approach for the review team to be confident in the alignment with the WWC's effect size procedures.

The WWC generally accepts the study-reported *p* values and statistical significance for study findings. However, a reviewer should favor the WWC computed statistical significance in four common cases:

1. Study authors did not include statistical significance estimates.
2. Study authors' calculations have a known problem such as not applying a required adjustment for baseline differences.
3. Study authors did not account for clustering in an individual-level analysis for a cluster-level assignment study (see appendix E for more detail).
4. Study authors reported statistical significance based on unadjusted analyses, but analyses required adjustment for baseline differences.

If multiple main findings meet WWC standards in the same outcome domain, the WWC also will pool these findings to determine the effect size and statistical significance at the domain level, using the procedures described in appendix F. The composite, as opposed to main findings, will then be used to determine the study's effectiveness rating in the outcome domain.

TECHNICAL APPENDICES

APPENDIX A. PRINCIPLES FOR PRIORITIZING AND SEARCHING FOR STUDIES TO REVIEW

The What Works Clearinghouse (WWC) reviews existing, publicly available research in education to inform federal, state, and local decisionmakers. Given the vast research literature on education, the WWC must prioritize topics for systematic reviews as well as reviews of individual studies. This appendix describes the processes that govern the identification and prioritization of eligible research for WWC review.

Limiting reviews to eligible and accessible research

To be reviewed by the WWC, studies need to be eligible under the current version of the *Study Review Protocol* and publicly available. The WWC considers recent studies more relevant to decisionmakers than older studies and therefore, studies over 20 years old are ineligible for WWC review.

When selecting publicly available manuscripts for review, the WWC favors studies with a final study report or peer-reviewed manuscript in ERIC. ERIC is the Institute of Education Sciences' (IES') searchable, online bibliographic and full-text database of education research for educators, researchers, and the general public. If studies are nominated for WWC review through the Help Desk, the WWC will first encourage study authors to submit their study through ERIC's online submission system for prioritization for review, if the study is not published in a routinely indexed source.

Prioritizing topics for practice guides

The WWC conducts systematic reviews of studies in select topic areas for the purpose of communicating research evidence to practitioners through the publication of practice guides and intervention reports. IES has authorized the WWC to conduct systematic reviews in topic areas including: English language and literacy; science, technology, engineering, and mathematics (STEM); social, emotional, and behavioral interventions; and postsecondary education. The WWC identifies topics for future practice guides based on several considerations:

- The topic is aligned with a federal program's authorizing legislation or program design emphasizing the use of evidence-based practices; and
- The topic is one identified as a priority in education in federal legislation or by U.S. Department of Education leadership or by educators or school administrators; and
- The topic is one on which the WWC has not released a practice guide in the past 5 years; and
- There is evidence that at least two distinct interventions in the topic area as supported by strong, moderate, or promising evidence from multiple WWC individual study reviews, two or more WWC intervention reports, or two or more non-WWC systematic reviews of evidence that include high-quality research.

After IES selects a topic for a practice guide, the WWC will conduct a broad, systematic review of relevant evidence as specified in the corresponding topic area review protocol. The review will include the coding of intervention components for studies that meet WWC standards for possible meta-analytic synthesis to inform the expert panel's development of recommendations for the practice guide. Based on the systematic review of

relevant evidence, the WWC may identify a topic for an additional practice guide, as well as specific interventions for intervention reports, and will prioritize the preparation of intervention reports as described below.

Prioritizing interventions for intervention reports

The WWC conducts systematic reviews in select topic areas for the purpose of preparing intervention reports that summarize the research evidence on specific programs, policies, products, or practices. The WWC favors the preparation of intervention reports on interventions:

- That are replicable interventions of broad interest or in wide use (or both) according to topic area content experts or surveys of educators; and
- That are without an existing WWC intervention report released in the past 5 years; and
- For which the available research includes two or more studies that Meet WWC Standards With or Without Reservations, where at least one study is not included in a WWC intervention report; and
- For which the findings meeting WWC standards are based on either a multisite sample including at least 350 individuals across studies, or single-case design findings including at least 20 cases across studies and for which design-comparable effect-size estimation is feasible.

Each topic area review protocol may specify additional details of WWC procedures for prioritizing interventions for intervention reports. For each intervention to be included in such a report, the WWC will conduct a comprehensive literature search for all eligible studies on that intervention.

Prioritizing individual studies for review

The WWC uses the *Study Review Protocol* to review publicly available studies for a variety of reasons, giving priority to studies that:

- Have been identified in connection with a systematic review of evidence for the development of a WWC practice guide or intervention report; or
- Need to be assessed as strong evidence (Tier 1) or moderate evidence (Tier 2) as specified in the corresponding Notice Inviting Applications for a U.S. Department of Education grant competition; or
- Have been identified for review by U.S. Department of Education program office leadership or were funded by the Institute of Education Sciences; or
- Have not been previously reviewed by the WWC but have the potential to provide strong (Tier 1), moderate (Tier 2), or promising (Tier 3) evidence as defined in table 21 of this *Handbook*; or
- Have not been previously reviewed by the WWC but focus on replicable interventions of broad interest or in wide use (or both).

Literature search procedures

Systematic literature searches are a critical component for WWC intervention reports and practice guides. Review teams should work closely with their institutions' research librarians to design a comprehensive search strategy using best practices (see the Campbell methods guide on searching for studies [Kugley et al., 2017] and the WWC webinar on systematic literature searches using ERIC). The following section provides recommended practices for designing a search strategy, identifying databases to search, crafting search strings, conducting supplementary search methods, and reporting of the search strategy when designing and implementing a literature search for WWC products.

Designing a search strategy

Designing a literature search strategy for a systematic review requires balancing two competing considerations: comprehensiveness and efficiency. A truly comprehensive search would identify all relevant research for a review. A search that is not comprehensive fails to identify documents that are relevant to the review. An efficient search minimizes the number of irrelevant documents identified by the search. An inefficient search identifies documents that are not relevant to the review. A literature search for a systematic review seeks to optimize both comprehensiveness and efficiency, however this is often a tradeoff as increasing one can decrease the other. Following best practices (for example, see Kugley et al., 2017), review teams and librarians developing literature search strategies for the WWC should adopt the goal of identifying all eligible research. Therefore, the literature search strategy should prioritize comprehensiveness while simultaneously considering opportunities to maximize efficiency.

The WWC has several recommendations when a literature search is inefficient, that is, it identifies too many irrelevant studies, or the set of identified studies is unfeasibly large for the resources allocated to the review. Review teams should first investigate whether it makes sense to narrow the parameters of the research question, for example, by focusing on narrower aspects of the populations or interventions of interest. Search designers should consider narrowing the research question by focusing on more recent studies or studies with more rigorous research designs. Search designers should not, however, narrow the research question by focusing on published studies only, due to the likelihood of inducing publication bias (Polanin et al., 2016).

Identifying databases to search

The WWC has several recommendations regarding which online databases to search. Foremost, review teams should use ERIC as the primary database source when searching for studies. The public version of ERIC is an up-to-date index of education research and much of the gray literature. Review teams may choose to access ERIC through a search gateway that their institution subscribes to, such as EBSCOhost or ProQuest. These gateways can support more sophisticated search capabilities than the public version of ERIC. However, review teams should know that there is a lag in the indexing time between the public version of ERIC and the version indexed by these platforms and relying exclusively on commercial platforms may result in recent research being excluded from the search. In addition to searching ERIC, review teams should consider searching multidisciplinary databases that contain sources that are not comprehensively indexed in ERIC. For example, Academic Search Ultimate, ProQuest Dissertations and Theses, APA PsycInfo, Education Source, Education Research Complete, and EconLit may complement ERIC. When searching for dissertations and theses, it is important to utilize the ProQuest Dissertations

& Theses Global Database, as it is the most up-to-date and comprehensive index of these publication types. IES contractors should limit searching for dissertations and theses to ProQuest Dissertations & Theses Global, EBSCO Open Dissertations, ERIC, or another database specified in the topic area review protocol. IES contractors can utilize the National Library of Education's EBSCO Databases including Academic Search Ultimate, EconLit, Education Source, and ERIC, as well as ProQuest's Dissertations & Theses Global Database. For access to these databases, please send an email to AskALibrarian@ed.gov.

Review teams may consider searching multiple databases simultaneously but should recognize that there may be trade-offs. Gateways such as EBSCOhost and ProQuest allow review teams to create one search string that can be used simultaneously across multiple databases. Additionally, simultaneous searching enables automatic deduplication of citations across databases. However, simultaneously searching multiple databases restricts the fields available for searching to only those present in all of the databases included in the search. If, for example, one of the databases includes a field for study population that is not included in all of the other databases searched, the search will not be able to utilize this field.

Crafting the search string

Review teams should carefully consider the search terms and the search string created for the search. Review team leadership should begin by consulting the ERIC Thesaurus or the controlled vocabulary from complementary databases to identify relevant terms related to the intervention, population, and study designs of interest (see table A.1 for an example). To expand on this initial set of terms, consider adding synonyms (and applicable antonyms), related terms, natural language analogs, spelling variations (for example, those found in United States English versus United Kingdom English), and truncations. Review teams should be purposeful in selecting which fields to search (for example, title, abstract, subject heading) and whether the full text of the article will be searched, recognizing that this may increase the number of irrelevant articles identified by the search. Review teams and librarians developing a literature search strategy for the WWC should be careful about using filters and fields that may not be consistently populated or accurate in all databases, such as publication type, location, or methodology.

Table A.1. *Search term examples from the Adolescent Literacy Review protocol*

Category	Example search term
Intervention	Approach, curricula*, educational therapy, homework, improvement, instruct*, practice, program, remedial, school*, strategy, success*, teach*, treatment
Population	Adolescent*, eighth grade, elementary school, eleventh grade, fifth grade, fourth grade, grade 4, grade 5, grade 6, grade 7, grade 8, grade 9, grade 10, grade 11, grade 12, high school, junior high, K-12, middle grades, middle school, ninth grade, seventh grade, sixth grade, student*, summer school, tenth grade, twelfth grade
Study design	ABAB design, affect*, assignment, causal, comparison group, control*, counterfactual, effect*, efficacy, evaluation*, experiment*, impact*, matched group, meta analysis, meta-analysis, posttest, posttest, pretest, pre-test, QED, QES, quasi-experimental, quasiexperimental, random*, RCT, RDD, regression discontinuity, simultaneous treatment, SCD, single case, single subject, treatment, reversal design, withdrawal design

Note: This illustrative table is drawn from the Adolescent Literacy Review Protocol, Version 3.0, found at https://ies.ed.gov/ncee/wwc/Document/29. The asterisk (*) is a special character that allows the truncation of terms so that the search returns any word that begins with the specified letters. This feature varies across online databases and is not available in the public version of ERIC. Review teams should consult the specified online database to ensure accurate usage.

Developing a search string is typically an iterative process of identifying terms, conducting searches, evaluating results, and revising to improve accuracy and efficiency. Search strings can be improved using multiple methods. First, work with an institutional librarian to revise the search string to optimize comprehensiveness and efficiency. Second, have a colleague conduct a review of the search strategy using a check list such as the PRESS Evidence-Based Checklist (McGowan et al., 2016). Third, calibrate search strings to ensure they identify studies known to be relevant to the review. If studies that are known to be relevant are not identified with the search string, examine the searched fields of the relevant studies to identify additional terms to add to the string. To identify additional search terms that can be used for identifying relevant studies in an automated manner, review teams can use the R package *litsearchr* (Grames et al., 2019). This package uses text-mining to read a preliminary list of study abstracts and identify common words not included in the original search string.

Supplemental search strategies

In addition to searching ERIC and other databases, the WWC recommends several supplemental search strategies to ensure a highly sensitive literature search.

- **Gray literature.** Review teams should search specific websites or sources of gray literature that are not indexed in ERIC (the sources of grey literature currently indexed by ERIC can be found here). Websites of research firms, government agencies, nonprofit organizations, and other funders all may include eligible research.
- **Research registries**. Review teams should conduct searches of research registries like the Registry of Efficacy and Effectiveness Studies and Open Science Framework Registries.
- **Hand search.** Review teams can hand search specific journals that are particularly relevant to the topic.
- **Reference harvesting.** The literature search strategy should include forward and backward reference harvesting of relevant studies. Forward reference harvesting is when a member of the review team scans the titles of articles that cite a relevant study using Web of Science or Google Scholar. Backward reference harvesting is when a member of the review team scans the titles of articles included in the reference list of relevant studies.

Documenting the search strategy

Finally, review teams should document the implemented search strategy in topic area synthesis protocols with enough detail to support replication of the literature search. This documentation should include the databases searched; whether the databases were searched individually or simultaneously; the exact search string used for each database, containing Boolean operators and any special characters; and the supplemental search strategies that were implemented, including the specific sources of gray literature that were searched. Rethlefsen et al. (2021) provided a 16-item checklist for reporting literature searches.

APPENDIX B. PROCEDURES FOR SENDING AUTHOR QUERIES

The What Works Clearinghouse (WWC) reviews studies using information from the primary, publicly available study manuscript and related publicly available documents, but this information is sometimes insufficient to complete the review. In these cases, the WWC may send an **author query** to request information from the researchers who conducted the study. The WWC will not review findings sent through responses to author queries that are not publicly available and will encourage authors to make any unpublished findings publicly available through ERIC.

Author queries attempt to gather missing information needed to determine the study's eligibility, the WWC research rating for each finding in a study eligible for WWC review, an acceptable effect size estimate, or contextual information such as sample demographics. For instance, review teams may query study authors to determine whether the study sample demographics match eligibility criteria for the review protocol. Review teams also may ask study authors to provide information necessary to estimate effect sizes (such as unadjusted outcome standard deviations and covariate-adjusted means). However, the WWC does not ask study authors to conduct new analyses. The WWC summarizes and archives study authors' responses to WWC queries in the study review notes.

A typical author query process includes the following steps, though review team leadership has the discretion to adapt them as needed to support efficiency:

- The first WWC-certified reviewer performs the initial review using information available in the study. If key information is missing, the reviewer will notify review team leadership that the study's eligibility for review, its research rating, or the estimated effect size may differ if the study authors provide further information.
    - A senior member of the review team—such as a study reconciler—will either confirm that information from the study authors may change the study's eligibility, its rating, or calculation of effect size or will help the initial reviewer locate the necessary information in the study or related reports.
        - If the initial reviewer and senior member of the review team agree that an author query is needed, then they will draft the specific questions for the author(s) and may create a table in which the author(s) can fill in the missing information. If information regarding study context is not provided in the publicly available documents (such as demographic or geographic information about the study sample or the type of intervention), then the author query can request that information as well.
        - The WWC review team leadership locates contact information for the author(s) and sends the query via email. WWC review teams typically provide study authors two weeks to respond to queries. Review team leadership may give study authors additional time to answer the query if study authors request it.
        - On occasion, the reviewer and a senior member of the review team may determine that additional author queries are needed.
        - If study authors do not respond to a query or do not have the necessary information, then the initial reviewer will complete the review using just the information provided in the study and related reports. However, if the study authors' responses provide additional information, then the initial reviewer will incorporate that additional information into the review.

- If the initial reviewer concludes that the study should be rated ineligible or *Does Not Meet WWC Standards* based on information in the study, related manuscripts, and responses to author queries, then the reconciler will confirm the rating and finalize the review.
- If the initial reviewer concludes that the study should be rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations* based on information in the study, related manuscripts, and responses to author queries, then a second WWC-certified reviewer will make an independent determination of the study's eligibility, rating, and estimation of effect based on the same information used by the first reviewer.
    - After the second review, a reconciler identifies discrepancies in judgments between the first and second WWC-certified reviewers. The reconciler then consults with reviewers to determine the most appropriate judgments and finalizes the WWC review.

APPENDIX C. BOUNDARIES FOR DEFINING HIGH VERSUS LOW ATTRITION

The What Works Clearinghouse (WWC) must determine whether attrition is low or high for randomized controlled trials and regression discontinuity designs, as well as determine high or low representativeness for cluster-level assignment studies. The WWC has measured the levels of expected bias associated with different combinations of overall and differential attrition rates, under both optimistic and cautious assumptions. Review teams must choose between the cautious and optimistic attrition boundaries and provide a justification for their choice in the study review guide notes.

When the combination of overall and differential rates of attrition results in unacceptable levels of potential bias, the WWC labels the combination as high attrition. When the combination of overall and differential rates of attrition result in tolerable levels of potential bias, the WWC labels the combination as low attrition. For each overall attrition rate, table C.1 shows the highest differential attrition rate allowable to be considered low attrition. Note that the WWC also uses these attrition boundaries to assess whether the analytic sample of individuals from nonattriting clusters is representatives of those clusters.

Table C.1. *Highest differential attrition rate for a sample to maintain low attrition, by overall attrition rate, under cautious and optimistic assumptions*

	Differential attrition			Differential attrition			Differential attrition	
Overall attrition	Cautious boundary	Optimistic boundary	Overall attrition	Cautious boundary	Optimistic boundary	Overall attrition	Cautious boundary	Optimistic boundary
0	5.7	10.0	12	6.2	10.9	24	4.9	9.4
1	5.8	10.1	13	6.1	10.8	25	4.8	9.2
2	5.9	10.2	14	6.0	10.8	26	4.7	9.0
3	5.9	10.3	15	5.9	10.7	27	4.5	8.8
4	6.0	10.4	16	5.9	10.6	28	4.4	8.6
5	6.1	10.5	17	5.8	10.5	29	4.3	8.4
6	6.2	10.7	18	5.7	10.3	30	4.1	8.2
7	6.3	10.8	19	5.5	10.2	31	4.0	8.0
8	6.3	10.9	20	5.4	10.0	32	3.8	7.8
9	6.3	10.9	21	5.3	9.9	33	3.6	7.6
10	6.3	10.9	22	5.2	10.9	34	3.5	7.4
11	6.2	10.9	23	5.1	10.8	35	3.3	7.2

Continued on next page

Table C.1. *Highest differential attrition rate for a sample to maintain low attrition, by overall attrition rate, under cautious and optimistic assumptions (continued)*

	Differential attrition			Differential attrition			Differential attrition	
Overall attrition	Cautious boundary	Optimistic boundary	Overall attrition	Cautious boundary	Optimistic boundary	Overall attrition	Cautious boundary	Optimistic boundary
36	3.2	7.0	46	1.6	4.6	56	0.2	2.3
37	3.1	6.7	47	1.5	4.4	57	0.0	2.1
38	2.9	6.5	48	1.3	4.2	58	N/A	1.9
39	2.8	6.3	49	1.2	3.9	59	N/A	1.6
40	2.6	6.0	50	1.0	3.7	60	N/A	1.4
41	2.5	5.8	51	0.9	3.5	61	N/A	1.1
42	2.3	5.6	52	0.7	3.2	62	N/A	0.9
43	2.1	5.3	53	0.6	3.0	63	N/A	0.7
44	2.0	7.0	54	0.4	2.8	64	N/A	0.5
45	1.8	6.7	55	0.3	2.6	65	N/A	0.3

Note: Overall attrition rates are given as percentages. Differential attrition rates are given as percentage points. Not every combination of differential and overall attrition is possible for any given study. N/A is not applicable (if the total attrition rate is 58 percent or higher, no differential attrition rate will yield low attrition under the cautious boundary).

APPENDIX D. GLOSSARY OF SYMBOLS FOR STATISTICAL FORMULAS

The following technical appendices include the statistical formulas that guide the What Works Clearinghouse's (WWC's) review procedures. Table D.1 provides a glossary of symbols used in these formulas.

Table D.1. *Glossary of statistical formula symbols*

Symbol	Description	Equation number
β	Standardized regression coefficient for intervention impact estimate	E.15, E.16
$\hat{\beta}_1^{biased}$	Biased estimator of the true impact of an intervention	G.2
η	Design effect for a cluster-level assignment study	E.22, E.23, E.25, E.26, E.27, E.29, E.30, E.31, E.33, E.38
γ	Small numbers of clusters correction for a cluster-level assignment study	E.20, E.24, E.32, E.33
$\hat{\Delta}^{complier}$	Differential attrition rate for compliers	G.6, G.8
$\hat{\Delta}_{final}^{complier}$	Final differential attrition rate for all compliers for studies with three or more assignment groups	G.10
$\hat{\Delta}_{g,g-1}^{complier}$	Differential attrition rate for compliers pertaining to the comparison between groups (g - 1) and g	G.10
λ_j	Deviations from missing at random assumption for group j	H.1, H.3
ρ_{cor}	Correlation between the pretest and posttest measures	E.19, H.2, H.3, H.6, H.7, H.8, H.9, H.15, H.16, H.17, H.18, H.20, H.22, H.23, H.24, H.25, H.26, H.28, H.29, H.30, H.32, H.34, H.35, H.36, H.37, H.38, H.40, H.41, H.42
ρ	Average correlation among outcome measures	F.4
ρ_{ICC}	Intraclass correlation coefficient	E.20, E.21, E.22
$\hat{\sigma}^2$	Estimated level-1 variance for variability within individuals	E.40, E.41, E.42
$\hat{\tau}^2$	Estimated level-2 variance for variability across individuals	E.40, E.41, E.42
ω	Small sample bias correction term	E.1, E.3, E.5, E.6, E.7, E.8, E.9, E.11, E.12, E.13, E.16, E.20, E.23, E.24, E.25, E.26, E.27, E.29, E.30, E.31, E.32, E.40, E.42, H.7, H.8, H.9, H.16, H.17, H.18, H.22, H.23, H.24, H.25, H.27, H.28, H.29, H.30, H.34, H.35, H.36, H.37, H.39, H.40, H.41, H.42
$\Phi(\cdot)$	Cumulative distribution function of the standard normal distribution	G.2, G.4, G.5
$\Phi^{-1}(\cdot)$	Inverse of the cumulative distribution function of the standard normal distribution	G.3

Symbol	Description	Equation number
$\bar{A}_c$	Attrition rate in the comparison group	G.6
$\bar{A}_{c0}$	Attrition rate in the comparison group of those who did not receive the intervention (no take-up)	G.7
$\bar{A}_{i,}$	Attrition rate in the intervention group	G.6
$\bar{A}_{i0}$	Attrition rate in the intervention group of those who did not receive the intervention (no take-up)	G.7
a	Regression intercept for the average outcome in absence of the intervention	E.39
B1, B2, B3	The largest bias in the outcome effect size due to deviation from the missing-at-random assumption when the baseline measure is observed for all subjects	H.7, H.8, H.9
B1*, B2*, B3*	The largest bias in the outcome effect size due to deviation from the missing-at-random assumption when the baseline measure is imputed or missing for some subjects	H.16, H.17, H.18
b	Unstandardized intervention-comparison group mean difference of the outcome	E.1, E.14, E.15, E.20, E.27, E.39, E.40
$bias$	Expected bias in standard deviation units	G.2, G.4, H.6, H.15, H.26, H.38
C1, C2, C3, C4	Bounds on the baseline effect size using the complete-case baseline effect size and assuming no missing outcome data	H.22, H.23, H.24, H.25
C1*, C2*, C3*, C4*	Bounds on the baseline effect size using the complete-case baseline effect size and assuming missing outcome data	H.34, H.35, H.36, H.37
c	Unstandardized coefficient from a regression of the outcome y on baseline measure x	H.4, H.5, H.13
D1, D2, D3, D4	Bounds on the baseline effect size using the imputed baseline effect size and assuming no missing outcome data	H.27, H.28, H.29, H.30
D1*, D2*, D3*, D4*	Bounds on the baseline effect size using the imputed baseline effect size and assuming missing outcome data	H.39, H.40, H.41, H.42
$\bar{D}_{c,an}$	Intervention take-up rate for individuals in the comparison analytic sample	G.1
$\bar{D}_{c,ran}$	Intervention take-up rate for individuals assigned to the comparison group	G.6, G.7
$\bar{D}_{i,an}$	Intervention take-up rate for individuals in the intervention analytic sample	G.1
$\bar{D}_{i,ran}$	Intervention take-up rate for individuals assigned to the intervention group	G.6, G.7
$\bar{D}_{j,ran}$	Intervention take-up rate for individuals assigned to group j	G.11
$\bar{D}_{ran}$	Intervention take-up rate for the entire random assigned sample	G.11

Symbol	Description	Equation number
df	Degrees of freedom	E.3, E.4, E.21, E.23, E.24, E.25, E.26, E.28, E.29, E.30, E.41, E.42
e_{ij}	Level-1 error term for case *i* at time *j* in a multilevel model	E.39
F	*F* statistic	E.8, E.13, G.1
$f_c(\rho_{cor})$	Function of the baseline-outcome correlation for the comparison group used in relating the comparison group complete case outcome mean to the full-sample mean	H.5, H.6, H.14, H.15, H.21, H.26, H.33, H.38
$f_i(\rho_{cor})$	Function of the baseline-outcome correlation for the intervention group used in relating the intervention group complete case outcome mean to the full-sample mean	H.5, H.6, H.14, H.15, H.21, H.26, H.33, H.38
$f_j(\rho_{cor})$	Function of the baseline-outcome correlation for group *j* used in relating the complete case outcome mean to the full-sample mean	H.2, H.3, H.11, H.12, H.19, H.31
G	Number of mutually exclusive subgroups	F.1, F.2, G.10, G.11
$\bar{G}$	Fixed-effects meta-analytic average effect size	F.6
$\bar{G}^*$	Reweighted meta-analytic average effect size such that studies rated *Meets WWC Standards Without Reservations* receive most of the meta-analytic weight	F.8
g	Hedge's *g* standardized effect size	E.1, E.5, E.6, E.7, E.8, E.9, E.11, E.12, E.13, E.16, E.20, E.23, E.24, E.25, E.26, E.27, E.29, E.30, E.31, E.32, E.36, E.40, E.42, E.43
g_i	Effect size for the *i*th main finding in a study	F.3
g_{MAR}	Outcome effect size obtained using an imputation method based on the missing-at-random assumption	H.4, H.13
g_{NMAR}	Outcome effect size accounting for when the missing-at-random assumption does not hold	H.5, H.14
g_{xI}	Baseline effect size based on imputed data	H.27, H.28, H.29, H.30, H.39, H.40, H.41, H.42
g_{xMAR}	Baseline effect size obtained using an imputation method based on the missing-at-random assumption	H.20, H.21, H.32
g_{xNMAR}	Baseline effect size accounting for when the missing-at-random assumption does not hold	H.33
g_{xR}	Baseline effect size based on complete-case baseline data	H.20, H.21, H.22, H.23, H.24, H.25
$g_{xR(xy)}$	Baseline effect size based on complete-case baseline and outcome data	H.32, H.33, H.34, H.35, H.36, H.37
$\bar{g}$	Domain-level composite effect size for a study	F.3
$\bar{g}_s$	Domain-level composite effect size for study *s*	F.6, F.8

Symbol	Description	Equation number
$h(.)$	Function that relates the probability of observing an outcome given a baseline and outcome value	H.1
$I(.)$	Indicator function equal to 1 if the logical statement inside the function is true and 0 if the statement is false	I.1, I.2, I.3, I.4
J	Total number of studies	F.6, F.7, F.8, F.9
K	Total number of main findings meeting WWC standards for an outcome domain in a single study	F.3, F.4
M	Total number of clusters in a cluster-level assignment study	E.20, E.21, E.22, E.27, E.28, E.30
MDES	Minimum effect size that can be detected using a two-tailed test	G.3
m_A	Number of data points in the A series ("baseline" phase) in single-case designs	I.1, I.2, I.3, I.4
m_B	Number of data points in the B series ("intervention" phase) in single-case designs	I.1, I.2
N	Total number of individuals in the analytic sample	E.4, E.14, E.16, E.20, E.21, E.22, E.27, E.30
N_{ran}	Total number of individuals randomly assigned	G.11
NAP	Nonoverlap of all pairs in single-case designs	I.1, I.2, I.3, I.4
n_c	Number of individuals in the comparison group analytic sample	E.2, E.5, E.6, E.7, E.8, E.9, E.11, E.12, E.13, E.14, E.16, E.17, E.18, E.23, E.26, E.29, E.31, E.32, E.33, E.34, E.37, E.38, G.1, H.13, H.14, H.15, H.17, H.18, H.32, H.33, H.34, H.35, H.36, H.37, H.38, H.41, H.42
$n_{c,ran}$	Number of individuals randomly assigned to the comparison group	G.9
n_{cx}	Number of individuals in the comparison group with observed baseline data	H.13, H.14, H.15, H.17, H.18
n_{cy}	Number of individuals in the comparison group with observed outcome data	H.32, H.33, H.34, H.35, H.36, H.37, H.38, H.41, H.42
n_{gc}	Number of individuals in subsample g in the comparison group	F.2
n_{gi}	Number of individuals in subsample g in the intervention group	F.2
n_{gj}	Number of individuals in subsample g in group j	F.1
n_i	Number of individuals in the intervention group analytic sample	E.2, E.5, E.6, E.7, E.8, E.9, E.11, E.12, E.13, E.14, E.16, E.17, E.18, E.23, E.26, E.29, E.31, E.32, E.33, E.34, E.37, E.38, G.1, H.13, H.14, H.15, H.16, H.18, H.32, H.33, H.34, H.35, H.36, H.37, H.38, H.40, H.42

Symbol	Description	Equation number
$n_{i,ran}$	Number of individuals randomly assigned to the intervention group	G.9
n_{ix}	Number of individuals in the intervention group with observed baseline data	H.13, H.14, H.15, H.16, H.18
n_{iy}	Number of individuals in the intervention group with observed outcome data	H.32, H.33, H.34, H.35, H.36, H.37, H.38, H.40, H.42
n_j	Number of individuals in the analytic sample for group j	H.10, H.11, H.12, H.31
$n_{j,ran}$	Number of individuals assigned to group j	G.11
n_{jx}	Number of individuals in group j with observed baseline data	H.10, H.11, H.12
n_{jy}	Number of individuals in group j with observed outcome data	H.31
$Odds_c$	Odds for the comparison group	E.35, E.36
$Odds_i$	Odds for the intervention group	E.35, E.36
OR	Ratio between the odds for the two groups	E.35, E.36
P_c	Probability of a positive outcome combined across all subsamples in the comparison group	F.2
P_i	Probability of a positive outcome combined across all subsamples in the intervention group	F.2
p_c	Average probability of a positive outcome for comparison group	E.35, E.37, E.38
p_{gc}	Probability of a positive outcome in subsample g in the comparison group	F.2
p_{gi}	Probability of a positive outcome in subsample g in the intervention group	F.2
p_i	Average probability of a positive outcome for intervention group	E.35, E.37, E.38
$p_j(x,y)$	Probability of observing an outcome for group j given baseline value x and outcome value y	H.1
R^2	Multiple correlation between the covariates and the outcome	E.11, E.12, E.13, E.18, E.26
$\hat{R}_c^{complier}$	Attrition rate for compliers in the comparison group	G.7, G.8, G.9
$\hat{R}_i^{complier}$	Attrition rate for compliers in the intervention group	G.8, G.9
$\hat{R}_{overall}^{complier}$	Overall attrition rate for compliers	G.9
SD	Total individual-level standard deviation of the outcome	E.14
SD_B	Pooled within-group, cluster-level standard deviation of the outcome	E.27, E.30
SD_c	Individual-level standard deviation of the comparison group	E.2, E.5, E.17, E.18, E.33, E.34
SD_{gj}	Outcome standard deviation for subsample g in assignment group j	F.1
SD_i	Individual-level standard deviation of the intervention group	E.2, E.5, E.17, E.18, E.33, E.34
SD_j	Outcome standard deviation combined across all subsamples for assignment group j	F.1

Symbol	Description	Equation number
SD_p	Pooled within-group, individual-level standard deviation of the outcome	E.1, E.2, E.9, E.14, E.19, E.20, E.24, E.25
$SD_{p,gain}$	Pooled within-group, individual-level standard deviation of the pre-post gain score	E.19
SD_x	Standard deviation of the baseline measure	E.15, H.1, H.2, H.5, H.6, H.7, H.8, H.9, H.11, H.12, H.14, H.15, H.16, H.17, H.18, H.19, H.31, H.32, H.33, H.34, H.35, H.36, H.37
SD_y	Standard deviation of the outcome measure	E.15, H.1, H.2, H.4, H.5, H.11, H.12, H.13, H.14, H.19, H.20, H.21, H.22, H.23, H.24, H.25, H.26, H.28, H.29, H.30, H.31, H.32, H.33, H.34, H.35, H.36, H.37, H.38, H.40, H.41, H.42
$SE[b]$	Standard error of the unstandardized mean difference	E.9, E.10, E.30, E.42
$SE_{CC}[b]$	Cluster-corrected standard error of the unstandardized mean difference in a cluster-level assignment study	E.24
$SE_{UC}[b]$	Standard error uncorrected for clustering of the unstandardized mean difference in a cluster-level assignment study	E.25
$SE[g]$	Standard error of the Hedges' *g* effect size	E.6, E.9, E.11, E.23, E.24, E.25, E.26, E.29, E.30, E.37, E.38, E.42, E.43, G.3
$SE[\bar{g}]$	Standard error of the domain-level average effect size in a single study	F.4
$SE[g_i]$	Standard error of the *i*th effect size in a study	F.4
$SE[\bar{g}_s]$	Standard error of the domain-level average effect size for study *s*	F.5, F.9
$SE[\bar{G}]$	Standard error of the fixed-effects meta-analytic average	F.7
$SE[\bar{G}^*]$	Standard error of the reweighted meta-analytic average such that studies rated *Meets WWC Standards Without Reservations* receive most of the meta-analytic weight	F.9
$SE[y_i]$	Standard error of the intervention group mean	E.10, E.17, E.18
$SE[y_c]$	Standard error of the comparison group mean	E.10, E.17, E.18
$SE_{CC}[y_i]$	Cluster-corrected standard error of the intervention group mean in a cluster-level assignment study	E.33
$SE_{CC}[y_c]$	Cluster-corrected standard error of the comparison group mean in a cluster-level assignment study	E.33
$SE_{UC}[y_i]$	Standard error uncorrected for clustering of the intervention group mean in a cluster-level assignment study	E.34

Symbol	Description	Equation number
$SE_{UC}[y_c]$	Standard error uncorrected for clustering of the comparison group mean in a cluster-level assignment study	E.34
$SE[\hat{\beta}_1^{biased}]$	Standard error for the biased estimator of the true impact of an intervention	G.2
T_{ij}	Dummy indicator for receiving the intervention for case i at time j	E.39
t	t test statistic for group mean difference	E.7, E.12, E.43
t_{CC}	Cluster-corrected t test statistic for group mean difference in cluster-level assignment study	E.31
t_{UC}	t test statistic uncorrected for clustering for group mean difference in cluster-level assignment study	E.32
u_i	Level-2 error term for case i in a multilevel model	E.39
$V[\bar{g}_s]$	Variance of the domain-level average effect size for study s	F.5
W_s	Meta-analytic inverse variance weight for study s	F.5, F.6, F.7
W_s^*	Rescaled meta-analytic weight for study s that ensures that most the meta-analytic weight goes to studies rated *Meets WWC Standards Without Reservations*	F.8, F.9
w_g	Weight on the comparison between groups (g - 1) and g	G.10, G.11
x	Value of the baseline measure for a subject	H.1
$\bar{x}_c$	Full-sample baseline mean for the comparison group	H.4, H.5, H.6, H.7, H.9
$\bar{x}_{cR}$	Complete-case baseline means for the comparison group	H.4, H.5, H.6, H.7, H.9
$\bar{x}_i$	Full-sample baseline mean for the intervention group	H.4, H.5, H.6, H.8, H.9
$\bar{x}_{cx}$	Comparison baseline mean for individuals with observed baseline data but possibly missing outcome data	H.13, H.14, H.15, H.17, H.18
$\bar{x}_{cxy}$	Comparison baseline mean for individuals with observed baseline and outcome data	H.13, H.14, H.15, H.17, H.18, H.32, H.33, H.34, H.35, H.36, H.37
$\bar{x}_{c\sim y}$	Comparison group baseline mean for sample with observed outcome data	H.32, H.33, H.34, H.35, H.36, H.37
$\bar{x}_{iR}$	Complete-case baseline mean for the intervention group	H.4, H.5, H.6, H.8, H.9
$\bar{x}_{ix}$	Intervention baseline mean for individuals with observed baseline data but possibly missing outcome data	H.13, H.14, H.15, H.16, H.18
$\bar{x}_{ixy}$	Intervention baseline mean for individuals with observed baseline and outcome data	H.13, H.14, H.15, H.16, H.18, H.32, H.33, H.34, H.35, H.36, H.37
$\bar{x}_{i\sim y}$	Intervention group baseline mean for sample with observed outcome data	H.32, H.33, H.34, H.35, H.36, H.37
$\bar{x}_j$	Full-sample baseline mean for group j	H.2, H.19, H.31
$\bar{x}_{jR}$	Complete-case baseline mean for group j	H.2, H.19

Symbol	Description	Equation number
$\bar{x}_{jx}$	Baseline mean for individuals in group *j* with observed baseline data but possibly missing outcome data	H.11, H.12
$\bar{x}_{j\sim y}$	Baseline mean for sample *j* with observed outcome data	H.31
$\bar{x}_{jxy}$	Baseline mean for individuals in group *j* with observed baseline and outcome data	H.11, H.12, H.31
$\bar{Y}_j$	Combined group mean across all subsamples in assignment group *j*	F.1
y	Value of the outcome measure for a subject	H.1
y_{ij}	Observation of case *i* at time *j*	E.39
y_i^A	Data point *i* in the A series ("baseline" phase) in single-case designs	I.1, I.2, I.3, I.4
y_j^B	Data point *j* in the B series ("intervention" phase) in single-case designs	I.1, I.2
$\bar{y}_c$	Comparison group mean	E.5, H.20, H.21, H.22, H.23, H.24, H.25, H.26, H.29, H.30
$\bar{y}_{cR}$	Complete case outcome mean for the comparison group	H.4, H.5, H.20, H.21, H.22, H.23, H.24, H.25, H.26, H.29, H.30
$\bar{y}_{cxy}$	Outcome mean for the individuals in the comparison group with observed baseline and outcome data	H.13, H.14, H.32, H.33, H.34, H.35, H.36, H.37, H.38, H.41, H.42
$\bar{y}_{cy}$	Comparison group outcome mean for sample with observed outcome data but possibly missing baseline data	H.32, H.33, H.34, H.35, H.36, H.37, H.38, H.41, H.42
$\bar{y}_{c\sim x}$	Outcome mean for the individuals in the comparison group analytic sample missing the baseline measure	H.13, H.14
$\bar{y}_{gj}$	Outcome mean for subsample *g* in assignment group *j*	F.1
$\bar{y}_i$	Intervention group mean	E.5, H.20, H.21, H.22, H.23, H.24, H.25, H.26, H.28, H.30
$\bar{y}_{iR}$	Complete case outcome mean for the intervention group	H.4, H.5, H.20, H.21, H.22, H.23, H.24, H.25, H.26, H.28, H.30
$\bar{y}_{ixy}$	Outcome mean for the individuals in the intervention group with observed baseline and outcome data	H.13, H.14, H.32, H.33, H.34, H.35, H.36, H.37, H.38, H.40, H.42
$\bar{y}_{iy}$	Intervention group outcome mean for sample with observed outcome data but possibly missing baseline data	H.32, H.33, H.34, H.35, H.36, H.37, H.38, H.40, H.42
$\bar{y}_{i\sim x}$	Outcome mean for the individuals in the intervention group analytic sample missing the baseline measure	H.13, H.14
$\bar{y}_j$	Full-sample outcome mean for group *j*	H.2, H.10, H.11, H.12, H.19
$\bar{y}_{jR}$	Complete case outcome mean for group *j*	H.2, H.19
$\bar{y}_{jx}$	Outcome mean for the individuals in the analytic sample for group *j* with observed baseline data	H.10

Symbol	Description	Equation number
$\bar{y}_{jy}$	Outcome mean for sample *j* with observed outcome data but possibly missing baseline data	H.31
$\bar{y}_{jxy}$	Outcome mean for the individuals in group *j* with observed baseline and outcome data	H.11, H.12, H.31
$\bar{y}_{j\sim x}$	Outcome mean for the individuals in the analytic sample for group *j* missing the baseline measure	H.10, H.11, H.12
z_q	*q*th quantile of the standard normal distribution	G.2, G.4, G.5

APPENDIX E. STATISTICAL FORMULAS FOR EACH FINDING IN A STUDY

Study authors may report their analyses in many ways, with varying degrees of comparability and utility. The What Works Clearinghouse (WWC) aims to report study findings in a consistent way, using a common metric and accounting for differences across analyses that may affect their results. This appendix describes how the WWC computes effect sizes, which capture the magnitude of intervention effects, and standard errors, which capture the uncertainty in estimating those effects.

This appendix includes the following sections: (1) individual-level assignment studies with continuous outcomes, (2) cluster-level assignment studies with continuous outcomes, (3) dichotomous outcomes, (4) difference-in-differences adjustments, (5) design-comparable effect sizes from single-case designs, (6) regression discontinuity designs, (7) computing *p* values and statistical significance, and (8) computing the improvement index.

Individual-level assignment studies with continuous outcomes

The most straightforward scenario for computing WWC effect sizes and standard errors is for individual-level assignment studies with continuous outcomes. For instance, a study could compare means across an intervention group and a comparison group for a continuous measure of student reading achievement. One challenge in characterizing the intervention effect is comparability across measures. The raw mean difference may not be comparable to mean differences for other reading achievement measures with different scales or scoring procedures.

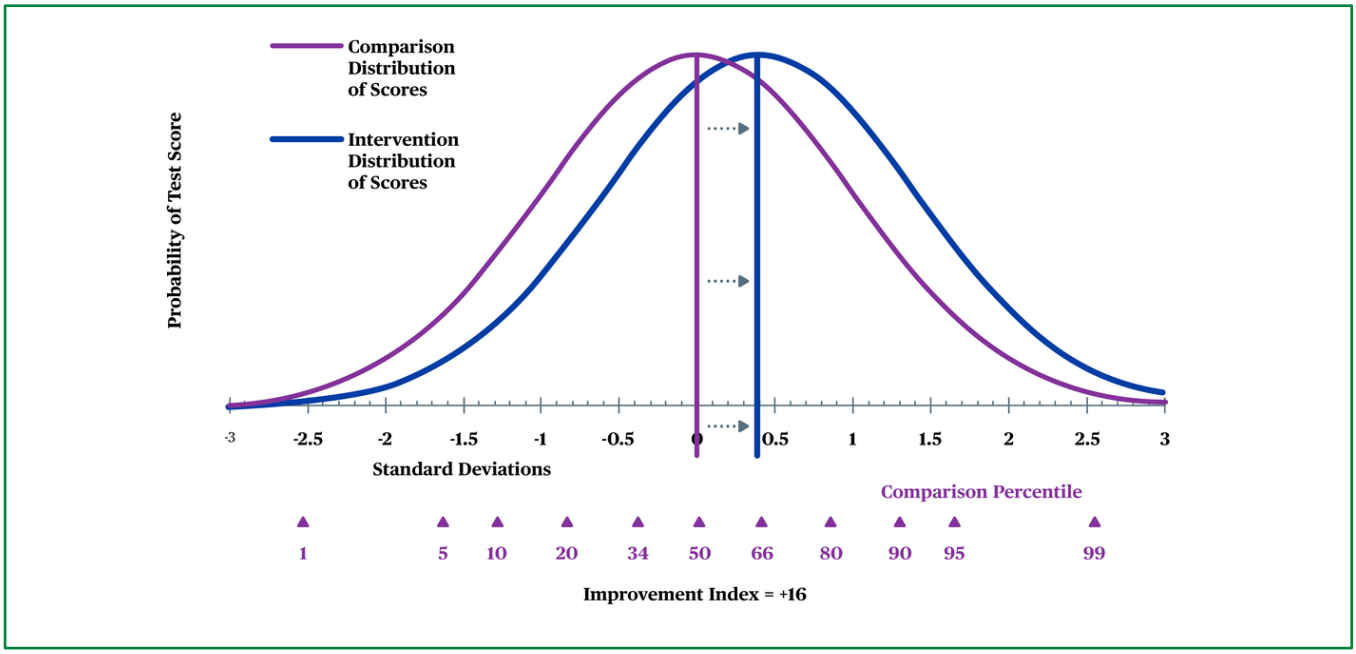
The WWC uses the Hedges' *g* effect size metric as a standardized measure of intervention effects. This metric represents the mean difference between intervention and comparison groups in standard deviation units. That is, the raw mean difference is divided by the variability within the groups, placing effect sizes on a common scale. This approach allows the WWC to synthesize findings across outcome measures and studies.

Information to compute Hedges' *g*

- An estimate of the unstandardized mean difference (preferably based on covariate adjustment)
- Standard deviations separately by group (always based on unadjusted statistics)
- Sample sizes separately by group

Figure E.1 graphically represents an effect size of $g = 0.40$, such as improving reading test scores by 0.4 standard deviation. The intervention group distribution (dotted line) is shifted to the right relative to the comparison group distribution (solid line). Consider the magnitude of this effect for a typical student who scores at the median in the comparison group. An improvement of 0.4 standard deviation would increase this student's score from the 50th percentile to the 66th percentile. The WWC calls this increase of +16 in the percentile rank an improvement index, as detailed in more depth later in this appendix.

Figure E.1. *An improvement of 0.4 standard deviation*



General effect size formulas

The WWC computes the Hedges' *g* effect size by dividing an estimate of the unstandardized mean difference *b* by the pooled within-group standard deviation SD_p (a type of weighted average of the variability within the intervention and comparison groups). For the unstandardized mean difference *b*, the WWC prefers covariate-adjusted estimates that control for baseline differences. These estimates are often more precise and have less bias than unadjusted mean differences. However, the standard deviations used to compute the standardized effect size should always be based on unadjusted statistics to ensure comparability across studies. Standard deviations for an outcome from national, state, or district norms are also accepted by the WWC.

The Hedges' *g* computation also includes multiplication by a small-sample correction factor ω, which is usually close to 1 but slightly less than 1 when the sample size is small (Hedges, 1981). This correction factor is needed to produce unbiased estimates of the population effect size. The general formula for computing Hedges' *g* for continuous outcomes in individual-level assignment studies can be written as follows:

[E.1] $$g = \frac{\omega b}{SD_p}$$

Formulas for the pooled standard deviation SD_p and small-sample correction factor ω are given by the following:

[E.2] $$SD_p = \sqrt{\frac{(n_i - 1)SD_i^2 + (n_c - 1)SD_c^2}{n_i + n_c - 2}}$$

[E.3] $$\omega = 1 - \frac{3}{4df - 1}$$

where n_i and n_c are the sample sizes for the intervention and comparison groups, SD_i and SD_c are the intervention and comparison standard deviations, and df is the degrees of freedom. The following formula provides the degrees of freedom for individual-level assignment studies:

[E.4] $$df = N - 2$$

where $N = n_i + n_c$ is the total number of individuals. The Hedges' g metric is very similar to the well-known Cohen's d metric, except that Hedges' g corrects for small-sample bias with the ω term. The WWC always applies this additional term, including for larger samples. However, the term will be very close to 1 in larger samples. For instance, this correction factor will be ω = 0.99 for an individual-level assignment study with a total sample size of 100 students. If the study authors reported a bias-uncorrected standardized effect size (such as Cohen's d), the WWC will apply the bias correction term ω to the author-reported effect size if the effect size cannot be computed another way (such as using the original means and standard deviations).

The following subsections present variations of the preceding formulas that account for differences in how study authors may report their findings for individual-level assignment studies with continuous outcomes. Later sections in this appendix address extensions for cluster-level assignment studies and dichotomous outcomes.

Unadjusted mean differences

The WWC allows using unadjusted means to compute effect sizes in some limited scenarios such as for randomized controlled trials (RCTs) with a low risk of bias due to compositional change. Unadjusted means are acceptable, although not preferred, in such cases because those studies' designs limit bias in estimating intervention effects. Such RCTs yield unbiased estimates on average, even without baseline adjustment.

The WWC will therefore use unadjusted means to compute effect sizes if (a) the study did not report sufficient information to compute covariate-adjusted effect sizes and (b) statistical adjustment for baseline differences is not required to meet WWC standards. Substituting the unadjusted mean difference, $\bar{y}_i - \bar{y}_c$, as b in equation E.1 yields the Hedges' g effect size:

[E.5] $$g = \frac{\omega(\bar{y}_i - \bar{y}_c)}{\sqrt{\frac{(n_i - 1)SD_i^2 + (n_c - 1)SD_c^2}{n_i + n_c - 2}}}$$

This equation includes the formula for the pooled standard deviation in the denominator to explicitly show what statistics are required to compute the Hedges' g value with this approach. This formula cannot be used if the study authors do not provide one of the statistics included in it, such as the intervention and comparison group standard deviations.

The following formula provides the standard error for effect sizes calculated using unadjusted means (Borenstein & Hedges, 2019)[35]:

$$[E.6] \qquad SE[g] = \omega \sqrt{\frac{n_i + n_c}{n_i n_c} + \frac{g^2}{2(n_i + n_c)}}$$

The standard error expresses the uncertainty in the effect size estimate. Larger standard errors indicate greater uncertainty (and typically smaller sample sizes). The next appendix describes how the WWC uses the standard error to compute the statistical significance of study findings and their meta-analytic weight in cross-study syntheses.

Consider an RCT with a low risk of bias due to compositional change that gives the following reading achievement statistics in table E.1:

Table E.1. *Descriptive statistics for a low-attrition randomized controlled trial*

Group	*M*	*SD*	*n*
Intervention	59.12	12.32	39
Comparison	54.34	11.57	41

M is mean, *SD* is standard deviation, and *n* is sample size.

The small-sample correction term is $\omega = 1 - \frac{3}{4(39+41-2)-1} = 0.99$.

The Hedges' g is calculated as follows:

$$g = \frac{0.99(59.12 - 54.34)}{\sqrt{\frac{(39-1)12.32^2 + (41-1)11.57^2}{39+41-2}}} = \frac{0.99 \times 4.78}{11.94} = 0.40$$

The standard error is then:

$$SE[g] = 0.99 \sqrt{\frac{39+41}{39 \times 41} + \frac{0.40^2}{2(39+41)}} = 0.22$$

Test statistics for unadjusted mean comparisons

For RCTs with low attrition, when means or standard deviations are not reported, the WWC can compute Hedges' g based on t test or analysis of variance (ANOVA) F test results. The following formulas in this subsection

[35] The WWC uses the bias-corrected effect size g to compute the right-hand g^2 term in the standard error formula. This approach contrasts slightly with Borenstein and Hedges' (2019) approach, which uses the effect size uncorrected for bias (Cohen's d) to compute that right-hand term. Both approaches are approximations because the population variance formula uses the population effect size (Hedges & Olkin, 1985), which is not known in practice. These distinctions tend not to matter much because the left-hand term is usually a much larger contributor to the variance than the right-hand term (and ω is usually close to 1).

apply to only simple unadjusted between-group comparisons of outcome means (separate formulas apply to test statistics based on covariate adjustment).

This following formula applies to effect sizes based on *t* test results:

[E.7] $$g = \omega t \sqrt{\frac{n_i + n_c}{n_i n_c}}$$

This following formula applies to effect sizes based on ANOVA *F* test results:

[E.8] $$g = \omega \sqrt{\frac{F(n_i + n_c)}{n_i n_c}}$$

where the sign is determined by the sign of the main difference. The appropriate standard error formula for both formulas remains equation E.6, which assumes an effect size based on unadjusted comparisons.

The example study in table E.1 could have instead reported the test statistics $t = 1.79$ or $F = 3.20$; the effect size would still be calculated as $g = 0.40$ with those alternative statistics.

Last, an exact two-tailed *p* value can be converted into a *t* statistic and then used in equation E.7. For instance, a *p* value of .077 for a total sample size of 80 students (see table E.1) corresponds to $t = 1.79$, using 78 as the degrees of freedom (see equation E.4). One-tailed *p* values can be converted into two-tailed *p* values by multiplying by two. The WWC, however, will not apply this *p*-to-*t* conversion for inexact *p* values such as "$p < .05$" because they can correspond to a range of effect sizes.

Covariate-adjusted mean differences

The preceding formulas based on unadjusted mean comparisons cannot be used for studies that require baseline adjustment, such as for quasi-experimental designs (QEDs) and high-attrition RCTs. Effect sizes from those studies must instead use covariate-adjusted statistics that control for baseline differences. The WWC also prefers covariate-adjusted effect sizes for RCTs with a low risk of bias due to compositional change to increase the precision of the effect size estimate.

Using covariate-adjusted means to compute effect sizes relies on the same general Hedges' *g* formula introduced in equation E.1. The unstandardized mean difference *b* would instead be based on the covariate-adjusted mean difference. For instance, the example study in table E.1 could have used multiple regression to control for baseline reading achievement, reporting an unstandardized impact estimate (regression coefficient) of 3.59 for the covariate-adjusted mean difference. The Hedges' *g* would be computed as follows:

$$g = \frac{0.99 \times 3.59}{11.94} = 0.30$$

The only difference with the earlier computation based on unadjusted means is that the mean difference *b* is now the regression coefficient as opposed to the unadjusted mean difference. Note that the small-sample

correction ω and the pooled within-group standard deviation remain the same (for instance, the pooled standard deviation remains based on unadjusted statistics).

Alternatively, the study could have conducted an analysis of covariance (ANCOVA) and reported covariate-adjusted means of 58.53 and 54.94 for the intervention and comparison groups, respectively. The covariate-adjusted mean difference would remain 3.59 in that case and the covariate-adjusted Hedges' g would remain the same as well.

If the study authors reported the standard error $SE[b]$ for the unstandardized mean difference b, then the standard error for the standardized Hedges' g effect size can be computed as follows:

[E.9] $$SE[g] = \omega \sqrt{\left(\frac{SE[b]}{SD_p}\right)^2 + \frac{g^2}{2(n_i + n_c)}}$$

The left-hand term in this formula standardizes the unstandardized standard error by dividing by the pooled standard deviation. The right-hand term is the same as in equation E.6 and reflects the uncertainty introduced by using the pooled standard deviation (a sample statistic) in the effect size computation. Consider if the example study in table E.1 reported the regression coefficient standard error as 2.07. The effect size standard error would then be:

$$SE[g] = 0.99 \sqrt{\left(\frac{2.07}{11.94}\right)^2 + \frac{0.40^2}{2(39 + 41)}} = 0.17$$

Note that this value of 0.17 is smaller than the standard error of 0.22 computed earlier for the effect size based on unadjusted means. This difference reflects that covariate adjustment typically reduces uncertainty in the intervention effect estimate, reducing the standard error and reflecting increased precision.

If the t statistic for an unstandardized regression coefficient is reported instead, the standard error for the unstandardized mean differences can be calculated using $SE[b] = b/t$. Studies using ANCOVA also may report the intervention mean standard error $SE[y_i]$ and comparison mean standard error $SE[y_c]$. In this case, the standard error of the mean difference can be computed as follows and substituted into equation E.9:

[E.10] $$SE[b] = \sqrt{SE[y_i]^2 + SE[y_c]^2}$$

A study may not report sufficient information to determine the covariate-adjusted standard error of the mean difference, rendering equation E.9 unusable. However, if the study authors reported the multiple correlation R^2 between the covariates and the outcome, the covariate-adjusted standard error for the effect size can instead be calculated as follows:

[E.11] $$SE[g] = \omega \sqrt{\frac{n_i + n_c}{n_i n_c}(1 - R^2) + \frac{g^2}{2(n_i + n_c)}}$$

Compared with unadjusted R^2 values, the WWC prefers adjusted R^2 values that account for the number of entered predictors and correct for overconfidence in the model predictions. The WWC will treat negative adjusted R^2 values as 0 percent. However, the WWC will use unadjusted R^2 values if the adjusted value is unavailable. Consider if the example study authors reported the adjusted R^2 value as 40 percent. The effect size standard error would be computed as 0.17 using equation E.11.

If the unstandardized standard error and R^2 value are both unavailable, then the WWC will take a cautious approach to calculating the effect size standard error and assume a value of 0 percent for R^2 using equation E.11. This cautious approach will overestimate the magnitude of the standard error but protects against type I error. Table E.2 summarizes the prioritization of different formulas based on what statistics the study authors report.

Table E.2. *Use cases for covariate-adjusted standard error formulas*

SE[*b*]	R^2	Appropriate standard error formula
✓ Reported[a]	✓ Reported	Use equation E.9
✓ Reported[a]	✗ Not reported	Use equation E.9
✗ Not reported	✓ Reported	Use equation E.11
✗ Not reported	✗ Not reported	Use equation E.11 imputing R^2 = 0 percent

a. Includes cases in which *SE*[*b*] can be derived from the standard errors of group means.

Test statistics for covariate-adjusted mean comparisons

A study could conduct a covariate-adjusted analysis such as multiple regression but not report sufficient information to compute an effect size using the preceding formulas. For instance, the study article could leave the unadjusted standard deviations unreported. In such cases, the WWC could compute the Hedges' *g* effect size using a covariate-adjusted *t* test or *F* statistic as follows:

[E.12] $$g = \omega t \sqrt{\frac{n_i + n_c}{n_i n_c}(1 - R^2)}$$

[E.13] $$g = \omega \sqrt{\frac{F(n_i + n_c)}{n_i n_c}(1 - R^2)}$$

Using these formulas requires that the study authors report the multiple correlation R^2. This requirement is needed because covariate-adjusted test statistics are based on the covariate-adjusted within-group variance, not the unadjusted within-group variance. The $1 - R^2$ term corrects for this issue, making the effect size comparable with other effect sizes based on unadjusted standard deviations. As noted earlier, the WWC also will allow converting exact two-tailed *p* values into *t* statistics to apply these formulas, but the WWC will not apply the *p*-to-*t* conversions for inexact *p* values such as $p < .05$.

Alternative ways to compute the pooled standard deviation

The preceding sections introduce alternative approaches (based on test statistics) to compute the Hedges' g effect size if the pooled within-group standard deviation cannot be calculated directly using equation E.2. This section further expands on this idea to describe alternative ways of computing the pooled standard deviation SD_p. These conversions aim to yield standard deviations of the outcome measure collected at the follow-up time point without adjustment for baseline measures.

Total standard deviation. A study might report the total standard deviation but not the standard deviation for each group separately. The total standard deviation can be converted to the pooled within-group standard deviation as follows:

$$SD_p = \sqrt{\frac{N-1}{N-2}SD^2 - \frac{n_i n_c}{N(N-2)}b^2} \quad \text{[E.14]}$$

where $N = n_i + n_c$ is the total study sample size. This transformation is exact if b is the unadjusted mean difference (not adjusted for covariates) and nearly equivalent if b is the covariate-adjusted mean difference. The WWC will allow using this formula in both cases.

Consider if the example study in table E.1 reported the total standard deviation as $SD = 12.11$ but did not report the standard deviation separately by group. The pooled within-group standard deviation would be calculated as follows (using $b = 4.78$ as the unadjusted mean difference):

$$SD_p = \sqrt{\frac{80-1}{80-2}12.11^2 - \frac{39 \times 41}{80(80-2)}4.78^2} = 11.94$$

This pooled within-group standard deviation of 11.94 is identical to the value computed using equation E.2 with the standard deviations separately by group.

Standardized z scores. A study might standardize the outcome response scale by subtracting the grand mean and dividing by the total (unadjusted) standard deviation to generate what are often called z scores. The WWC can substitute $SD = 1$ into equation E.14 to address this case.

Standardized regression coefficients. A study might report standardized regression coefficients (sometimes called beta weights). These standardized coefficients differ from unstandardized coefficients based on z scores. To illustrate why, the standardized coefficient β for an outcome with total standard deviation SD_y and predictor with standard deviation SD_x can be written as follows:

$$\beta = b\frac{SD_x}{SD_y} \quad \text{[E.15]}$$

This formula illustrates that standardized coefficients and z scores both involve dividing by the total outcome standard deviation, but standardized coefficients also involve multiplying by the predictor standard deviation. An effect size formula for Hedges' g that corrects for both issues is given by:

[E.16] $$g = \frac{\omega\beta}{\sqrt{1-\beta^2}}\sqrt{\frac{N(N-2)}{n_i n_c}}$$

This formula also applies to point-biserial correlations such as the Pearson's r correlation coefficient between a dichotomous treatment indicator and continuous outcome variable (where $\beta = r$).

Standard errors of group means. A study may report the standard error for group means but not the standard deviations separately by group. For unadjusted analyses, the standard errors for each group can be converted to the standard deviations as follows:

[E.17] $$SD_i = SE[y_i]\sqrt{n_i}\ and\ SD_c = SE[y_c]\sqrt{n_c}$$

For covariate-adjusted analyses, an additional $1 - R^2$ term is required:

[E.18] $$SD_i = SE[y_i]\sqrt{n_i(1-R^2)}\ and\ SD_c = SE[y_c]\sqrt{n_c(1-R^2)}$$

The study must report the R^2 value to make this conversion to standard deviations for covariate-adjusted analyses. The converted standard deviations for each group can then be used to compute the within-group pooled standard deviation.

Gain score standard deviations. A study may report only the means and standard deviations for the gain score (posttest minus pretest score). Gain score standard deviations are typically smaller than the unadjusted posttest standard deviations, inflating the effect size. If the study authors reported the baseline-outcome correlation ρ_{cor}, the WWC will correct for this issue as follows:

[E.19] $$SD_p = \frac{SD_{p,gain}}{\sqrt{2(1-\rho_{cor})}}$$

where $SD_{p,gain}$ is the pooled within-group standard deviation for the gain score. This conversion approximates the pooled standard deviation of the posttest scores. The effect size standard error for analyses based on gain scores are detailed in a later subsection about difference in differences.

The WWC will not apply this conversion if the study did not report the baseline-outcome correlation. The WWC also will not report effect sizes based on unconverted gain score standard deviations, given their lack of comparability with other WWC effect sizes. Hence, studies based on gain scores need to report either the unadjusted posttest standard deviations (preferable) or the gain score standard deviations and the baseline-outcome correlation (the information required for equation E.19).

Baseline standard deviations. The WWC will allow review team leadership to use baseline standard deviations in place of outcome standard deviations if (a) the outcome standard deviations are missing and (b) the baseline and outcome measure are the same (and use the same scoring procedures). In some cases, review team leadership may have a specific concern that the baseline standard deviations may differ substantially from the outcome standard deviations. One concern is floor effects, such as testing novel content yielding baseline performance close to the minimum. Another concern is ceiling effects, such as training students to mastery

yielding outcome performance close to the maximum. Although these concerns may be uncommon, review team leadership has the discretion not to use baseline standard deviations, even if all other standard deviations are not available (forcing the effect size to be missing). If the review team does not identify such a concern, the WWC will allow using baseline standard deviations to compute outcome effect sizes.

Single-group standard deviations. In some uncommon cases, a study might report standard deviations for only one group, such as for the comparison group but not intervention group. Although the WWC prefers computing effect sizes using the pooled standard deviation, it is important to note that standard statistical approaches routinely invoke the homogeneity of variance assumption—that is, assume the population standard deviations are the same across intervention and comparison groups. In this respect, the comparison standard deviation estimates the same population standard deviation as the pooled sample standard deviation; the only difference is yielding slightly less efficient effect size estimates. The WWC will therefore allow computing the effect size based on a single group's standard deviation, although using the pooled standard deviation is generally preferable.

Cluster-level assignment studies with continuous outcomes

All preceding statistical formulas focus on individual-level assignment studies. The statistical considerations are more complex for cluster-level assignment studies, such as assigning schools to conditions and measuring student outcomes. Compared with individual-level assignment, cluster-level assignment generally yields more uncertainty (larger standard errors) for intervention effect estimates.

The following section describes the WWC's approach to handling such issues, including accounting for variation in how study authors may conduct and report their analyses. For instance, study authors could analyze individual-level data and account for clustering, analyze individual-level data and not account for clustering, or analyze cluster-level means. This section focuses on continuous outcomes; a later section presents extensions for dichotomous outcomes.

Defining the effect size of interest for cluster-level assignment studies

The WWC defines the effect size of interest for cluster-level assignment studies using the total variability, both between and within clusters, as the standardizer (see equation 3 in Hedges, 2007). Using the between- and within-cluster variability as the standardizer yields effect sizes comparable with individual-level assignment studies conducted in several sites. This point is important because most of the individual-level assignment studies that the WWC reviews are conducted in more than one school. In contrast, using only the between-cluster variability as the standardizer would yield larger effect sizes than those typically found in individual-level assignment studies.

A common way to compute effect sizes from cluster-level assignment studies is using an estimate of the unstandardized mean difference b and the pooled individual-level standard deviation SD_p:

[E.20] $$g = \frac{\omega b}{SD_p}\sqrt{\gamma} = \frac{\omega b}{SD_p}\sqrt{1 - \frac{2\left(\frac{N}{M} - 1\right)\rho_{ICC}}{N - 2}}$$

where N is the total number of individuals, M is the total number of clusters, and ρ_{ICC} is the intraclass correlation that represents the degree of clustering. This formula is appropriate for both unadjusted and covariate-adjusted estimates of the mean difference b. The formula is similar to the general Hedges' g formula for individual-level assignment (see equation E.1) but adds another bias correction term $\sqrt{\gamma}$ (see equation 15 in Hedges, 2007). The added term can be viewed as a small number of clusters adjustment. Although typically small, this adjustment is needed to yield approximately unbiased estimates of the population effect size, especially for studies with few clusters.

The small number of clusters correction $\sqrt{\gamma}$ is in addition to the small sample size adjustment ω, provided earlier in equation E.3. The formula for the degrees of freedom df is different, however, for effect sizes based on individual-level standard deviations in cluster-level assignment studies:

[E.21] $$df = \frac{\left[(N-2) - 2\left(\frac{N}{M} - 1\right)\rho_{ICC}\right]^2}{(N-2)(1-\rho_{ICC})^2 + \frac{N}{M}\left(N - 2\frac{N}{M}\right)\rho_{ICC}^2 + 2\left(N - 2\frac{N}{M}\right)\rho_{ICC}(1-\rho_{ICC})}$$

This formula for df can then be substituted into equation E.3 to determine ω.

The intraclass correlation and application to an example

The previous subsection introduced a key parameter: the intraclass correlation ρ_{ICC}. This parameter can range from 0 to 1 and represents the degree of statistical clustering of individuals within clusters. More technically, it is the proportion of the total variance attributable to the between-group variance. The WWC will use the study author-reported intraclass correlation when available, but this value is often not reported. Based on empirical literature in the field of education, the WWC has adopted the default intraclass correlation values of .20 for achievement outcomes and .10 for all other outcomes (Hedges & Hedberg, 2007; Schochet, 2008). The topic area team leadership may set different defaults in the review protocol with justification.

Consider if the example study in table E.1 on reading achievement was a cluster-level assignment study with M = 8 classrooms randomly assigned to conditions. Assume that the study did not report the intraclass correlation, meaning that the WWC would use the default intraclass correlation of .20 for achievement outcomes. The preliminary calculations for calculating the effect size can be written as follows:

$$\sqrt{\gamma} = \sqrt{1 - \frac{2\left(\frac{80}{8} - 1\right).20}{80-2}} = 0.98$$

$$df = \frac{\left[(80-2) - 2\left(\frac{80}{8} - 1\right).20\right]^2}{(80-2)(1-.20)^2 + \frac{80}{8}\left(80 - 2\frac{80}{8}\right).20^2 + 2\left(80 - 2\frac{80}{8}\right).20(1-.20)} = 59.44$$

$$\omega = 1 - \frac{3}{4 \times 59.44 - 1} = 0.99$$

The pooled individual-level standard deviation remains the same as calculated before ($SD_p = 11.94$), and the unadjusted mean difference based on the table E.1 statistics is $b = 4.78$. The unadjusted Hedges' g effect size for this example cluster-level assignment study therefore would be:

$$g = \frac{0.99 \times 4.78}{11.94} 0.98 = 0.39$$

Note that this effect size of $g = 0.39$ is the nearly same as the value $g = 0.40$ calculated earlier based on assuming individual-level assignment. Similarly, if $b = 3.59$ was used instead as a covariate-adjusted mean difference, then the effect size based on equation E.20 would be $g = 0.29$, which is very close to the analogous value of $g = 0.30$ computed earlier for individual-level assignment. However, the following subsection details that the standard errors for these effect sizes are much larger for cluster-level assignment than individual-level assignment.

Standard errors for unadjusted analyses for cluster-level assignment studies

Assigning clusters of individuals to intervention and comparison groups reduces a study's effective sample size. Intuitively, the number of randomized units is smaller for cluster-level assignment than individual-level assignment. This issue increases standard errors and could increase false positive rates if the study authors do not account for clustering in their analyses.

The increase in effect size variance depends on the intraclass correlation. The variance remains the same if the intraclass correlation is 0, reflecting no statistical clustering. The variance increases more for larger intraclass correlation values. The design effect η approximately represents the multiplicative increase in the effect variance relative to individual-level assignment:

[E.22] $$\eta = 1 + \left(\frac{N}{M} - 1\right)\rho_{ICC}$$

For the example study in the previous subsection, this parameter is given by:

$$\eta = 1 + \left(\frac{80}{8} - 1\right).20 = 2.80$$

This value means that, for the same number of students, the effect size variance will be approximately 2.8 times as large for cluster-level assignment than individual-level assignment. The standard error will be approximately $\sqrt{2.80} = 1.67$ times as large because the standard error is the square root of the variance.

For unadjusted analyses, the standard error for the effect size given in equation E.20 is given by:

[E.23] $$SE[g] = \omega\sqrt{\frac{n_i + n_c}{n_i n_c}\eta + \frac{g^2}{2df}}$$

The degrees of freedom df for this formula is given by equation E.21. This formula is appropriate regardless of whether the study authors' analysis accounted for clustering. In other words, the calculated effect size standard error will account for clustering, even if the study authors' analysis did not. This formula assumes analyses

without baseline covariate adjustment. This formula also assumes that the number of individuals is the same for each cluster, but it still will be approximately correct for studies with unequal cluster sizes[36] (Hedges, 2007).

For the previous example study, the standard error for the unadjusted analysis would be:

$$SE[g] = 0.99\sqrt{\frac{39+41}{39\times 41}2.8 + \frac{0.39^2}{2\times 59.44}} = 0.37$$

This standard error of 0.37 is much larger than the analogous value of 0.22 computed earlier for individual-level assignment, reflecting increased uncertainty about the intervention effect estimate.

Standard errors for covariate-adjusted analyses for cluster-level assignment studies

The WWC will compute covariate-adjusted standard errors for cluster-level assignment studies, extending the approaches presented previously based on regression coefficient standard errors (equation E.9) and R^2 values (equation E.11). One complication is whether the study authors' analysis model appropriately accounted for clustering, which can substantially change the author-reported standard errors. The following formulas account for these different analytic choices.

Study authors can account for clustering in multiple acceptable ways. For instance, they could include random effects at the unit of assignment level in a hierarchical linear model, which are sometimes also called mixed-effects or multilevel models (Raudenbush & Bryk, 2002). Study authors also could conduct a single-level ordinary least-squares regression model and then adjust the standard errors at the unit of assignment level using cluster-robust standard errors (McNeish et al., 2017). Either approach can yield regression coefficient standard errors that account for clustering.

If the study authors reported a cluster-corrected standard error $SE_{CC}[b]$ for the unstandardized mean difference, the WWC will compute the effect size standard error as follows:

[E.24] $$SE[g] = \omega\sqrt{\left(\frac{SE_{CC}[b]}{SD_p}\right)^2\gamma + \frac{g^2}{2df}}$$

where γ is the small number of clusters correction introduced earlier in equation E.20.

Study authors will often not account for clustering in cluster-level assignment studies, even though they should. For instance, a study might conduct an ordinary least-squares regression model without adjusting the standard errors for clustering in a cluster-level assignment study. Standard errors from such models cannot be substituted into equation E.24 because the standard error would be too small. However, the WWC can compute the effect size standard error as follows if the study authors reported an uncorrected standard error $SE_{UC}[b]$ for the unstandardized mean difference:

[36] These studies often do not report sufficient information to apply more complicated formulas that explicitly address unequal cluster sizes (Hedges, 2007). Hence, the WWC will apply equation E.23 and other formulas in this section to studies with equal and unequal cluster sizes.

[E.25] $$SE[g] = \omega \sqrt{\left(\frac{SE_{UC}[b]}{SD_p}\right)^2 \eta + \frac{g^2}{2df}}$$

where η is the design effect introduced in equation E.22. The design effect accounts for the author-reported standard error being too small because the design effect is often considerably larger than 1.

A study might also report the multiple correlation R^2 from a single-level model (such as ANCOVA or ordinary least-squares regression) and not report the regression coefficient standard error. In this case, the WWC can compute the covariate-adjusted standard error as follows:

[E.26] $$SE[g] = \omega \sqrt{\frac{n_i + n_c}{n_i n_c}(1 - R^2)\eta + \frac{g^2}{2df}}$$

This standard error formula is appropriate regardless of whether the study authors' model accounted for clustering (because this formula does not rely on the author-reported standard error). Following similar guidelines for individual-level assignment, the WWC will substitute 0 percent for R^2 in this formula if both the regression coefficient and R^2 are not reported, effectively making the formula the same as the standard error for unadjusted analyses (equation E.23). The WWC also will follow the same relative prioritization of standard error formulas noted in table E.2 (favoring the regression coefficient standard error approach over the R^2 approach).

Consider if the example study introduced earlier reported a regression coefficient b = 3.59 for the covariate-adjusted unstandardized mean difference and a coefficient standard error $SE_{CC}[b]$ = 3.54 that accounted for clustering. The covariate-adjusted effect size would be g = 0.29 based on equation E.20. The covariate-adjusted standard error for the effect size would be calculated as follows:

$$SE[g] = 0.99 \sqrt{\left(\frac{3.54}{11.94}\right)^2 \left(1 - \frac{2\left(\frac{80}{8} - 1\right).20}{80 - 2}\right) + \frac{0.29^2}{2 \times 59.44}} = 0.29$$

This covariate-adjusted effect size standard error of 0.29 is smaller than the unadjusted standard error of 0.37 for the cluster-level assignment design without covariates, indicating increased precision. However, it is still larger than the standard error for both the unadjusted and covariate-adjusted individual-level assignment equivalent (0.22 and 0.17, respectively), reflecting the important role of cluster-level assignment in determining effect size precision.

Alternatively, the study could have reported an uncorrected standard error $SE_{UC}[b]$ = 2.07, which would have yielded the same 0.29 effect size standard error using equation E.25 (and η = 2.80 as calculated earlier). Reporting R^2 = 40 percent and using equation E.26 would have yielded the same effect size standard error.

Analyses of cluster-level data

The previous subsections assume that the study authors analyzed individual-level data or otherwise had access to individual-level standard deviations. However, these data may often be unavailable. Many analyses of

educational data use publicly available school-level means and may not have access to individual-level data. This limitation means the study authors would have access to only cluster-level standard deviations, which are usually much smaller than individual-level standard deviations.

Cluster-level and individual-level standard deviations are therefore not interchangeable. Effect sizes will be much larger for those based on cluster-level, rather than individual-level, standard deviations. The WWC will correct for this issue by using the intraclass correlation to adjust cluster-level effect sizes and yield an individual-level effect size equivalent (see equations 25 and 26 in Hedges, 2007). This individual-level Hedges' *g* equivalent can be written as follows:

[E.27] $$g = \frac{\omega b}{SD_B}\sqrt{\frac{M\eta}{N}}$$

where SD_B is the pooled cluster-level standard deviation (based on applying equation E.2 to the cluster-level standard deviations and number of clusters within each group). The degrees of freedom *df* to compute ω for this formula is:

[E.28] $$df = M - 2$$

This formula for the degrees of freedom contrasts with the formula given in equation E.21. The difference arises from the pooled cluster-level standard deviation being distributed with fewer degrees of freedom than the pooled individual-level standard deviation.

For unadjusted analyses, the standard error for the individual-level effect size is given by:

[E.29] $$SE[g] = \omega\sqrt{\left(\frac{n_i + n_c}{n_i n_c}\right)\eta + \frac{g^2}{2df}}$$

Note that this standard error formula is nearly identical to the one for effect sizes based on individual-level standard deviations (equation E.23). The only difference is that the degrees of freedom for the right-hand term is $df = M - 2$ rather than equation E.21. This difference tends not to matter much because the right-hand term is generally a small contributor to the overall effect size variance. In other words, cluster-level effect sizes converted into the individual-level equivalents have nearly the same variability as effect sizes computed based on individual-level standard deviations.

For covariate-adjusted analyses, the effect size standard error can be computed as follows if the standard error of the unstandardized mean difference $SE[b]$ is reported:

[E.30] $$SE[g] = \omega\sqrt{\left(\frac{SE[b]}{SD_B}\right)^2\frac{M\eta}{N} + \frac{g^2}{2df}}$$

If the study instead reported the multiple correlation R^2, then the effect size standard error can be computed using equation E.26 (but substituting $df = M - 2$ for the degrees of freedom).

The computed effect size will have some sensitivity to the chosen intraclass correlation value, but the determination of statistical significance will be robust to this value. Both the effect size and standard error depend on the intraclass correlation by the exact same multiplicative amount. This dependence exactly cancels out when computing the *t* test ratio—that is, the effect size divided by its standard error—used to determine the WWC *p* value. Hence, the WWC-calculated *p* value for effect sizes based on cluster-level standard deviations does not depend on the intraclass correlation value.

This conversion provides a route for studies without access to individual-level data to contribute to WWC syntheses. Otherwise, such studies would be excluded when computing meta-analytic effect sizes. One important consideration is that the computed effect size may either be systematically too large or too small depending on how the estimated or imputed intraclass correlation compares to the true population intraclass correlation. If the computed effect size is too small, then the meta-analytic effect size will be conservative. If the computed effect size is too large, the bias in the meta-analytic estimate will be partially offset by the study having standard errors that are also too large, reducing the meta-analytic weight the study receives.

Consider if the study authors reported the following statistics for reading achievement in table E.3.

Table E.3. *Descriptive statistics for a cluster-level analysis*

Group	Cluster-level standard deviation	Number of clusters	Number of individuals
Intervention	3.42	12	121
Comparison	3.89	10	99

The effect size for a mean difference of $b = 2.56$ would be computed using equation E.27 as follows:

$$SD_B = \sqrt{\frac{(12-1)3.42^2 + (10-1)3.89^2}{12+10-2}} = 3.64$$

$$df = 12 + 10 - 2 = 20$$

$$\omega = 1 - \frac{3}{4 \times 20 - 1} = 0.96$$

$$\eta = 1 + \left(\frac{220}{22} - 1\right).20 = 2.80$$

$$g = \frac{0.96 \times 2.56}{3.64}\sqrt{\frac{22 \times 2.80}{220}} = 0.36$$

For unadjusted analyses, the effect size standard error would be computed using equation E.29:

$$SE[g] = 0.96\sqrt{\left(\frac{121+99}{121\times 99}\right)2.80 + \frac{0.36^2}{2\times 20}} = 0.22$$

For covariate-adjusted analyses, consider if the study authors reported the unstandardized standard error as 1.05. The effect size standard error would be computed using equation E.30:

$$SE[g] = 0.96\sqrt{\left(\frac{1.05}{3.64}\right)^2\frac{22\times 2.80}{220} + \frac{0.36^2}{2\times 20}} = 0.16$$

Other cluster-level extensions for continuous outcomes

Alternative formulas for computing the effect size or pooled within-group standard deviation can extend to cluster-level assignment studies. For instance, the previous formulas for using the total standard deviation (equation E.14) or gain score standard deviation (equation E.19) apply equally to individual-level and cluster-level assignment studies.

Formulas based on test statistics or standard errors, however, have additional considerations when applied to cluster-level assignment studies. For instance, consider the approach presented previously to compute the effect size using a *t* statistic for unadjusted mean comparisons (equation E.7). When applied to cluster-level assignment, this approach depends on whether the study authors reported a cluster-corrected *t* statistic t_{CC} or an uncorrected statistic t_{UC} that does not account for clustering. The Hedges' *g* effect size can be computed as follows in these two different scenarios:

[E.31] $$g = \omega t_{CC}\sqrt{\eta\frac{n_i+n_c}{n_i n_c}}$$

[E.32] $$g = \omega t_{UC}\sqrt{\gamma\frac{n_i+n_c}{n_i n_c}}$$

Similar extensions apply to covariate-adjusted test statistics (same as above, but add a $\sqrt{1-R^2}$ term) or *F* statistics (replace *t* with $\sqrt{F}$). Test statistics based on cluster-level data should be treated as cluster-corrected statistics.

The approach to compute standard deviations based on author-reported group mean standard errors (equations E.17 and E.18) also depends on whether the study authors' analysis accounted for clustering. For group mean standard errors without covariate adjustment, the group standard deviations can be computed as follows:

[E.33] $$SD_i = SE_{CC}[y_i]\sqrt{\frac{\gamma n_i}{\eta}} \text{ and } SD_c = SE_{CC}[y_c]\sqrt{\frac{\gamma n_c}{\eta}}$$

[E.34] $$SD_i = SE_{UC}[y_i]\sqrt{n_i} \text{ and } SD_c = SE_{UC}[y_c]\sqrt{n_c}$$

Similar extensions apply to covariate-adjusted group mean standard errors (same as above but adds a $\sqrt{1-R^2}$ term).

Dichotomous outcomes

Educational research studies sometimes use dichotomous outcomes such as dropping out versus staying in school, grade promotion versus retention, and passing versus failing a test. Whenever possible, the WWC will present results for dichotomous outcomes using percentages. For instance, a study might find college enrollment rates of 50 percent versus 40 percent in the intervention versus comparison group, respectively. This approach is suitable for presenting individual study findings for a single outcome measure.

The WWC, however, also must synthesize findings across studies (or even for multiple outcomes within an outcome domain in a single study). These syntheses sometimes need to combine findings for both continuous and dichotomous outcomes. The WWC therefore requires a common effect size metric for dichotomous outcomes that is comparable with Hedges' *g* for continuous outcomes. For this reason, the WWC has adopted the Cox index as the default effect size for dichotomous outcomes. This metric aims to yield effect sizes comparable with Hedges' *g* for continuous outcomes (Sanchez-Meca et al., 2003).

An example based on proficiency rates can help illustrate the intuition behind the Cox index. A study could report the means and standard deviations for a continuously scaled mathematics achievement measure (assumed to be normally distributed). These statistics permit computation of the Hedges' *g* effect size. This study, however, could instead dichotomize the achievement measure and present only the percentage proficient in each group, omitting the continuous means. The Cox index aims to recreate the Hedges' *g* effect size (based on the continuous measure) using only the percentages for the dichotomous measure[37] (see Sanchez-Meca et al., 2003, for simulations demonstrating this point).

Computing the Cox index for dichotomous outcomes

Computing the Cox index effect size starts with first understanding the concept of *odds*, which compares the probability of an event occurring with the probability of it not occurring. The odds for a rate of 75 percent would be .75 / .25 = 3, indicating that the event is three times as likely to occur than not occur (3:1 odds).

The odds ratio *OR* is the ratio of the odds for two groups being compared:

[E.35] $$OR = \frac{Odds_i}{Odds_c} = \frac{p_i/(1-p_i)}{p_c/(1-p_c)}$$

[37] For truly dichotomous outcomes such as college enrollment, one could imagine an underlying normally distributed continuous variable that determines the dichotomous outcome based on a certain threshold. The Cox index aims to estimate the Hedges' *g* for that underlying latent continuous variable. Although less intuitive than percentages, this approach places effect sizes for dichotomous and continuous measures on a similar scale, enabling synthesis across findings and outcome measures.

where p_i is the intervention probability and p_c is the comparison probability (and $Odds_i$ and $Odds_c$ are the odds in each group). Consider a study that had 50 percent proficiency rate in the intervention group (1:1 odds) and 40 percent in the comparison group (1:1.5 odds). The odds ratio would be:

$$OR = \frac{.50/(1-.50)}{.40/(1-.40)} = 1.5$$

The Cox index is computed using the natural logarithm of the odds ratio (also called the log odds ratio) as follows:

[E.36] $$g = \frac{ln(OR)}{1.65} = \frac{ln(Odds_i) - ln(Odds_c)}{1.65}$$

The symbol *g* also is used here for the Cox index to note its comparability with the Hedges' *g* effect size. The right-hand part of this formula writes the Cox index as a mean difference (difference in log odds) divided by a standardizing term (see Cox, 1970, p. 21), highlighting its similarity to the functional form for the standardized mean difference for continuous outcomes. The WWC will apply standard continuity corrections to findings that have group means of 0 percent or 100 percent.[38]

For unadjusted analyses in individual-level assignment studies, the standard error for the Cox index is given by the following:

[E.37] $$SE[g] = \frac{1}{1.65}\sqrt{\frac{1}{p_i n_i} + \frac{1}{(1-p_i)n_i} + \frac{1}{p_c n_c} + \frac{1}{(1-p_c)n_c}}$$

Consider a difference of 50 percent versus 40 percent for an intervention group with 30 students and comparison group with 25 students, respectively. As noted previously, the odds ratio would be $OR = 1.5$. The Cox index would therefore be:

$$g = \frac{ln(1.5)}{1.65} = 0.25$$

The Cox index standard error would be:

$$SE[g] = \frac{1}{1.65}\sqrt{\frac{1}{.50 \times 30} + \frac{1}{(1-.50) \times 30} + \frac{1}{.40 \times 25} + \frac{1}{(1-.40) \times 25}} = 0.33$$

[38] Continuity corrections address the problem of infinite log odds ratios for means of 0 percent or 100 percent—that is, at least one zero-count cell in a 2×2 table of group crossed by outcome status. Consistent with standard meta-analytic procedures, the WWC will add 0.5 to each cell of a 2×2 table that requires this correction (Weber et al., 2020). Proportions based on these corrected counts will be used to compute effect sizes and standard errors, but the WWC will report the original uncorrected proportions on the WWC website. Some methodologists have raised concerns about applying continuity corrections to outcomes that routinely yield 0 percent or 100 percent means (such as cancer diagnosis rates in the general population; Efthimiou, 2018), but these types of rare (or ubiquitous) outcomes should be uncommon in educational research.

Dichotomous outcomes can sometimes yield large Cox index effect sizes, especially for base rates close to 0 percent or 100 percent. For instance, the difference between a 99.8 percent versus 99.5 percent rate corresponds to a Cox index of 0.56. The WWC will still use the Cox index transformation for statistical analysis in these cases, but practitioners may find the raw percentages to be more informative and interpretable than a converted effect size. This point underscores why the WWC will present results for dichotomous outcomes as percentages, whenever possible.

Covariate-adjusted analyses for dichotomous outcomes

Like for continuous outcomes, the WWC will not report effect sizes based on unadjusted dichotomous means if the study requires baseline adjustment such as for QEDs and high-attrition RCTs. Effect sizes for these studies must instead incorporate covariate adjustment. The WWC will allow three categories of adjustment approaches for dichotomous outcomes: (a) average predicted probabilities, (b) linear probability models, and (c) logistic regression coefficients.

Average predicted probabilities. Study authors can compute average predicted probabilities from logistic (or probit) regression models that adjust for baseline differences (Gelman & Pardoe, 2007). This approach is analogous to using adjusted means from ANCOVA models for continuous outcomes. The adjusted probabilities are based on using the regression model and covariate values to predict the average outcome if the entire sample received the intervention (adjusted intervention mean) or did not receive the intervention (adjusted comparison mean). The WWC can then use these adjusted probabilities to compute the covariate-adjusted Cox index using equation E.36.

As a conservative assumption, the WWC will compute the Cox index standard error assuming unadjusted analyses—that is, using equation E.37—due to lack of guidance in the field on how to compute standard errors for covariate-adjusted Cox indices.

Linear probability models. Linear probability models use standard regression models, such as ordinary least squares, that assume a linear relationship between the predictors and the probability of a dichotomous outcome. Typically, the adjusted intervention mean is computed by adding the regression coefficient for the intervention effect (denoting the adjusted percentage-point difference) to the observed comparison group mean.[39] For instance, the adjusted intervention mean would be 45 percent if the comparison mean was 40 percent and the regression coefficient for the impact estimate was .05 (indicating a covariate-adjusted mean difference of 5 percentage points). The WWC can then use the two mean values (45 percent and 40 percent) to compute the covariate-adjusted Cox index using equation E.36.

Study authors considering this approach should be aware of its functional form assumptions, especially when applied to QEDs and high-attrition RCTs. Linear probability models can introduce bias for these research designs if the assumed linear relationship between continuous baseline covariates and average probabilities poorly

[39] One issue is that linear probability models can sometimes yield out-of-bounds predictions (outside of the range 0 percent to 100 percent). In general, this issue is not a major concern for the WWC because making predictions for individual students is not relevant to the WWC's context. However, the adjusted intervention mean could also be out of bounds. In such cases, the WWC will truncate any model-adjusted means to their nearest boundary (0 percent or 100 percent).

approximates the true data generating mechanisms (Deke, 2014). However, this consideration also remains for logistic regression models, which also can yield bias due to functional form misspecification or omitted variables. The maximum WWC research rating of *Meets WWC Standards With Reservations* reflects these considerations for studies that require baseline adjustment. In contrast, functional form assumptions are less of a concern for studies such as RCTs with a low risk of bias due to compositional change that are eligible for a rating of *Meets WWC Standards Without Reservations* (Deke, 2014; Gomilla, 2021; Hellevik, 2007). These types of studies can yield unbiased estimates of intervention effects even with unadjusted statistics.

Like average predicted probabilities from logistic regression models, the WWC will compute the standard error for the Cox index assuming unadjusted analyses (using equation E.37).

Logistic regression coefficients. If the study did not report covariate-adjusted means, the WWC will allow computing the covariate-adjusted Cox index based on logistic regression coefficients. This approach requires that the study also reports the standard error for the logistic regression coefficient.

If reported on a log odds ratio scale, logistic regression coefficients can be directly substituted as $ln(OR)$ in equation E.36 to compute the Cox index—that is, divide the coefficient by 1.65. For instance, a log odds ratio coefficient of 0.78 would yield a Cox index of 0.78 / 1.65 = 0.47. The Cox index standard error would then be the logistic regression coefficient standard error divided by 1.65. Additional conversion is needed if the study authors report results on an odd ratios scale instead. An odds ratio of 0.78 (notice the lack of "log" in "odds ratio") instead corresponds to *ln*(0.78) / 1.65 = -0.15.

For this approach, the study authors must provide the standard error for the logistic regression coefficient. The WWC will author query for this standard error if it is not reported in the study article or request the average predicted probabilities from the logistic regression model instead. This requirement is important because, unlike ordinary least-squares regression, controlling for covariates will always increase the standard error for the logistic regression coefficient; see the mathematical proof by Robinson and Jewell (1991). Hence, assuming an unadjusted standard error for the logistic regression coefficient will be anticonservative—that is, inflate false positive rates. Though counterintuitive, this point reflects that logistic regression coefficients estimate a slightly different type of log odds ratio than those based on covariate-adjusted means, as detailed next.

Prioritization of covariate-adjusted approaches. A study might report multiple covariate-adjustment approaches for dichotomous outcomes. The WWC will prioritize these approaches in the following order:

1. Average predicted probabilities
2. Linear probability models
3. Logistic regression coefficients

The WWC favors Cox indices based on covariate-adjusted means—that is, average predicted probabilities or those from a linear probability model. Odds ratios based on these means are sometimes called marginal odds ratios and are conceptually appropriate effect sizes for characterizing effects on heterogenous groups of students (Daniel et al., 2020). This point aligns well with the WWC's focus on practitioner audiences who aim to improve outcomes for diverse groups of students, such as students with varying likelihoods of college enrollment.

In contrast, odds ratios based on logistic regression coefficients are sometimes called conditional odds ratios and are conceptually appropriate for characterizing effects on an individual student (Daniel et al., 2020). The conditional log odds ratio is always larger in magnitude than, or equal to, the marginal log odds ratio computed based on average predicted probabilities from the same logistic regression model (Norton & Dowd, 2018). Hence, both conditional and marginal odds ratios can incorporate covariate adjustment, but they estimate different quantities. This distinction may not matter much in practice, but it is the reason why the WWC prefers Cox indices based on covariate-adjusted means over those based on logistic regression coefficients.

Cluster-level assignment studies with dichotomous outcomes

The WWC will use the same Cox index formula (equation E.36) to compute the effect size for dichotomous outcomes in cluster-level assignment studies. The WWC will use this formula for analyses of individual-level and cluster-level data (the only difference is in the weighting of clusters with unequal sizes). Special formulas are not required for cluster-level assignment studies because, unlike continuous outcomes, the effect size for dichotomous outcomes does not rely on standard deviations.

The effect size standard error, however, must reflect the increased variance of intervention effect estimates in cluster-level assignment studies. This standard error is given as follows:

[E.38] $$SE[g] = \frac{\sqrt{\eta}}{1.65}\sqrt{\frac{1}{p_i n_i} + \frac{1}{(1-p_i)n_i} + \frac{1}{p_c n_c} + \frac{1}{(1-p_c)n_c}}$$

where η is the design effect term introduced earlier (equation E.22). Study authors looking for guidance on computing intraclass correlation values for dichotomous outcomes can consult Goldstein et al. (2002).

Design-comparable effect sizes from single-case designs

For single-case design (SCD) studies rated *Meets WWC Standards With Reservations* or *Meets WWC Standards Without Reservations*, the WWC will calculate a design-comparable effect size where feasible and appropriate in the judgment of review team leadership. Effect size estimation should be performed at the end of the review process, when data from all findings within a given domain are available to allow for an assessment of the appropriate functional form of the effect size model, as discussed later in this section. The design-comparable effect size is comparable, in principle, to a standardized mean difference effect size (Hedges' g) from a group design study, if that group design study could be conducted with the same population of participants, intervention procedures, and outcome assessment procedures. One major advantage of the design-comparable effect size is it can used in meta-analysis alongside the Hedges' *g* from group-design studies.

A design-comparable effect size can be computed for a study that has three or more units (usually individuals, but sometimes also classrooms or other clusters) in a design that is multiple baseline across individuals, multiple probe across individuals, or a treatment reversal design that is replicated across three or more individual units. SCDs involve multiple observations in intervention and comparison conditions for each unit. Despite the name, SCDs typically involve data from several units.

Computing the design-comparable effect size requires access to raw outcome data by unit, by observation occasion, and by intervention phase. If the study manuscript or related documentation does not provide the raw data in tabular form, the WWC prefers contacting the study authors for the raw data. If the study authors do not provide the raw data, then WWC reviewers may use graph-digitizing software to extract the individual data points from a graph.

The WWC uses the multilevel modeling framework described in Pustejovsky et al. (2014) to estimate parameters for computing the design-comparable effect size. This approach uses restricted maximum likelihood estimation and is more flexible than earlier approaches based on method of moments estimators (Hedges et al., 2012, 2013).

The simplest multilevel model for estimating the design-comparable effect size can be written as follows (corresponding to model MB1 in Pustejovsky et al., 2014):

[E.39] $$y_{ij} = a + bT_{ij} + u_i + e_{ij}$$

where y_{ij} is the observation of unit i at time j, a is the average outcome in the absence of the intervention, b is the intervention effect assumed to be constant across units, and T_{ij} is a dummy indicator for receiving the intervention. The level-1 error term e_{ij} is assumed have a mean of zero, a variance of σ^2, and a first-order autocorrelation of ϕ. The level-2 error term, u_i, is the deviation from the average level for unit i, which is assumed to be normally distributed with a mean of zero and a variance of τ^2. This model assumes there are no time trends at baseline or any later phases; though this assumption may not always be appropriate, this parsimonious model can provide one starting point for computing the effect size.

The design-comparable effect size g for this model is (derived from equation 14 in Pustejovsky et al., 2014):

[E.40] $$g = \omega \frac{b}{\sqrt{\hat{\tau}^2 + \hat{\sigma}^2}}$$

where $\hat{\tau}^2$ and $\hat{\sigma}^2$ are variance estimates from the model in equation E.39. This effect size formula is similar to equation E.2, except that the standardizing denominator term is different. For SCDs, the variances between and within units are separate estimates, which are then combined in standardizing the intervention effect. For group designs, only the total variability is estimated based on a single point in time (the pooled outcome standard deviation implicitly includes variability within participants because the raw variability across participants includes measurement error). Though the estimation approach differs, the conceptual effect size quantity is similar across SCDs and group designs.

The small-sample bias correction term[40] ω is defined in equation E.3, where the degrees of freedom df is (derived equation 13 in Pustejovsky et al., 2014):

[E.41] $$df = \frac{2(\hat{\tau}^2 + \hat{\sigma}^2)^2}{Var[\hat{\tau}^2 + \hat{\sigma}^2]}$$

[40] Pustejovsky et al. (2014) used $J(v)$ to refer to the small-sample bias correction term and $\boldsymbol{\omega}$ to refer to a vector of variance components. The WWC instead uses the term ω to refer to the sample-sample bias correction term.

where $Var[\hat{\tau}^2 + \hat{\sigma}^2]$ is the variance of the sum of the two estimated variance components. In practice, the degrees of freedom can be small, making the bias correction quite consequential.

The approximate standard error $SE[g]$ for the design-comparable effect size is given by (derived from equation 15 from Pustejovsky et al., 2014):

$$SE[g] = \omega \sqrt{\frac{SE[b]^2}{\hat{\tau}^2 + \hat{\sigma}^2}\left(\frac{df}{df - 2}\right) + g^2\left(\frac{df}{df - 2} - \frac{1}{\omega^2}\right)} \quad \text{[E.42]}$$

where $SE[b]$ is the standard error of the unstandardized intervention effect estimate.

The parsimonious model in equation E.39 assumes that (a) the intervention effect is constant across units and (b) there is "no trend" at baseline or any later phases. Review team leadership should consider the appropriateness of these assumptions by reviewing the pattern of responding in consultation with a visual analysis expert or other appropriate algorithms used by SCD researchers. Existing empirical research also may provide guidance on modeling time trends for the types of outcomes that are under review.

If the underlying data do not conform to the most parsimonious model specifications, the review team may consult with content and methodological experts to alter the underlying model to fit the data more appropriately (for details, see Pustejovsky et al., 2014). If there is ambiguity about the appropriate model, the most parsimonious model should be preferred. The review team also may elect to not compute the design-comparable effect size if an appropriate method is not available or if an appropriate model specification cannot be identified. The WWC will document the rationale for the chosen approach to compute the design-comparable effect size (or the reasons for not computing it).

In the case of SCDs of cluster-level data, the transformation factor used in E.27 and E.30, $\sqrt{\frac{M\eta}{N}}$, may be applied to E.40 and E.42 in the same way to estimate the D-CES and its standard error, which approximates effects using individual-level data.

Software products to implement these estimation approaches include the scdhlm R package or a web application available at https://jepusto.shinyapps.io/scdhlm/ (Pustejovsky et al., 2021).

Regression discontinuity designs

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

For continuous outcomes, the numerator of the effect size is the difference between the predicted group means, with each mean estimated using data from the corresponding side of the cutoff. The standard deviations and sample sizes used to estimate the effect size should be the standard deviations and sample sizes of the treatment and comparison groups from the full sample (as opposed to just those units within an optimal bandwidth that

weights observations relative to their distance from the cutoff). If it might be possible to compose more than one treatment and comparison group (such as with a fuzzy regression discontinuity design), then the treatment and comparison groups should be formed based on treatment assignment.

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

Special considerations apply to calculating standard errors from regression discontinuity designs. The strong correlation between the forcing variable and intervention assignment increases the standard error of intervention effect estimate. For this reason, regression discontinuity designs must report the model-based standard error of the unstandardized intervention effect that accounts for this correlation. The WWC will then use equation E.9 (for individual-level assignment) or equation E.24 (for cluster-level assignment) to compute the standard error of the Hedges' *g* effect size. If the study authors do not report the model-based standard error or *t* statistic, then the WWC will not compute the effect size standard error; the finding will not contribute to cross-study meta-analytic averages in this case.

Study authors must account for clustering in cluster-level assignment regression discontinuity designs; otherwise, the WWC will not use the model-based standard errors. Appropriate methods to address clustering include boot-strapping, multilevel linear modeling, or the method proposed by Lee and Card (2008).

Computing *p* values and statistical significance

The WWC presents study findings by focusing on domain-level average effect sizes that summarize findings separately by outcome domain. A study could have multiple main findings within an outcome domain, such as having two distinct measures of reading comprehension. Appendix F details how the WWC computes these domain-level averages and determines their statistical significance.

In addition, the WWC will present the statistical significance for each finding in a study, including for both main and supplemental findings. The finding-level statistical significance serves as additional contextual information, but it does not affect the WWC's qualitative conclusions. For instance, the WWC's characterization of study findings depends on the domain-level, but not finding-level, statistical significance.

The WWC generally accepts the author-reported *p* values and statistical significance for study findings. However, the WWC will compute the statistical significance in three common cases:

1. The study does not include statistical significance estimates; or
2. The study calculations have a known problem such as not applying a required adjustment for baseline differences; or
3. The study did not account for clustering in an individual-level analysis for a cluster-level assignment study.

An example study could report unadjusted analyses but require adjustment for baseline differences. The WWC could apply a difference-in-differences adjustment in this case, but the author-reported statistical significance would be unacceptable to use for official WWC purposes.

In these cases, the WWC determines the statistical significance of each study finding by first computing the t test ratio between the effect size g and its standard error $SE[g]$:

[E.43] $$t = \frac{g}{SE[g]}$$

The degrees of freedom for determining the p value for this t statistic is based on the formulas presented earlier for individual-level assignment (equation E.4) and cluster-level assignment (equation E.21; however, note that equation E.28 applies instead for effect sizes based on cluster-level standard deviations).

As an example, an earlier section calculated an effect size $g = 0.40$ and standard error $SE[g] = 0.22$ for individual-level assignment study with a total sample size of 80 students (corresponding to 78 degrees of freedom). The t statistic for computing the WWC-calculated p value would therefore be $t = \frac{0.40}{0.22} = 1.82$. The p value could be calculated in Microsoft Excel using *T.DIST.2T(t, df)*, which would yield *T.DIST.2T*(1.82, 78) = .07. Alternatively, the p value could be calculated in the statistical software R using 2**pt(-abs(t), df)*, which would yield 2**pt(-abs*(1.82), 78) = .07. The WWC would not consider this finding to be statistically significant because the p value is larger than the conventional threshold of .05.

The WWC does not apply multiple comparison corrections to the finding-level statistical significance. The WWC instead addresses the issue of multiple comparisons by focusing the interpretation of evidence on the domain-level average effect size. This aggregate can summarize across multiple main findings within an outcome domain, as detailed more in appendix F.

Computing the improvement index

The WWC may translate effect sizes into improvement indices to help readers judge the practical importance of the magnitude of intervention effects. The improvement index is the average expected change in the percentile rank for an average comparison group student that then receives the intervention (or also the difference in percentile ranks for an average intervention versus comparison group student). The beginning of this appendix described an example effect size g = 0.40 translating into an improvement index of +16 percentile points.

Computing the improvement index has two steps:

1. **Convert the effect size to Cohen's U_3 index.** Cohen's U_3 index is the fraction of comparison group students outperformed by the average intervention group student. An effect size of 0.40 corresponds to a U_3 index of 0.66, meaning that the average intervention group student scores higher than 66 percent of comparison group students (assuming normally distributed data[41]). Alternatively, the average intervention group student performs in the 66th percentile of the comparison group distribution. The U_3 index is computed based on the proportion of the area under a standard normal curve that is below the value of the

[41] The U_3 index computations assume equal population variances for the intervention and comparison groups. Although the U_3 estimate is not an unbiased estimate of the tail area, the bias is typically small for studies with sample sizes likely to be found in WWC reviews (Hedges & Olkin, 2016).

effect size. For instance, the U_3 index for an effect size of 0.40 can be computed using the command *pnorm(0.40)* in the statistical software R or *NORM.S.DIST(0.4, TRUE)* in Microsoft Excel.

2. **Compute improvement index = U_3 – 50 percent**. The WWC computes the improvement index by subtracting 50 percent from the U_3 index. This index therefore represents the difference in percentile rank for an average intervention group student versus an average comparison group student (with the percentiles based on the comparison group distribution). For instance, the improvement index for an effect size of 0.40 can be computed in percentage point units using *100*pnorm(0.40) - 50* in R or *100*NORM.S.DIST(0.4, TRUE) - 50* in Microsoft Excel.

The WWC also may compute improvement indices for the domain-level average effect sizes in a study (potentially combining across multiple study findings or outcome measures) and meta-analytic effect sizes that synthesize findings across studies. The WWC will first synthesize the effect sizes and then compute the improvement index for the average effect size, rather than directly average the improvement indices. Appendix F describes the WWC's approach to synthesizing effect sizes.

APPENDIX F. STATISTICAL FORMULAS FOR AGGREGATING STUDY FINDINGS

To determine the magnitude of an aggregate effect, the WWC combines findings in three situations: (a) across subsamples for a single outcome measure within a study, (b) across a study's multiple main findings within an outcome domain, and (c) across studies.

Aggregating across subsamples for a single outcome measure in a study

Some studies present findings separately for several subsamples of subjects without presenting an aggregate result. Examples include presenting results separately for students in grades 6, 7, and 8; high and low-risk students; or demographic subsamples such as boys and girls.

In these situations, the What Works Clearinghouse (WWC) may query authors to learn whether they conducted an analysis on the full sample. The study's analysis is preferred, as it may be more precise than the WWC's computation. If the WWC cannot obtain aggregate results from the author, then the WWC averages results across subsamples for a single outcome measure within a study.

The WWC will use the equations in this subsection if the subsample findings meet all of the following criteria:

- Are independent (nonoverlapping) subsamples.
- Have the same outcome measure.
- Have the same follow-up period.
- Use the same outcome scoring procedures across subsamples.
- Report unadjusted or adjusted means (continuous or dichotomous).

For continuous outcomes, the following equations provide the group mean ($\bar{Y}_j$) combined across all G mutually exclusive subsamples and the combined standard deviation (SD_j):

[F.1] $$\bar{Y}_j = \frac{\sum_{g=1}^{G} n_{gj}\bar{y}_{gj}}{\sum_{g=1}^{G} n_{gj}} \text{ and } SD_j = \sqrt{\frac{\sum_{g=1}^{G}\left[(n_{gj}-1)SD_{gj}^2 + n_{gj}(\bar{Y}_j - \bar{y}_{gj})^2\right]}{\sum_{g=1}^{G} n_{gj} - 1}}$$

where n_{gj}, $\bar{y}_{gj}$, and SD_{gj} are the sample size, outcome mean, and standard deviation for subsample g in group j (intervention or comparison group), respectively.

For dichotomous outcomes, the following equations provide the combined intervention probability P_j and combined comparison probability P_c:

[F.2] $$P_i = \frac{\sum_{g=1}^{G} n_{gi}p_{gi}}{\sum_{g=1}^{G} n_{gi}} \text{ and } P_c = \frac{\sum_{g=1}^{G} n_{gc}p_{gc}}{\sum_{g=1}^{G} n_{gc}}$$

where p_{gi} and p_{gc} are the probabilities of the occurrence of a positive outcome for the intervention and the comparison groups for subsample g, respectively.

The WWC will report the combined finding as a main finding and the original subsample findings as supplemental findings. The effect sizes and standard errors for the combined finding will come from applying the appendix E formulas to the combined sample. For instance, substituting the combined means and standard deviations from equation F.1 into equation E.5 will yield the effect size for the combined finding for continuous outcomes in individual-level assignment studies.

Aggregating a study's multiple main findings within an outcome domain

The WWC will compute a domain-level average effect size if study authors report multiple main findings that meet WWC standards within an outcome domain. Consider a study that reports multiple outcome measures such as two standardized measures of mathematics achievement. The WWC will average the effect sizes for the two measures, provided that the findings for both meet WWC standards and satisfy criteria for being main findings. The WWC's presentation of a study's findings, especially for the characterization of intervention effectiveness, will focus on these domain-level averages rather than on the individual main findings that compose the averages. If a study includes only one main finding within an outcome domain, then the domain-level finding is the same as that single main finding.

The domain-level average effect size is the simple, unweighted average of the effect sizes for the individual main findings that meet WWC standards within an outcome domain:

$$\bar{g} = \frac{1}{K}\sum_{i=1}^{K} g_i \qquad \text{[F.3]}$$

where $\bar{g}$ is the domain-level average effect size, K is the number of main findings that meet WWC standards within an outcome domain, and g_i is the ith main finding. The WWC excludes findings that do not meet WWC standards from this average.

The standard error for this domain-level average effect size is given by the following:

$$SE[\bar{g}] = \frac{1}{K}\sqrt{\sum_{i=1}^{K} SE[g_i]^2 + \rho \sum_{i \neq j} SE[g_i]SE[g_j]} \qquad \text{[F.4]}$$

where ρ is the average correlation among outcome measures, and $SE[g_i]$ and $SE[g_j]$ are the ith and jth effect size standard error.[42] This formula is an approximation in part because it assumes that all interoutcome correlations

[42] The $i \neq j$ summation notation treats pairs as unordered (for example, $i = 2$ and $j = 4$ is distinct from $i = 4$ and $j = 2$), meaning that $\sum_{i \neq j} 1 = k(k-1)$.

are the same.[43] Any missing study correlations relevant to ρ are assumed to be 1.0. This general formula is applicable to any of the effect size and standard error estimators in appendix E, including for individual-level and cluster-level assignment studies.

The WWC determines the statistical significance of the domain-level finding by first computing the *t* test ratio between the domain-level effect size $\bar{g}$ and its standard error $SE[\bar{g}]$ using the formula $t = \bar{g}/SE[\bar{g}]$. The degrees of freedom for determining the *p* value for this *t* statistic is based on averaging the degrees of freedom for the contributing main findings.

The WWC uses the domain-level *t* statistic and degrees of freedom to compute the domain-level *p* value. One example way to compute this *p* value is using the *t*-distribution function in Microsoft Excel: *p = T.DIST.2T(t, df)*. Another example way is to use the statistical software R: *p = 2*pt(-abs(t), df)*. If the *p* value from a two-tailed *t* test is less than .05, then the domain-level finding is statistically significant. No corrections for multiple comparisons are required because the domain-level finding is a single aggregate summarizing across multiple main findings within an outcome domain.

> *Example.* Consider an individual-level assignment randomized controlled trial (RCT) that reported two main findings for the mathematics achievement outcome domain. The findings represent two distinct standardized measures with correlation $\rho = .70$. Both of their effect sizes and standard errors (table F.1) were based on regression-adjusted statistics using relevant formulas from appendix E. Both findings had sample sizes of 50 intervention students and 50 comparison students for each measure, yielding *df* = 98 for the domain-level degrees of freedom.

Table F.1. *Domain-level computations for an example study with two main findings*

Finding	*g*	*SE*	*p*
Mathematics achievement measure 1	0.446	0.153	N/A
Mathematics achievement measure 2	0.348	0.142	N/A
Domain-level average	0.397	0.136	.004

N/A is not applicable.

The domain-level effect size is the simple average of the two effect sizes. The domain-level standard error is based on applying equation F.4 as follows:

$$SE[\bar{g}] = \frac{1}{2}\sqrt{0.153^2 + 0.142^2 + .70(0.153 \times 0.142 + 0.142 \times 0.153)} = 0.136$$

[43] The correlations between outcome measures are used in place of the correlations between effect sizes. In general, the two correlations are very similar, especially when the correlation between measures is positive, which is reasonable in this context. When they differ, the correlation between outcome measures will be slightly larger than the correlation between effect sizes, resulting in a slightly conservative standard error estimate (Thompson & Becker, 2014).

The domain-level *t* statistic is therefore *t* = 0.397/0.136 = 2.919 with 98 degrees of freedom. Using the Excel formula of *T.DIST.2T(2.919, 98)* yields .004 as the *p* value, indicating a statistically significant domain-level finding. Using the R formula of *2*pt(-abs(*2.919), 98*)* yields the same *p* value of .004.

Aggregating findings across studies within an outcome domain

The WWC combines effect sizes across studies for WWC products that include more than one study, such as for intervention reports and practice guides. The WWC computes cross-study, domain-level average effect sizes using a fixed-effects meta-analysis approach.

The WWC chose the fixed-effects model because the WWC aims to make inferences about the studies in WWC intervention reports and practice guides. Unlike the fixed-effect (singular) model, the fixed-effects (plural) model does not assume that the studies are estimating a common effect. Instead, the fixed-effects model assumes that the observed variation among the effect sizes in the meta-analysis reflects the true variation in population effects. Accordingly, inferences to larger study populations are constrained to those that share the same patterns of important study characteristics that are related to effect size.

The WWC's meta-analytic approach gives more weight to studies with more precisely estimated effects. For example, a simple randomized experiment with 300 students will have approximately three times the weight of a simple randomized experiment with 100 students. This approach is similar to how schools compute grade point averages: A grade earned in a three-credit-hour course will have three times the weight of a grade earned in a one-credit-hour course.

Inverse-variance weighting. Effect sizes are weighted by the inverse of their variances (which are largely determined by sample size and design). This procedure is known as inverse-variance weighting. The variance is the square of the standard error. The meta-analytic weight W_s for study s can therefore be written as follows:

$$W_s = \frac{1}{V[\bar{g}_s]} = \frac{1}{SE[\bar{g}_s]^2} \quad \text{[F.5]}$$

where W_s is the meta-analytic weight, $V[\bar{g}_s]$ is the variance, and $SE[\bar{g}_s]$ is the standard error for the domain-level average effect size $\bar{g}_s$ for study s. The study-specific, domain-level effect size and standard error come from equations F.3 and F.4.

The weighted meta-analytic average $\bar{G}$ that combines effect sizes across J studies is then:

$$\bar{G} = \frac{\sum_{s=1}^{J} W_s \bar{g}_s}{\sum_{s=1}^{J} W_s} \quad \text{[F.6]}$$

Only domain-level findings that met WWC standards are included in these averages. The WWC conducts these calculations separately by outcome domain.

The standard error $SE[\bar{G}]$ of the meta-analytic average $\bar{G}$ is given by the following:

$$SE[\bar{G}] = \sqrt{\frac{1}{\sum_{s=1}^{J} W_s}} \quad \text{[F.7]}$$

A statistically significant meta-analytic average is one for which the null hypothesis was rejected using a two-sided z test and a type I error rate of $\alpha = .05$. The z test statistic is computed using the ratio of the meta-analytic effect size and standard error: $z = \bar{G}/SE[\bar{G}]$.

> *Example.* Consider three RCTs that all met WWC standards without reservations and evaluated the same intervention for effects on student mathematics achievement. Table F.2 provides the domain-level effect size and standard error for each study. The study-specific, domain-level averages could each summarize multiple main findings in a study. For instance, the first row corresponds to the domain-level average computed in table F.1.

Table F.2. *Example of fixed-effects meta-analysis*

Finding	*g*	*SE*	*p*
Study 1 domain-level average	0.397	0.136	.004
Study 2 domain-level average	0.413	0.178	.022
Study 3 domain-level average	0.156	0.054	.004
Meta-analytic average	0.205	0.048	<.001

The smaller standard error for study 3 indicates that its effect size was more precisely estimated than for study 1 and study 2, likely reflecting a larger sample size. Hence, study 3 receives more meta-analytic weight. Using equation F.5, the meta-analytic weights for the first, second, and third studies are 54.066, 31.562, and 342.936, respectively.

The meta-analytic average is based on applying equation F.6 as follows:

$$\bar{G} = \frac{54.066 \times 0.397 + 31.562 \times 0.413 + 342.936 \times 0.156}{54.066 + 31.562 + 342.936} = 0.205$$

The meta-analytic standard error is based on applying equation F.7 as follows:

$$SE[\bar{G}] = \sqrt{\frac{1}{54.066 + 31.562 + 342.936}} = 0.048$$

These calculations do not directly depend on the studies' sample sizes or unit of random assignment because the standard errors already incorporate those features. Hence, the study-specific, domain-level standard error acts as a summary statistic that supports meta-analysis across a diverse range of research design features.

Last, the z test statistic is z = 0.205 / 0.048 = 4.251, which corresponds to a p value less than .001, indicating a statistically significant meta-analytic average.

Reweighting by research design of findings. To constrain the bias from findings that have a rating of *Meets WWC Standards With Reservations*, the WWC will reweight the fixed-effects meta-analytic synthesis in certain instances where findings that have a rating of *Meets WWC Standards Without Reservations* do not account for the majority of the meta-analytic weight. This procedure will ensure that the meta-analytic average is based primarily on effect sizes from findings that most credibly facilitate causal inference.

The WWC implements this reweighing procedure only when all the following conditions are met:

1. The synthesis includes one or more findings rated *Meets WWC Standards With Reservations* and one or more findings rated *Meets WWC Standards Without Reservations*.
2. Findings rated *Meets WWC Standards With Reservations* account for over 50 percent of the default inverse-variance meta-analytic weights.
3. The total sum of default inverse weights for findings rated *Meets WWC Standards Without Reservations* is 87.2 or greater, which is the threshold needed to detect an effect size of 0.30 standard deviations at 80 percent power.

If any one of these criteria are not met, then the WWC will synthesize the effect sizes from the corresponding findings using the default inverse-variance weights alone.

If all these criteria are met, the following formula is used to reweight the meta-analytic average:

[F.8] $$\bar{G}^* = \frac{\sum_{s=1}^{J} W_s^* \bar{g}_s}{\sum_{s=1}^{J} W_s^*}$$

The following formula provides the variance of the reweighted meta-analytic average:

[F.9] $$SE[\bar{G}^*] = \frac{\sqrt{\sum_{s=1}^{J} W_s^{*2} SE[\bar{g}_s]^2}}{\sum_{s=1}^{J} W_s^*}$$

where W_i^* is a set of weights rescaled to sum to .49 for studies rated *Meets WWC Standards With Reservations* and .51 for studies rated *Meets WWC Standards Without Reservations*.

Reweighting example. Consider if the first and second findings in table F.2 were rated *Meets WWC Standards Without Reservations*, but the third was rated *Meets WWC Standards With Reservations*. Although the third finding has the lowest research rating, it would carry most of the weight without reweighting, calculated as 342.936 / (54.066 + 31.562 + 342.936) = 80.0% of the total meta-analytic weight.

However, the first and second finding could be reweighted to receive most of the meta-analytic weight instead:

$$W_1^* = .51\frac{54.066}{54.066 + 31.562} = .322$$

$$W_2^* = .51\frac{31.562}{54.066 + 31.562} = .188$$

$$W_3^* = 0.49\frac{342.936}{342.936} = .490$$

The reweighted meta-analytic mean and standard error would be calculated as follows:

$$\bar{G}^* = \frac{.322 \times 0.397 + .188 \times 0.413 + .490 \times 0.156}{0.322 + 0.188 + .490} = 0.282$$

$$SE[\bar{G}^*] = \frac{\sqrt{.322^2 \times 0.136^2 + .188^2 \times 0.178^2 + .490^2 \times 0.054^2}}{.322 + .188 + .490} = 0.061$$

Moderator analysis. In any instance that effect sizes from findings rated *Meets WWC Standards With Reservations* are combined with effect sizes from findings rated *Meets WWC Standards Without Reservations*, the WWC will conduct a test of research rating as an effect size moderator.

When the meta-analytic synthesis includes a combination of studies that used both independent and nonindependent outcome measures, the WWC will conduct a test of the relationship between measure composition (the proportion of independent measures in a study) and effect sizes.

Both sets of moderator tests will use the default inverse-variance weights (specified in equation F.5) to maximize the statistical power of detecting differences in effect sizes.

APPENDIX G. ADDITIONAL DETAIL FOR ANALYSES OF COMPLIER AVERAGE CAUSAL EFFECTS

Reporting requirements for complier average causal effect estimates

Computing the complier average causal effect estimate

Chapter V defines the complier average causal effect (CACE) estimate in general terms. When there is only one instrument, the two-stage least-squares estimate is the same as a ratio in which the numerator is the intent-to-treat estimate and the denominator is the estimated effect of intervention assignment on take-up from the first-stage equation. This ratio is similar to, but more general than, the Bloom (1984) adjustment. The Bloom (1984) estimator is the intent-to-treat estimate divided by the take-up rate in the intervention group. It is equivalent to the two-stage least-squares estimator when there is no take-up in the comparison group and no baseline covariates are included in the analysis. When these two conditions hold, these standards can be applied to studies that use the Bloom adjustment.[44]

Although the two-stage least-squares estimator is the most widely used approach to CACE estimation, other methods exist. Alternative methods include limited information maximum likelihood (Anderson & Rubin, 1949), generalized method of moments (Hansen, 1982), and missing-data methods based on Bayesian procedures or the expectation-maximization algorithm (Imbens & Rubin, 1997a). Because these methods have not been used frequently in education evaluations, the WWC does not include standards that apply to these methods, making analyses using these alternate methods ineligible for review.

Reporting of complier average causal effect estimates

Among randomized controlled trials (RCTs), any CACE analysis that addresses a research topic relevant to a What Works Clearinghouse (WWC) product will be reviewed, so long as it meets the eligibility requirements outlined in Chapter V. However, the ways in which a study's CACE estimates are reported in WWC products will vary depending on the type and focus of the product and the availability of intent-to-treat estimates.

RCTs that report both an intent-to-treat and a CACE estimate on the same outcome. For this type of study, both the intent-to-treat and CACE estimate will be reviewed under their respective standards. The WWC will report the estimates and their ratings as follows:

- If the study is being reviewed for an intervention report or a practice guide, then only one of the two types of estimates will contribute to the effectiveness rating in intervention reports or the level of evidence in practice guides. The lead methodologist for the intervention report, or the evidence coordinator for the practice guide, has the discretion to choose which estimate is used. For example, these individuals may choose based on which type of research question—effects of being assigned to an intervention versus effects of receiving an intervention—is the most common question addressed by other studies in the WWC product. Alternatively,

[44] When members of the assigned comparison group take up the intervention, the Bloom adjustment is not applicable. When the structural equation has baseline covariates, the Bloom adjustment implicitly excludes those covariates from the first-stage equation, leading to underidentification.

the review team leaders may choose based on which type of research question is deemed to be of greatest interest to decisionmakers. After a particular type of estimate (intent-to-treat or CACE) is selected, the other estimate will be mentioned only in a footnote or an appendix.

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

Reporting requirements for variances of complier average causal effect estimates

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

- **Two-stage least-squares asymptotic standard errors.** These standard errors reflect all types of error discussed above. Standard statistical packages report them for two-stage least-squares estimation.
- **Delta method.** In CACE estimates that use a single instrument, the two-stage least-squares estimate is the ratio of the intent-to-treat estimate and the estimated first-stage coefficient on the instrument. The delta method, described by Greene (2000), can be used to express the variance of the CACE estimator as a function of these coefficients, the variance of the intent-to-treat estimator, the variance of the first-stage coefficient, and the covariance between the intent-to-treat estimator and the first-stage coefficient.

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

As for other research designs, the research rating that a CACE estimate receives will not depend on whether standard errors are valid. However, if study authors report an invalid standard error, then the WWC will not use the reported statistical significance of the CACE estimate in characterizing the study's findings.

Rating complier average causal effect estimates when attrition is low

No clear violations of the exclusion restriction (Criterion 1)

Under the exclusion restriction, the only channel through which assignment to the intervention or comparison groups can influence outcomes is by affecting take-up of the intervention being studied (Angrist et al., 1996). The exclusion restriction implies that always-takers in the intervention and comparison groups should not differ in outcomes because their assignment status did not influence their take-up status. Likewise, never-takers in the

intervention and comparison groups should not differ in outcomes. When this condition does not hold, group differences in outcomes would be attributed to the effects of taking up the intervention when they may be attributable to other factors differing between the intervention and comparison groups.

The exclusion restriction cannot be completely verified, as it is impossible to determine whether the effects of assignment on outcomes are mediated through unobserved channels. However, it is possible to identify clear violations of the exclusion restriction—in particular, situations in which groups face different circumstances beyond their differing take-up of the intervention of interest.

Existing WWC standards that prohibit "confounding factors"—that is, factors that differ completely between the assigned groups—already rule out many violations of the exclusion restriction. For example, if groups differ in their eligibility for interventions other than the intervention being studied, then the implied violation of the exclusion restriction also is a confounding factor that, under current WWC group design standards, would cause a study to be rated *Does Not Meet WWC Standards*.

Some scenarios that would be violations of the exclusion restriction do not represent confounding factors in intent-to-treat studies. For instance, the exclusion restriction would be violated if take-up were defined inconsistently between the assigned intervention group and assigned comparison group. For example, suppose that take-up in the assigned intervention group was defined as enrolling in the intervention being studied, such as an intensive afterschool program, whereas take-up in the assigned comparison group was defined as enrolling in the specified intervention or similar interventions, such as attending any program after school. In this case, differences in outcomes between assigned groups might not be attributable solely to differences in rates of take-up as defined by the study because the two take-up rates measure different concepts.

Another violation of the exclusion restriction that does not necessarily stem from a confounding factor is the scenario in which assignment to the intervention group changes the behavior of subjects even if they do not take up the intervention itself. For example, in an experiment to test the effectiveness of requiring unemployed workers to receive job-search and training services, assignment to the intervention group might motivate subjects to search for a job to avoid having to participate in the intervention services. In this case, the intervention assignment might affect outcomes through channels other than the take-up rate. Judgment is required to determine whether a potential unintended channel for group status to influence outcomes is important enough to undermine the internal validity of a CACE estimate. Under this guidance, the WWC's lead methodologist for a review has the responsibility to make this judgment.

Sufficient instrument strength (Criterion 2)

The condition of sufficient instrument strength requires that the group assignment indicators—that is, the instrumental variables—collectively serve as strong predictors of take-up, the endogenous independent variable. As discussed next, this condition is necessary for conventional statistical tests based on two-stage least-squares estimators to have low type I (false positive) error rates.

The need for sufficient instrument strength stems from the statistical properties of two-stage least-squares estimators. An extensive statistical literature has demonstrated that, in finite samples, two-stage least-squares

estimators of CACE impacts include part of the bias of ordinary least-squares estimates (Basmann, 1974; Bloom et al., 2010; Bound et al., 1995; Buse, 1992; Nelson & Startz, 1990; Richardson, 1968; Sawa, 1969).[45] Moreover, in finite samples, two-stage least-squares estimators do not have a normal distribution (that is, the distribution typically used to construct confidence intervals). For these reasons, conventional statistical tests—such as *t* tests and *F* tests—based on two-stage least-squares estimators in finite samples have actual type I error rates that generally are higher than the assumed type I error rates (Stock & Yogo, 2005). For instance, a *t* test conducted at an assumed 5 percent significance level will have an actual type I error rate exceeding 5 percent.

In a limited set of circumstances, the WWC will be able to calculate the first-stage *F* statistic even if this statistic is not reported by the study and cannot be obtained through an author query. Specifically, in the case of a study with no clustering and one instrumental variable that distinguishes a single intervention group and a single comparison group, the WWC can obtain a conservative, lower-bound value for the first-stage *F* statistic if certain information is available. The needed information is the take-up rate for analysis sample members in the intervention group ($\overline{D}_{i,an}$), the take-up rate for analysis sample members in the comparison group ($\overline{D}_{c,an}$), the number of analysis sample members in the intervention group (n_i), and the number of analysis sample members in the comparison group (n_c). The first-stage *F* statistic is represented as:

[G.1] $$F = \frac{(\overline{D}_{i,an} - \overline{D}_{c,an})^2}{\frac{\overline{D}_{i,an}(1-\overline{D}_{i,an})}{n_i} + \frac{\overline{D}_{c,an}(1-\overline{D}_{c,an})}{n_c}}$$

which is a lower-bound value because it does not take into account precision gains from controlling for other covariates in the first-stage equation.

The bias issue with two-stage least-squares estimators shrinks as the instruments become stronger predictors of the endogenous independent variable. An instrument is considered a stronger predictor of an endogenous independent variable if the association between the instrument and endogenous independent variable is large or the association is precisely estimated. In the context of estimating CACE effects, group status is a stronger instrument when group take-up rates differ more and when sample sizes are larger.

Instruments must be strong enough for statistical tests of two-stage least-squares estimators to have "acceptably" low type I error rates. As instruments become stronger, the probability distributions of two-stage least-squares estimators converge to normal distributions centered on the true CACE impact. Type I error rates follow suit and converge to their assumed levels. The previous sentence puts "acceptably" in quotes because defining what is acceptable requires its own standard, which is explained next.

[45] As discussed by Bloom et al. (2010), the finite-sample bias of instrumental variable estimators originates from sampling error. Due to finite samples, random assignment will produce intervention and comparison groups that, by chance, are not fully identical on the characteristics of group members. Some of these unobserved characteristics exert influences on both take-up and outcomes. For illustrative purposes, suppose take-up and outcomes are positively correlated due to these unobserved influences. When sampling error leads to greater (or smaller) differences in take-up between the intervention and comparison groups, greater (or smaller) differences in outcomes arise. Although both types of differences result from random imbalances, the differences are systematically related, creating a spurious association between take-up and outcomes.

Selecting the maximum tolerable type I error rate is the first step in establishing a criterion for sufficient instrument strength. WWC standards do not provide a precedent for acceptable rates of type I error but do provide a precedent for acceptable levels of bias in impact estimates, which is 0.05 standard deviation. The following section uses this precedent to set acceptable type I error rates, using a statistical framework that links type I error rates to estimation bias. Using this framework, for a *t* test whose assumed type I error rate is .05, ensuring a bias of less than 0.05 standard deviation implies actual type I error rates of less than .10. Thus, the guidelines for instrument strength specified here are based on an upper limit of .10 for the type I error rate.

Linking complier average causal effect estimation bias with type I error rates

This section provides a statistical framework for deriving the relationship between the bias of an impact estimator and the estimator's type I error rate, focusing on a conventional *t* test. Setting a maximum tolerable bias—for which there is precedent in WWC standards—implies setting a maximum tolerable type I error rate.

Consider a situation in which the true impact of an intervention β_1 is zero. A biased estimator of this impact $\hat{\beta}_1^{biased}$ will have a distribution centered on a value different from zero. Larger bias increases type I error. As the distribution of the estimator lies further away from zero, there is a greater likelihood of incorrectly rejecting the hypothesis of a zero impact, assuming correct variances are estimated.

To derive the relationship between bias and type I error rates, the distribution of the two-stage least-squares estimator is not useable because its distribution has neither an expected value, when only one instrument is employed, nor a familiar distribution in finite samples (Stock et al., 2002). Instead consider a generic estimator expressed in effect size units $\hat{\beta}_1^{biased}$ that is distributed normally with an expected value equal to $bias > 0$ standard deviations when the true impact is zero. The probability of a type I error using a 5 percent significance test is:

[G.2]
$$\begin{aligned} Type\ I\ error\ rate &= Pr\left(\frac{\hat{\beta}_1^{biased}}{SE\left(\hat{\beta}_1^{biased}\right)} > z_{0.975}\right) + Pr\left(\frac{\hat{\beta}_1^{biased}}{SE\left(\hat{\beta}_1^{biased}\right)} < z_{0.025}\right) \\ &= 1 - \Phi\left(z_{0.975} - \frac{bias}{SE\left(\hat{\beta}_1^{biased}\right)}\right) + \Phi\left(z_{0.025} - \frac{bias}{SE\left(\hat{\beta}_1^{biased}\right)}\right) \end{aligned}$$

where $SE(\bullet)$ denotes the standard error of an estimator, z_q is the qth quantile of the standard normal distribution, and $\Phi(\bullet)$ is the cumulative distribution function of the standard normal distribution.

Equation G.2 provides the relationship between the type I error rate and bias as long as the standard error of the biased estimator is known. Therefore, to specify this relationship fully, a value for the standard error is needed. The standard error can vary depending on sample size, covariates, degree of clustering, and other factors. Picking a standard error essentially entails choosing a "benchmark" level of precision to complete the specification of equation G.2.

As the benchmark, assume a level of precision corresponding to a study for which the minimum detectable effect size is 0.25 standard deviation. A value for minimum detectable effect size, in turn, directly implies a value for the standard error. Specifically, the minimum effect size that can be detected using a two-tailed test at a

5 percent significance level with 80 percent power can be expressed as a function of the standard error of the effect size $SE[g]$ (see Bloom, 2005):

[G.3] $$MDES = SE[g][\Phi^{-1}(1 - 0.05/2) + \Phi^{-1}(0.8)] = 2.802 \times SE[g]$$

Using equation G.3, a study designed to have a minimum detectable effect size of 0.25 is expected to have a standard error of 0.09 standard deviation (0.25 / 2.802).

Substituting the benchmark standard error, 0.09 standard deviation, for $SE(\hat{\beta}_1^{biased})$ in equation G.2 completely specifies the relationship between the type I error rate and the amount of bias. Equation G.2 becomes:

[G.4] $$Type\ I\ error\ rate = 1 - \Phi\left(z_{0.975} - \frac{bias}{0.09}\right) + \Phi\left(z_{0.025} - \frac{bias}{0.09}\right)$$

The final step is to substitute into equation G.4 a value for *bias* that represents the maximum tolerable bias. As discussed, the maximum value for *bias* that is acceptable to the WWC is 0.05 standard deviation. Setting *bias* = 0.05 in equation G.4 obtains a maximum tolerable type I error rate equal to:

[G.5] $$Maximum\ tolerable\ type\ I\ error\ rate = 1 - \Phi\left(z_{0.975} - \frac{0.05}{0.09}\right) + \Phi\left(z_{0.025} - \frac{0.05}{0.09}\right) = .086$$

The maximum tolerable type I error rate then determines the minimum required first-stage *F* statistic for sufficient instrument strength. For a given number of instruments, Stock and Yogo (2005) calculated several different values for the minimum required first-stage *F* statistic, depending on whether the maximum tolerable type I error rate is .10, .15, .20, or .25. For setting the WWC standard, the preceding calculations yield a maximum tolerable type I error rate of .086, which is rounded to .10, the closest value addressed by Stock and Yogo (2005). This value determines the minimum required first-stage *F* statistic based on Stock and Yogo's (2005) calculations.

Calculating attrition in a complier average causal effect analysis

When there are two random assignment groups

When rating CACE estimates, the basic approach to determining whether attrition is low or high will follow the usual attrition standard for RCTs (see Chapter III, Compositional change). Both overall and differential attrition must be calculated. Table C.1 in appendix C will determine whether the combination of overall and differential attrition is considered low or high.

However, the specific method for calculating attrition rates when rating CACE estimates is different from the method used when rating intent-to-treat estimates. When rating intent-to-treat estimates, the overall attrition rate is the fraction of the entire randomly assigned sample that did not contribute outcome data to the final analysis. Likewise, the differential attrition rate is the difference in attrition rates between the entire assigned intervention group and entire assigned comparison group. It is appropriate to measure attrition for the entire sample when rating intent-to-treat estimates because those estimates are intended to represent how assignment to the intervention would, on average, affect all subjects.

In contrast, a CACE estimate represents the average effect of taking up the intervention for compliers only. Accordingly, when rating a CACE estimate, the WWC will calculate overall and differential attrition rates that pertain specifically to compliers. Because compliers cannot be directly identified, as discussed previously, the attrition rates for compliers likewise cannot be directly calculated. Instead, the attrition rates must be estimated on the basis of specific assumptions, discussed next.

For the usual scenario in which there are two assigned groups—that is, the intervention group, denoted by $Z = i$, and the comparison group, denoted by $Z = c$, the differential attrition rate for compliers $\hat{\Delta}^{complier}$ will be estimated as:

$$[G.6] \qquad \hat{\Delta}^{complier} = \frac{\bar{A}_i - \bar{A}_c}{\bar{D}_{i,ran} - \bar{D}_{c,ran}}$$

where $\bar{A}_z$ is the attrition rate in the assigned group $Z = z$, and $\bar{D}_{z,ran}$ is the fraction of the assigned group $Z = z$ that took up the intervention. The numerator of equation G.6 is the differential attrition rate that the WWC calculates when rating intent-to-treat estimates. The denominator is the difference in take-up rates between assigned groups. Equation G.6 provides a consistent estimate of the differential attrition rate for compliers under the assumption that attrition rates for always-takers and never-takers do not differ by assigned status. More generally, equation G.6 provides a conservative, upper-bound estimate of the differential attrition rate for compliers under the assumption that differential attrition rates for always-takers and never-takers, if nonzero have the same sign as the differential attrition rate for compliers. The WWC regards the latter assumption as reasonable and realistic. It is difficult to identify scenarios in which assignment to an intervention would influence attrition patterns in opposite ways for always-takers and never-takers.[46]

Calculating the overall attrition rate for compliers involves calculating the attrition rate for compliers in the intervention and comparison groups separately, then taking a weighted average of the two attrition rates, with weights equal to group size. Let $\bar{A}_{zd}$ be the observed attrition rate for people with assignment status $Z = z$ and take-up status $D = d$, with $D = 1$ denoting receipt of the intervention and $D = 0$ denoting nonreceipt. Following Imbens and Rubin (1997b), the attrition rate for compliers in the comparison group $R_c^{complier}$ will be estimated as[47]:

$$[G.7] \qquad \hat{R}_c^{complier} = \frac{\left(1 - \bar{D}_{c,ran}\right)\bar{A}_{c0} - \left(1 - \bar{D}_{i,ran}\right)\bar{A}_{i0}}{\bar{D}_{i,ran} - \bar{D}_{c,ran}}$$

[46] In most cases, attrition is due to missing outcome data. Less frequently, attrition may be due to missing data on take-up status. If some members of the randomly assigned sample are missing take-up status, then the WWC will not have all the information needed for calculating the denominator of equation G.6. In this case, the WWC assumes a worst-case scenario, in which individuals in the intervention group with missing take-up status truly did not take up the intervention and individuals in the comparison group with missing take-up status truly took up the intervention. This worst-case scenario minimizes the denominator in equation G.6 and, therefore, leads to an upper-bound for the differential attrition rate.

[47] The intuition behind equation G.7 is roughly as follows. Members of the assigned comparison group who do not take up the intervention consist of a mix of compliers and never-takers. Starting from the attrition rate for this mixed group, the first term in the numerator of equation G.7, remove the contribution coming from comparison-group never-takers, which is assumed to be equivalent to the observed attrition rate of never-takers in the intervention group, the second term in the numerator of equation G.7. The resulting difference is an estimate of the attrition rate for comparison-group compliers.

The attrition rate for compliers in the intervention group, $R_i^{complier}$, will then be estimated as:

$$[G.8] \qquad \hat{R}_i^{complier} = \hat{R}_c^{complier} + \hat{\Delta}_{complier}$$

The overall attrition rate $R_{overall}^{complier}$ will then be estimated as:

$$[G.9] \qquad \hat{R}_{overall}^{complier} = \frac{\hat{R}_i^{complier} n_{i,ran} + \hat{R}_c^{complier} n_{c,ran}}{n_{i,ran} + n_{c,ran}}$$

where $n_{i,ran}$ and $n_{c,ran}$ are the number of sample members randomly assigned to the intervention and comparison groups, respectively.

The procedure described thus far in this section is equivalent to using the units of analysis to estimate a two-stage least-squares regression in which attrition is the outcome, a take-up indicator is the endogenous independent variable, and an indicator for assignment to the intervention group is the instrumental variable. The estimated coefficient on the take-up indicator is equivalent to the differential attrition rate shown in equation G.6. The WWC will use the result from this two-stage least-squares regression as the measure of differential attrition if it is provided.

When there are three or more random assignment groups

This section considers the scenario in which each sample member could be randomly assigned to three or more groups. Consider an example of three groups including a group that is ineligible for the intervention, a group that has low priority for the intervention, and a group that has high priority for the intervention. Even though there are multiple assigned groups, only one intervention is studied. Hence, there is still only a single measure of take-up—a binary variable for taking up any portion of the intervention.

In this case, the WWC's approach is equivalent to estimating a two-stage least-squares regression in which attrition is the outcome, a take-up indicator is the endogenous independent variable, and multiple assignment group indicators (one for each group except an omitted reference group) are the instrumental variables. This approach involves first ordering the assigned groups from the lowest to the highest take-up rate. For each comparison between consecutively ordered groups, the WWC will apply equations G.6 through G.9 to obtain differential and overall attrition rates for compliers relevant to that comparison—that is, for subjects who are induced to take up the intervention by being assigned to the higher-ordered group instead of the lower-ordered group. The next step is then taking a weighted average of both the overall and differential attrition rate across those different comparisons, using weights defined in the following section (Imbens & Angrist, 1994).

First, order the assigned groups with the index $g = 0, 1, 2, \ldots G$ from lowest to highest take-up rate. Assume that any sample member who would take up the intervention if assigned to group g would also take up the intervention if assigned to a group ordered after g. This assumption is called the monotonicity assumption (Imbens & Angrist, 1994). For each comparison between group $(g - 1)$ and group g, compliers are defined as those who would take up the intervention if assigned to group g but not if assigned to group $(g - 1)$. The two-stage least-squares estimator of the CACE is a weighted average of complier impacts across these comparisons, with weights

given by Imbens and Angrist (1994). The WWC's method for calculating attrition follows the same approach. The WWC calculates attrition (both overall and differential) for each comparison between consecutively ordered groups. A weighted average is taken across those comparisons, using the same weights as those in the two-stage least-squares estimator.

Specifically, let $\hat{\Delta}_{g,g-1}^{\text{complier}}$ be the differential attrition rate for compliers pertaining to the comparison between groups (g - 1) and g, based on applying equation G.6. The final differential attrition rate for all compliers $\hat{\Delta}_{\text{final}}^{\text{complier}}$ is calculated as:

[G.10] $$\hat{\Delta}_{final}^{complier} = \frac{\sum_{g=1}^{G} w_g \hat{\Delta}_{g,g-1}^{complier}}{\sum_{g=1}^{G} w_g}$$

where w_g is the weight on the comparison between groups (g - 1) and g. Imbens and Angrist (1994) derived the weight to be:

[G.11] $$w_g = \left(\bar{D}_{g,ran} - \bar{D}_{g-1,ran}\right) \sum_{j=g}^{G} \frac{n_{j,ran}}{N_{ran}} \left(\bar{D}_{j,ran} - \bar{D}_{ran}\right)$$

where $\bar{D}_{j,ran}$ is the take-up rate for sample members assigned to group j, $\bar{D}_{ran}$ is the take-up rate in the entire randomly assigned sample, $n_{j,ran}$ is the number of sample members assigned to group j, and N_{ran} is the total number of sample members in the entire randomly assigned sample.

For calculating overall attrition, the same weights are used to take a weighted average of the overall complier attrition rates across all comparisons.

APPENDIX H. ADDITIONAL DETAIL FOR ANALYSES WITH MISSING DATA

Technical details on acceptable approaches for addressing missing data

Table H.1 elaborates on table 20 in Chapter V, providing additional detail for the acceptable approaches to addressing missing baseline or outcome data.

All but one of the acceptable approaches in table H.1 can provide unbiased estimates of the effectiveness of an intervention based on the assumption that the missing data do not depend on unmeasured factors. The exception is complete case analysis, which requires a more restrictive assumption that the missing data also do not depend on measured factors. Because of this, many researchers have recommended against using complete case analysis to address missing data (for example, Little et al., 2012; Peugh & Enders, 2004). Nevertheless, the WWC considers complete case analysis to be an acceptable approach for addressing missing data because possible bias due to measured factors can be assessed through the attrition standard and WWC's baseline adjustment requirement.

In addition, Jones (1996) and Allison (2002) raised concerns about using the approach in the last row of table H.1, imputation to a constant combined with including a missing data indicator, outside of randomized controlled trials (RCTs). Consequently, the WWC considers this approach acceptable for any baseline data in RCTs regardless of their sample attrition. However, in a QED or compromised RCT, the approach is acceptable only when applied to baseline measures not specified in this *Handbook* as required for assessing baseline equivalence.

Table H.1. *Acceptable approaches for addressing missing baseline or outcome data*

Approach	Description	WWC requirements	Statistical significance
Complete case analysis	Exclusion of observations with missing outcome and/or baseline data from the analysis.	None.	The WWC has no additional requirements for reporting statistical significance from analyses that use this method.

Continued on next page

Table H.1. *Acceptable approaches for addressing missing baseline or outcome data (continued)*

Approach	Description	WWC requirements	Statistical significance
Regression imputation	A regression model to predict imputed values for the missing data. This process includes estimating imputed values from a single regression model, and multiple imputation, which involves generating multiple datasets that contain imputed values for missing data through the repeated application of an imputation algorithm, such as chained equations.	The imputation regression model must: • Be conducted separately for the intervention and comparison groups or include an indicator variable for intervention status, • Include all of the covariates that are used for statistical adjustment in the impact estimation model, and • Include the outcome when imputing missing baseline data.	Standard errors must be computed using a method that reflects the missing information, such as a bootstrap method, or multiple imputation. For multiple imputation, the statistical significance calculation must: • Be based on at least five sets of imputations, and • Account for (1) the within-imputation variance component, (2) the between-imputation variance component, and (3) the number of imputations. Most established multiple imputation routines satisfy this requirement. A cluster-level assignment study with missing outcome data, analyzed using individual-level data, must provide evidence that the approach appropriately adjusts the standard errors for clustering by citing a peer-reviewed journal article or textbook that describes the procedure and demonstrates its effectiveness.
Maximum likelihood	An iterative routine to estimate model parameters and impute values for the missing data. Some examples are the expectation-maximization algorithm and full information maximum likelihood.	The procedure must use a standard statistical package or be supported with a citation to a peer-reviewed methodological journal article or textbook.	Standard errors must be computed using a method that reflects the missing information, such as a bootstrap method, or estimates based on the information matrix.

Continued on next page

Table H.1. *Acceptable approaches for addressing missing baseline or outcome data (continued)*

Approach	Description	WWC requirements	Statistical significance
Nonresponse weights	Use of weights based on estimated probabilities of having a nonmissing outcome, yielding greater weight for subjects with a higher probability of having missing outcome data. For example, the probabilities may be estimated from a logit or probit model.	Acceptable only for missing outcome data, not for missing baseline data. The estimated probabilities used to construct the weights must: • Be estimated separately for the intervention and comparison groups or include an intervention status indicator, and Include all baseline measures that are specified in the *Handbook* as required for use in statistical adjustments. Including additional covariates is acceptable but not required.	The analysis must properly account for the stratified sampling associated with the weights (as discussed in Wooldridge, 2002, p. 594). A cluster-level assignment study with missing outcome data, analyzed using individual-level data, must provide evidence that the approach appropriately adjusts the standard errors for clustering by citing a peer-reviewed journal article or textbook that describes the procedure and demonstrates its effectiveness.
Replacing missing data with a constant combined with including a missing data indicator	Setting all missing values for a baseline measure to a single value and including an indicator variable for records missing data on the measure in the impact estimation model.	• Acceptable only for missing baseline data, not for missing outcome data. When applied to a measure required for baseline adjustment, the method is acceptable only in RCTs regardless of sample attrition, but not in QEDs or compromised RCTs.	The WWC has no additional requirements for reporting statistical significance from analyses that use this method.

QED is quasi-experimental design. RCT is randomized controlled trial. WWC is What Works Clearinghouse.
Note: Requirements in this table are based on recommendations in several sources, including Allison (2002); Azur et al. (2011); Little and Rubin (2002); Puma et al. (2009); Rubin (1987); Schafer (1999); and Wooldridge (2002).

Assessing the bias from imputed outcome data when the baseline measure is observed for all subjects in the analytic sample

The imputation methods the WWC considers acceptable require assuming that data are missing at random, which means the missing data depend on measured factors but not on unmeasured factors. If that assumption does not hold, then the impact estimates may be biased. Therefore, QEDs and high-attrition RCTs that use acceptable approaches to impute outcome data must demonstrate that they limit the potential bias from using imputed data to measure impacts. Specifically, potential bias due to deviations from the missing-at-random assumption must not exceed 0.05 standard deviation.

The WWC uses a proxy pattern-mixture modeling approach to estimate the largest possible bias in an impact estimate under a set of reasonable assumptions about how the missing data are related to measured and unmeasured factors (Andridge & Little, 2011).

Assume that the probability of observing an outcome for a given subject is related to the baseline measure and the outcome, which may be unmeasured in some cases. This probability in the intervention group ($j = i$) or comparison group ($j = c$) is given by the following function h:

[H.1] $$p_j(x, y) = h\left(\frac{x}{SD_x} + \lambda_j \frac{y}{SD_y}\right)$$

where x is the baseline measure for a subject, y is the outcome measure for the subject, SD_x and SD_y are the standard deviations of the baseline and outcome measures, and λ_j measures the deviations from the missing-at-random assumption for group j. When λ_j = 0, the missing-at-random assumption holds for group j because the missing data depend only on measured baseline data. As λ_j increases, the missingness depends more strongly on the outcome, which may be unmeasured.

Following Andridge and Little (2011), the unmeasured full-sample outcome mean in a group $(\bar{y}_j)$ can be written as a function of the complete case outcome mean $(\bar{y}_{jR})$, the full-sample and complete case baseline means $(\bar{x}_j \text{ and } \bar{x}_{jR})$, and the correlation between the outcome and the baseline measure ρ_{cor}:

[H.2] $$\bar{y}_j = \bar{y}_{jR} + f_j(\rho_{cor}) \frac{SD_y}{SD_x} [\bar{x}_j - \bar{x}_{jR}]$$

where the function of ρ_{cor} is assumed to be:

[H.3] $$f_j(\rho_{cor}) = \frac{\lambda_j + \rho_{cor}}{\lambda_j \rho_{cor} + 1}$$

In many cases, the value of $\bar{y}_j$ will deviate more from the observed mean of $\bar{y}_{jR}$ when there is a larger absolute difference between the full-sample and complete case baseline means. Intuitively, this is because a larger difference means that the subjects with missing outcome data appear different from those with observed outcomes.

When data are missing at random, $f_i(\rho_{cor}) = f_c(\rho_{cor}) = \rho_{cor}$ (because $\lambda_i = \lambda_c = 0$), and the expected value of $\bar{y}_j$ is equal to what a researcher would obtain for the full-sample outcome mean when imputing missing values of the outcome measure with predicted values from a regression of the outcome on the baseline measure. But as λ_i or λ_c become larger, the value of $f_i(\rho_{cor})$ becomes larger (approaching $1/\rho_{cor}$), and the outcome mean for the full sample will deviate from the researcher's estimate of the mean using imputed data.

The effect size obtained using an imputation method based on the missing-at-random assumption can be written as the difference in the estimated full-sample intervention and comparison group outcome means with an adjustment for the baseline measure, given by:

[H.4] $$g_{MAR} = \frac{1}{SD_y}(\{\bar{y}_{iR} + c[\bar{x}_i - \bar{x}_{iR}]\} - \{\bar{y}_{cR} + c[\bar{x}_c - \bar{x}_{cR}]\} - c[\bar{x}_i - \bar{x}_c])$$

where c is the coefficient from a regression of y on x, and is equal to $\rho_{cor}(SD_y/SD_x)$

But this equation can be generalized to when the missing-at-random assumption does not hold:

[H.5] $$g_{NMAR} = \frac{1}{SD_y}\left(\left\{\bar{y}_{iR} + f_i(\rho_{cor})\frac{SD_y}{SD_x}[\bar{x}_i - \bar{x}_{iR}]\right\} - \left\{\bar{y}_{cR} + f_c(\rho_{cor})\frac{SD_y}{SD_x}[\bar{x}_c - \bar{x}_{cR}]\right\} - c[\bar{x}_i - \bar{x}_c]\right)$$

Comparing g_{MAR} and g_{NMAR} gives the bias due to deviations from the missing-at-random assumption:

[H.6] $$Bias_y = \frac{1}{SD_x}\{(f_i(\rho_{cor}) - \rho_{cor})[\bar{x}_i - \bar{x}_{iR}] - (f_c(\rho_{cor}) - \rho_{cor})[\bar{x}_c - \bar{x}_{cR}]\}$$

Because $f_j(\rho_{cor})$ is bounded between ρ_{cor} and $1/\rho_{cor}$, the largest bias, in absolute value, due to deviations from the missing-at-random assumption is given by the maximum of the values given by the following three equations:

[H.7] $$B1 = \omega\left|\frac{1}{SD_x}\frac{1-\rho_{cor}^2}{\rho_{cor}}[\bar{x}_c - \bar{x}_{cR}]\right|$$

[H.8] $$B2 = \omega\left|\frac{1}{SD_x}\frac{1-\rho_{cor}^2}{\rho_{cor}}[\bar{x}_i - \bar{x}_{iR}]\right|$$

[H.9] $$B3 = \omega\left|\frac{1}{SD_x}\frac{1-\rho_{cor}^2}{\rho_{cor}}[(\bar{x}_i - \bar{x}_{iR}) - (\bar{x}_c - \bar{x}_{cR})]\right|$$

The bounds in equations H.7, H.8, and H.9 will be calculated using data reported in studies or obtained from authors. The equations include the following data elements: (a) the means and standard deviations of the baseline measure for the analytic sample, separately for the intervention and comparison groups ($\bar{x}_i$, $\bar{x}_c$, and the standard deviations are used to calculate the pooled within-group standard deviation SD_x); (b) the means of the baseline measure for the subjects in the analytic sample with observed outcome data, separately for the intervention and comparison groups ($\bar{x}_{iR}, \bar{x}_{cR}$); and (c) the correlation between the baseline and the outcome measures (ρ_{cor}).

For simplicity, these bounds were derived for a single baseline measure. If multiple baseline measures were used to form the imputed values in a study, it is acceptable, but not required, to replace the baseline means with the average predicted value of the outcome, that is, the average of the values used to make adjustments to the outcome measure to produce an adjusted mean. In this case, $1/SD_x$ is removed from the calculation of the bounds and replaced with $1/SD_{\hat{y}}$ because the predicted values have units of the dependent variable, where $SD_{\hat{y}}$ is the standard deviation of the predicted values of the outcome. The ρ_{cor} correlation parameter would be replaced with $\sqrt{R^2}$ where R^2 is the multiple squared correlation between the baseline covariates and outcome. Additionally, for outcome domains that require baseline adjustments on multiple baseline measures, it is

required that the imputed values adjust for all baseline measures specified in the review protocol and that the bounds are calculated using the average of the predicted values.

Assessing the bias from imputed outcome data when the baseline measure is imputed or missing for some subjects in the analytic sample

When an analytic sample includes both imputed outcome data and missing or imputed baseline data, it is not possible to calculate the bounds in equations H.7, H.8, and H.9. This limitation arises because the means of the baseline measure are unknown for the analytic sample and are possibly unknown for the restricted sample of subjects with observed outcome data.

Instead, these bounds can be derived by first writing the full sample outcome mean as a weighted sum of the outcome mean for the sample with missing data on the baseline measure, and the sample with observed data on the baseline measure:

$$[H.10] \qquad \bar{y}_j = \left(\frac{n_j - n_{jx}}{n_j}\right)\bar{y}_{j\sim x} + \left(\frac{n_{jx}}{n_j}\right)\bar{y}_{jx}$$

where n_j is the number of observations in the analytic sample for group j, n_{jx} is the number of observations in the analytic sample for group j with an observed value of the baseline measure, $\bar{y}_{j\sim x}$ is the outcome mean for the observations in the analytic sample for group j missing the baseline measure, and $\bar{y}_{jx}$ is the outcome mean for the remaining members of the analytic sample for group j.

Assume that the analytic sample includes no cases where both the baseline and outcome data are missing, so $\bar{y}_{j\sim x}$ is observed. But $\bar{y}_{jx}$ is not observed because some cases with observed baseline data have missing outcome data. To address this, $\bar{y}_{jx}$ can be written as a function of observed measures:

$$[H.11] \qquad \bar{y}_j = \left(\frac{n_j - n_{jx}}{n_j}\right)\bar{y}_{j\sim x} + \left(\frac{n_{jx}}{n_j}\right)\left(\bar{y}_{jxy} + f_j(\rho_{cor})\frac{SD_y}{SD_x}\left[\bar{x}_{jx} - \bar{x}_{jxy}\right]\right)$$

where $\bar{y}_{jxy}$ is the outcome mean for the observations in the complete case analytic sample for group j observed at both baseline and for the collection of outcomes, $\bar{x}_{jxy}$ is the baseline mean for the same sample, and $\bar{x}_{jx}$ is the baseline mean for the sample with observed baseline data but possibly missing outcome data. This equation can be rewritten as:

$$[H.12] \qquad \bar{y}_j = \bar{y}_{jxy} + \left(\frac{n_j - n_{jx}}{n_j}\right)\left[\bar{y}_{j\sim x} - \bar{y}_{jxy}\right] + \left(\frac{n_{jx}}{n_j}\right)f_j(\rho_{cor})\frac{SD_y}{SD_x}\left[\bar{x}_{jx} - \bar{x}_{jxy}\right]$$

The effect size obtained using an imputation method based on the missing-at-random assumption ($f_j(\rho_{cor}) = \rho_{cor}$) can be written as the difference in the estimated full-sample intervention and comparison group outcome means, given by:

[H.13]
$$g_{MAR} = \frac{1}{SD_y}\left(\left\{\bar{y}_{ixy} + \left(\frac{n_i - n_{ix}}{n_i}\right)\left[\bar{y}_{i\sim x} - \bar{y}_{ixy}\right] + \left(\frac{n_{ix}}{n_i}\right)c\left[\bar{x}_{ix} - \bar{x}_{ixy}\right]\right\} - \left\{\bar{y}_{cxy} + \left(\frac{n_c - n_{cx}}{n_c}\right)\left[\bar{y}_{c\sim x} - \bar{y}_{cxy}\right] + \left(\frac{n_{cx}}{n_c}\right)c\left[\bar{x}_{cx} - \bar{x}_{cxy}\right]\right\}\right)$$

The more general equation that allows deviations from the missing-at-random assumption is given by:

[H.14]
$$g_{NMAR} = \frac{1}{SD_y}\left(\left\{\bar{y}_{ixy} + \left(\frac{n_i - n_{ix}}{n_i}\right)\left[\bar{y}_{i\sim x} - \bar{y}_{ixy}\right] + \left(\frac{n_{ix}}{n_i}\right)f_i(\rho_{cor})\frac{SD_y}{SD_x}\left[\bar{x}_{ix} - \bar{x}_{ixy}\right]\right\} - \left\{\bar{y}_{cxy} + \left(\frac{n_c - n_{cx}}{n_c}\right)\left[\bar{y}_{c\sim x} - \bar{y}_{cxy}\right] + \left(\frac{n_{cx}}{n_c}\right)f_c(\rho_{cor})\frac{SD_y}{SD_x}\left[\bar{x}_{cx} - \bar{x}_{cxy}\right]\right\}\right)$$

Comparing g_{MAR} and g_{NMAR} gives the bias due to deviations from the missing-at-random assumption:

[H.15]
$$bias = \frac{1}{SD_x}\left\{\left(\frac{n_{ix}}{n_i}\right)\left(f_i(\rho_{cor}) - \rho_{cor}\right)\left[\bar{x}_{ix} - \bar{x}_{ixy}\right] - \left(\frac{n_{cx}}{n_c}\right)\left(f_c(\rho_{cor}) - \rho_{cor}\right)\left[\bar{x}_{cx} - \bar{x}_{cxy}\right]\right\}$$

The absolute value of this bias is no greater than the maximum of *B*1* - *B*3*:

[H.16]
$$B1^* = \omega\left|\frac{1}{SD_x}\frac{1-\rho_{cor}^2}{\rho_{cor}}\left(\frac{n_{ix}}{n_i}\right)\left[\bar{x}_{ix} - \bar{x}_{ixy}\right]\right|$$

[H.17]
$$B2^* = \omega\left|\frac{1}{SD_x}\frac{1-\rho_{cor}^2}{\rho_{cor}}\left(\frac{n_{cx}}{n_c}\right)\left[\bar{x}_{cx} - \bar{x}_{cxy}\right]\right|$$

[H.18]
$$B3^* = \omega\left|\frac{1}{SD_x}\frac{1-\rho_{cor}^2}{\rho_{cor}}\left[\left(\frac{n_{ix}}{n_i}\right)\left(\bar{x}_{ix} - \bar{x}_{ixy}\right) - \left(\frac{n_{cx}}{n_c}\right)\left(\bar{x}_{cx} - \bar{x}_{cxy}\right)\right]\right|$$

In addition to (c) used in calculating *B*1 - *B*3 discussed above, the bounds in equations H.16, H.17, and H.18 include the following data elements: (d) the means of the baseline measure for the subjects in the analytic sample with observed baseline data, separately for the intervention and comparison groups ($\bar{x}_{ixy}$, and $\bar{x}_{cxy}$,); (e) the means of the baseline measure for the subjects in the analytic sample with observed baseline and outcome data, separately for the intervention and comparison groups ($\bar{x}_{ix}$, and $\bar{x}_{cx}$,); (f) the standard deviations of the baseline measure for either the sample of subjects in the analytic sample with observed baseline data or the sample with observed baseline and outcome data, separately for the intervention and comparison groups, which are used to calculate SD_x; and (g) the number of subjects with observed baseline data in the analytic sample by condition (n_{ix} and n_{cx}).

The formulas for *B*1* - *B*3* reduce to *B*1 - *B*3 when there are no missing baseline data.

Bounding the baseline difference when the outcome is observed for all subjects in the analytic sample

It is not possible to assess baseline equivalence using observed data for the analytic sample in QEDs and high-attrition RCTs that use acceptable approaches to impute baseline data or are missing some baseline data for the analytic sample.

The WWC uses the same proxy pattern-mixture modeling approach used to address imputed outcome data to estimate the largest possible baseline difference under a set of reasonable assumptions about how the missing data are related to measured and unmeasured factors (Andridge & Little, 2011).

The baseline mean for a sample with missing or imputed baseline data can be modelled using:

[H.19] $$\bar{x}_j = \bar{x}_{jR} + \frac{1}{f_j(\rho_{cor})}\frac{SD_x}{SD_y}[\bar{y}_j - \bar{y}_{jR}]$$

where $\bar{x}_j$ and $\bar{x}_{jR}$ are the full-sample and complete case baseline means, $\bar{y}_j$ and $\bar{y}_{jR}$ are the full-sample and complete case outcome means, ρ_{cor} is the correlation between the outcome and the baseline measure, and $f_j(\rho_{cor})$ is a function given equation H.3.

The full-sample baseline effect size obtained using an imputation method based on the missing-at-random assumption can be written as the baseline effect size for the observed sample g_{xR} with an adjustment for the difference between the full-sample and complete case outcome means in the intervention and comparison groups, given by:

[H.20] $$g_{xMAR} = g_{xR} + \frac{\rho_{cor}}{SD_y}([\bar{y}_i - \bar{y}_{iR}] - [\bar{y}_c - \bar{y}_{cR}])$$

where $g_{xR} = \frac{1}{SD_y}(\bar{x}_{iR} - \bar{x}_{cR})$. The more general equation for the baseline effect size that allows for deviations from the missing-at-random assumption is:

[H.21] $$g_{xNMAR} = g_{xR} + \frac{1}{SD_y}\left(\frac{[\bar{y}_i - \bar{y}_{iR}]}{f_i(\rho_{cor})} - \frac{[\bar{y}_c - \bar{y}_{cR}]}{f_c(\rho_{cor})}\right)$$

Because $f_j(\rho_{cor})$ is bounded between ρ_{cor} and $1/\rho_{cor}$, the largest baseline effect size (in absolute value) accounting for deviations from the missing-at-random assumption is given by the maximum of the values given by the following four equations:

[H.22] $$C1 = \omega\left|g_{xR} + \frac{\rho_{cor}}{SD_y}([\bar{y}_i - \bar{y}_{iR}] - [\bar{y}_c - \bar{y}_{cR}])\right|$$

[H.23] $$C2 = \omega\left|g_{xR} + \frac{1}{SD_y\rho_{cor}}([\bar{y}_i - \bar{y}_{iR}] - [\bar{y}_c - \bar{y}_{cR}])\right|$$

[H.24] $$C3 = \omega \left| g_{xR} + \frac{1}{SD_y}\left(\rho_{cor}[\bar{y}_i - \bar{y}_{iR}] - \frac{1}{\rho_{cor}}[\bar{y}_c - \bar{y}_{cR}]\right)\right|$$

[H.25] $$C4 = \omega \left| g_{xR} + \frac{1}{SD_y}\left(\frac{1}{\rho_{cor}}[\bar{y}_i - \bar{y}_{iR}] - \rho_{cor}[\bar{y}_c - \bar{y}_{cR}]\right)\right|$$

The first of these, C1, is $|g_{xMAR}|$, the estimate of the baseline effect size when the missing-at-random assumption holds.

The bounds in equations H.22 to H.25 will be calculated using data reported in studies or obtained from authors. The equations include the following data elements: (a) the means and standard deviations of the outcome measure for the analytic sample, separately for the intervention and comparison groups ($\bar{y}_i$, $\bar{y}_c$ and the standard deviations are used to calculate the pooled within-group standard deviation SD_y); (b) the means of the outcome measure for the subjects in the analytic sample with observed baseline data, separately for the intervention and comparison groups ($\bar{y}_{iR}$, $\bar{y}_{cR}$); (c) the correlation between the baseline and the outcome measures (ρ_{cor}); and (d) an estimate of the baseline difference based on study data (g_{xR}).

Applying the bounds in equations H.22 to H.25 does not require knowing the baseline effect size using imputed baseline data. Rather, these bounds use the complete case baseline effect size. When the study imputes the baseline data using an acceptable approach and reports the baseline effect size based on imputed data, g_{xI}, a different set of bounds should be used.

Comparing g_{xMAR} and g_{xNMAR}, the bias in the imputed baseline effect size due to deviations from MAR is given by:

[H.26] $$bias = \frac{1}{SD_y}\left\{\left(\frac{1}{f_i(\rho_{cor})} - \rho_{cor}\right)[\bar{y}_i - \bar{y}_{iR}] - \left(\frac{1}{f_c(\rho_{cor})} - \rho_{cor}\right)[\bar{y}_c - \bar{y}_{cR}]\right\}$$

Adding this bias to g_{xI} gives an alternative set of bounds for the baseline effect size:

[H.27] $$D1 = \omega|g_{xI}|$$

[H.28] $$D2 = \omega \left| g_{xI} + \frac{1}{SD_y}\frac{1-{\rho_{cor}}^2}{\rho_{cor}}[\bar{y}_i - \bar{y}_{iR}]\right|$$

[H.29] $$D3 = \omega \left| g_{xI} - \frac{1}{SD_y}\frac{1-{\rho_{cor}}^2}{\rho_{cor}}[\bar{y}_c - \bar{y}_{cR}]\right|$$

[H.30] $$D4 = \omega \left| g_{xI} + \frac{1}{SD_y}\frac{1-{\rho_{cor}}^2}{\rho_{cor}}[(\bar{y}_i - \bar{y}_{iR}) - (\bar{y}_c - \bar{y}_{cR})]\right|$$

For simplicity, the bounds *C*1 - *C*4 and *D*1 - *D*4 were derived based on an imputation model based only on the relationship between the outcome and the baseline measure. If the imputation model included baseline measures in addition to the outcome, then it is acceptable but not required to replace the outcome means with

the average predicted value of the baseline measure. In this case the formula should scale by $SD_{\hat{x}}$ instead of SD_y, where $SD_{\hat{x}}$ is the standard deviation of the predicted baseline scores.

Bounding the baseline difference when the outcome measure is imputed for some subjects in the analytic sample

When an analytic sample includes both imputed outcome data and missing or imputed baseline data, it is not possible to calculate the bounds *C*1 – *C*4 or *D*1 – *D*4. This is because the means of the outcome measure are unknown for the analytic sample and are possibly unknown for the restricted sample of subjects with observed baseline data.

The full sample baseline mean for group *j* can be written as:

[H.31] $$\bar{x}_j = \bar{x}_{jxy} + \left(\frac{n_j - n_{jy}}{n_j}\right)\left[\bar{x}_{j\sim y} - \bar{x}_{jxy}\right] + \left(\frac{n_{jy}}{n_j}\right)\left(\frac{1}{f_j(\rho_{cor})}\frac{SD_x}{SD_y}\left[\bar{y}_{jy} - \bar{y}_{jxy}\right]\right)$$

where $\bar{x}_{jxy}$ is the baseline mean for the observations in the complete case analytic sample for group j and is observed at both baseline and for the collection of outcomes, $\bar{y}_{jxy}$ is the outcome mean for the same sample, and $\bar{y}_{jy}$ is the outcome mean for the sample with observed outcome data but possibly missing baseline data.

The baseline effect size obtained using an imputation method based on the MAR assumption ($g_j(\rho) = \rho$) can be written as the difference in the estimated full-sample intervention and comparison group baseline means, given by:

[H.32] $$g_{xMAR} = g_{xR(xy)} + \frac{1}{SD_x}\left(\begin{array}{l}\left\{\left(\frac{n_i - n_{iy}}{n_i}\right)\left[\bar{x}_{i\sim y} - \bar{x}_{ixy}\right] + \left(\frac{n_{iy}}{n_i}\right)\frac{\rho_{cor}SD_x}{SD_y}\left[\bar{y}_{iy} - \bar{y}_{ixy}\right]\right\} \\ -\left\{\left(\frac{n_c - n_{cy}}{n_c}\right)\left[\bar{x}_{c\sim y} - \bar{x}_{cxy}\right] + \left(\frac{n_{cy}}{n_c}\right)\frac{\rho_{cor}SD_x}{SD_y}\left[\bar{y}_{cy} - \bar{y}_{cxy}\right]\right\}\end{array}\right)$$

where $g_{xR(xy)} = \frac{1}{SD_x}\left(\bar{x}_{ixy} - \bar{x}_{cxy}\right)$.

The more general formula that allows for deviations from the missing-at-random assumption is the following:

[H.33] $$g_{xNMAR} = g_{xR(xy)} + \frac{1}{SD_x}\left(\begin{array}{l}\left\{\left(\frac{n_i - n_{iy}}{n_i}\right)\left[\bar{x}_{i\sim y} - \bar{x}_{ixy}\right] + \left(\frac{n_{iy}}{n_i}\right)\frac{SD_x}{f_i(\rho_{cor})SD_y}\left[\bar{y}_{iy} - \bar{y}_{ixy}\right]\right\} \\ -\left\{\left(\frac{n_c - n_{cy}}{n_c}\right)\left[\bar{x}_{c\sim y} - \bar{x}_{cxy}\right] + \left(\frac{n_{cy}}{n_c}\right)\frac{SD_x}{f_c(\rho_{cor})SD_y}\left[\bar{y}_{cy} - \bar{y}_{cxy}\right]\right\}\end{array}\right)$$

The largest baseline effect size (in absolute value) accounting for deviations from the missing-at-random assumption is given by the maximum of the values from the following equations:

[H.34] $$C1^* = \omega \left| g_{xR(xy)} + \left(\begin{array}{l} \left\{ \left(\frac{n_i - n_{iy}}{n_i SD_x} \right) \left[\bar{x}_{i\sim y} - \bar{x}_{ixy} \right] + \rho_{cor} \left(\frac{n_{iy}}{n_i SD_y} \right) \left[\bar{y}_{iy} - \bar{y}_{ixy} \right] \right\} - \\ \left\{ \left(\frac{n_c - n_{cy}}{n_c SD_x} \right) \left[\bar{x}_{c\sim y} - \bar{x}_{cxy} \right] + \rho_{cor} \left(\frac{n_{cy}}{n_c SD_y} \right) \left[\bar{y}_{cy} - \bar{y}_{cxy} \right] \right\} \end{array} \right) \right|$$

[H.35] $$C2^* = \omega \left| g_{xR(xy)} + \left(\begin{array}{l} \left\{ \left(\frac{n_i - n_{iy}}{n_i SD_x} \right) \left[\bar{x}_{i\sim y} - \bar{x}_{ixy} \right] + \frac{1}{\rho_{cor}} \left(\frac{n_{iy}}{n_i SD_y} \right) \left[\bar{y}_{iy} - \bar{y}_{ixy} \right] \right\} - \\ \left\{ \left(\frac{n_c - n_{cy}}{n_c SD_x} \right) \left[\bar{x}_{c\sim y} - \bar{x}_{cxy} \right] + \frac{1}{\rho_{cor}} \left(\frac{n_{cy}}{n_c SD_y} \right) \left[\bar{y}_{cy} - \bar{y}_{cxy} \right] \right\} \end{array} \right) \right|$$

[H.36] $$C3^* = \omega \left| g_{xR(xy)} + \left(\begin{array}{l} \left\{ \left(\frac{n_i - n_{iy}}{n_i SD_x} \right) \left[\bar{x}_{i\sim y} - \bar{x}_{ixy} \right] + \rho_{cor} \left(\frac{n_{iy}}{n_i SD_y} \right) \left[\bar{y}_{iy} - \bar{y}_{ixy} \right] \right\} - \\ \left\{ \left(\frac{n_c - n_{cy}}{n_c SD_x} \right) \left[\bar{x}_{c\sim y} - \bar{x}_{cxy} \right] + \frac{1}{\rho_{cor}} \left(\frac{n_{cy}}{n_c SD_y} \right) \left[\bar{y}_{cy} - \bar{y}_{cxy} \right] \right\} \end{array} \right) \right|$$

[H.37] $$C4^* = \omega \left| g_{xR(xy)} + \left(\begin{array}{l} \left\{ \left(\frac{n_i - n_{iy}}{n_i SD_x} \right) \left[\bar{x}_{i\sim y} - \bar{x}_{ixy} \right] + \frac{1}{\rho_{cor}} \left(\frac{n_{iy}}{n_i SD_y} \right) \left[\bar{y}_{iy} - \bar{y}_{ixy} \right] \right\} - \\ \left\{ \left(\frac{n_c - n_{cy}}{n_c SD_x} \right) \left[\bar{x}_{c\sim y} - \bar{x}_{cxy} \right] + \rho_{cor} \left(\frac{n_{cy}}{n_c SD_y} \right) \left[\bar{y}_{cy} - \bar{y}_{cxy} \right] \right\} \end{array} \right) \right|$$

In addition to (c) and (d) used in calculating *C*1 - *C*4, the bounds in calculating *C*1* - *C*4* include the following data elements: (e) the means of the outcome measure for the subjects in the analytic sample with observed outcome data, separately for the intervention and comparison groups ($\bar{y}_{iy}$ and $\bar{y}_{cy}$); (f) the means of the outcome measure for the subjects in the analytic sample with observed baseline and outcome data, separately for the intervention and comparison groups ($\bar{y}_{ixy}$ and $\bar{y}_{cxy}$); (g) the standard deviations of the outcome measure for either the sample of subjects in the analytic sample with observed outcome data or the sample with observed baseline and outcome data, which are used to calculate SD_y; and (h) the number of subjects with observed outcome data in the analytic sample by condition (n_i and n_c).

Applying the bounds *C*1* - *C*4* requires knowing the complete case baseline effect size, not the baseline effect size using imputed baseline data. A different set of bounds apply when the study imputes the baseline data using an acceptable approach and reports the baseline effect size based on imputed data, g_{xI}.

Comparing g_{xMAR} and g_{xNMAR}, the bias in the imputed baseline effect size due to deviations from missing-at-random assumption is given by:

[H.38] $$bias = \frac{1}{SD_y} \left\{ \left(\frac{n_{iy}}{n_i} \right) \left(\frac{1}{f_i(\rho_{cor})} - \rho_{cor} \right) \left[\bar{y}_{iy} - \bar{y}_{ixy} \right] - \left(\frac{n_{cy}}{n_c} \right) \left(\frac{1}{f_c(\rho_{cor})} - \rho_{cor} \right) \left[\bar{y}_{cy} - \bar{y}_{cxy} \right] \right\}$$

Adding this bias to g_{xI} gives an alternative set of bounds for the baseline effect size *D*1* - *D*4*:

[H.39] $$D1^* = \omega |g_{xI}|$$

[H.40] $$D2^{*} = \omega \left| g_{xI} + \frac{1}{SD_y}\left(\frac{n_{iy}}{n_i}\right)\frac{1-\rho_{cor}{}^2}{\rho_{cor}}\left[\bar{y}_{iy} - \bar{y}_{ixy}\right]\right|$$

[H.41] $$D3^{*} = \omega \left| g_{xI} - \frac{1}{SD_y}\left(\frac{n_{cy}}{n_c}\right)\frac{1-\rho_{cor}{}^2}{\rho_{cor}}\left[\bar{y}_{cy} - \bar{y}_{cxy}\right]\right|$$

[H.42] $$D4^{*} = \omega \left| g_{xI} + \frac{1}{SD_y}\frac{1-\rho_{cor}{}^2}{\rho_{cor}}\left[\left(\frac{n_{iy}}{n_i}\right)\left(\bar{y}_{iy} - \bar{y}_{ixy}\right) - \left(\frac{n_{cy}}{n_c}\right)\left(\bar{y}_{cy} - \bar{y}_{cxy}\right)\right]\right|$$

The formulas for *C*1* - *C*4* and *D*1* - *D*4* reduce to *C*1 - *C*4 and *D*1 - *D*4 when there are no missing outcome data.

APPENDIX I. STATISTICAL FORMULAS FOR THE NONOVERLAP OF ALL PAIRS IN SINGLE-CASE DESIGNS

The What Works Clearinghouse (WWC) uses the nonoverlap of all pairs to assess the presence of baseline trend and reversibility in single-case designs. The nonoverlap of all pairs is one of several quantitative nonoverlap indices designed to mimic the judgments made by visual analysis. Research has shown that these indices are broadly consistent with visual analytic judgments (Parker et al., 2014), Their integration is intended to allow for a review process that is also broadly consistent with visual analytic judgments.

Although nonoverlap measures are most often used to compare baseline phases (the A phase) to treatment phases (the B phase), in principle they can be used to calculate the degree of overlap between any two series of data points. This appendix uses the convention that the first series of data points is the A series and the second series of data points is the B series.

The nonoverlap of all pairs was first conceived as a method that could be performed by hand, in concert with visual analysis. To calculate the nonoverlap of pairs by hand, take the first data point in the A series. Compare that data point with every data point in the B series, one by one. When the data point in the B series represents an improvement over the data point in the A series, record a score of 1 for that pair of data points. When the data point in the B series is tied with the data point in the A series, record a score of 0.5 for that pair of data points. When the data point in the B series represents worse performance than data point from the A series, record a score of 0 for that pair of data points.

Continue comparing data points from the A series to all data points in the B series until you have the total set of pairwise comparisons between the A series and the B series. Sum the scores, and then divide by the total number of pairwise comparisons. A value of 0 would indicate that all of the data points in the B series indicated a worse performance than the data points in the A series, or total overlap. A value of 1 would indicate that all of the data points in the B series represent an improvement over all of the data points in the A series, or total nonoverlap.

Calculating the nonoverlap of all pairs

Although the nonoverlap of all pairs can be calculated by hand, the WWC will always prefer to calculate this metric using tabular data provided by the primary study author or tabular data extracted from the plots in the original study. The following formulas are adapted from Pustejovsky (2019). For an outcome where an increase in the outcome is desirable, the nonoverlap of all pairs *NAP* can be calculated as:

$$NAP = \frac{1}{m_A m_B} \sum_{i=1}^{m_B} \sum_{j=1}^{m_A} \left[I\left(y_j^B > y_i^A\right) + 0.5\, I\left(y_j^B = y_i^A\right) \right] \quad \text{[I.1]}$$

Here, y_i^A is the set of data points in the A series from $i = 1, \ldots, m_A$ where m_A is the number of data points in the A series. Similarly, y_j^B is the set of data points in the B series from $j = 1, \ldots, m_B$ where m_B is the number of data points in the B series. The function $I()$ is an indicator function that is equal to 1 if the logical statement in the function is true or 0 if the statement is false.

The formula is similar for an outcome where a decrease in the outcome is desirable:

[I.2] $$NAP = \frac{1}{m_A m_B} \sum_{i=1}^{m_B} \sum_{j=1}^{m_A} \left[I\left(y_j^B < y_i^A\right) + 0.5\, I\left(y_j^B = y_i^A\right) \right].$$

Use for assessing baseline trend

In the context of single-case designs, if the data points in the baseline are trending in the expected direction for the intervention effect, this trend is a potential indicator that there is an issue with the design and a possible threat to internal validity. For study to be eligible to receive a rating of *Meets WWC Standards Without Reservations*, reviewers must ensure that there is not excessive trend in the baseline phase using the nonoverlap of all pairs. Studies with excessive baseline trend are still eligible to receive a rating of *Meets WWC Standards With Reservations*. Without evidence in the text of the study regarding some confound in the design, the presence of baseline trend alone is not enough to cause a design to receive a rating of *Does Not Meet WWC Standards*. For the purposes of this requirement, a nonoverlap of all pairs for baseline trend of 0.85 or lower is evidence that there is a lack of concerning baseline trend.[48]

When using the nonoverlap of all pairs for assessing baseline trend, reviewers will compare every data point in the baseline except the last three to the last three data points in the phase, in the direction of the expected effect of the intervention. In other words, if there are m_A observations in the baseline phase, reviewers will compare data points 1, …, m_A - 3 to data points m_A - 2, m_A - 1, and m_A in the direction of the expected effect. Equation I.1 can be rewritten as:

[I.3] $$NAP = \frac{1}{(m_A - 3)3} \sum_{i=i}^{(m_A-3)} \sum_{j=(m_A-2)}^{m_A} \left[I\left(y_j^A > y_i^A\right) + 0.5\, I\left(y_j^A = y_i^A\right) \right].$$

Equation I.2 can similarly be rewritten as:

[I.4] $$NAP = \frac{1}{(m_A - 3)3} \sum_{i=i}^{(m_A-3)} \sum_{j=(m_A-2)}^{m_A} \left[I\left(y_j^A < y_i^A\right) + 0.5\, I\left(y_j^A = y_i^A\right) \right].$$

Use for assessing reversibility

Single-case design with reversals or return-to-baseline phase(s) are most appropriate when the outcomes and interventions allow for the pattern of responding to return to the patterns observed in the initial baseline phase when the intervention is removed. If the intervention is likely to cause a permanent or at least a reasonably persistent change in the outcome of interest, designs with reversals are less appropriate. The observed intervention effect will likely be attenuated as a result. For a study to receive a rating of *Meets WWC Standards Without Reservations*, reviewers must ensure that there is evidence of at least minimal reversibility. Minimal reversibility for the WWC is defined as a nonoverlap of all pairs of 0.85 when comparing the initial baseline phase (the A data series in this case) separately to each reversal or return to baseline phase (the B data series) in the direction of the expected intervention effect.

[48] A small working group of methodologists and applied single-case design researchers arrived at this critical threshold by examining potential rank-ordered patterns of responding when there are six total data points in the baseline, the minimum required to receive a rating of *Meets WWC Standards Without Reservations*.

A single-case design with reversals that contains a phase with a nonoverlap of all pairs of greater than 0.85 with respect to the baseline phase may receive at best a rating of *Meets WWC Standards With Reservations*.[49]

Nonoverlap of all pairs examples

Consider the example of Figure 18 from Chapter VI. This is a treatment reversal design intended to examine the effectiveness of an intervention for reducing externalizing behaviors, so the baseline trend and the reversibility requirement need to be satisfied for this design to be eligible to receive a rating of *Meets WWC Standards Without Reservations*. In both cases, an improvement is a reduction in the observed externalizing behaviors.

Baseline trend example

To calculate the nonoverlap of all pairs for baseline trend, consider only the first six data points which make up the initial baseline. These data points are 15, 10, 14, 17, 13, and 12. Compare each of the first three data points to each of the last three observations to calculation the nonoverlap of all pairs. When a data point in the in the last three observations is smaller than an observation in the first three data points, give that pair a score of 1. When a data point in the last three data points is equal to an observation in the first three data points, give that pair a score of 0.5. When a data point in the last three data points is greater than a data point in the first three observations, give that pair a score of 0. Table I.1 contains results of these pairwise comparisons.

Table I.1. *Example of pairwise comparisons for baseline trend*

Initial baseline	Last three data points	Score
15	17	0
15	13	1
15	12	1
10	17	0
10	13	0
10	12	0
14	17	0
14	13	1
14	12	1

When all pairwise comparisons are complete, add up the results each of the pairwise comparisons and divide by the total number of comparisons, as demonstrated below.

$$NAP = \frac{0+1+1+0+0+0+0+1+1}{9} = .4\overline{4}$$

The result is a nonoverlap of pairs less than 0.85. Therefore, the experiment would be eligible to receive a rating of *Meets WWC Standards Without Reservations*.

[49] A small working group of methodologists and applied single-case design researchers arrived at this critical threshold by examining potential patterns of responding when there are six data points in the baseline phase and five data points in the subsequent reversal phase, the minimum number of data points required to receive a rating of *Meets WWC Standards Without Reservations*.

Reversibility example

To use the nonoverlap of all pairs to assess reversibility, compare the first baseline phase with data points 15, 10, 14, 17, 13, 12 to the second treatment phase with data points 9, 9, 11, 15, 20. As with the baseline trend example, compare each data point in the first baseline phase to each data point in the second baseline phase, scoring each pair in the same fashion. Table I.2 contains the results of these pairwise comparisons.

Table I.2. *Example of pairwise comparisons for reversibility*

First baseline phase	Second baseline phase	Score
15	9	1
15	9	1
15	11	1
15	15	.5
15	20	0
10	9	1
10	9	1
10	11	0
10	15	0
10	20	0
14	9	1
14	9	1
14	11	1
14	15	0
14	20	0
17	9	1
17	9	1
17	11	1
17	15	1
17	20	0
13	9	1
13	9	1
13	11	1
13	15	0
13	20	0
12	9	1
12	9	1
12	11	1
12	15	0
12	20	0

When all pairwise comparisons are complete, add up the results each of the pairwise comparisons and divide by the total number of comparisons, as demonstrated below.

$$NAP = \frac{\begin{matrix}1+1+1+.5+0+1+1+0+0+0+1+1+1+0+0+\\ 1+1+1+1+0+1+1+1+0+0+1+1+1+0+0\end{matrix}}{30} = .61\overline{6}$$

The result is a nonoverlap of pairs less than 0.85. Therefore, the experiment would be eligible to receive a rating of *Meets WWC Standards Without Reservations.*

REFERENCES

Allison, P. D. (2002). *Missing data* (Paper No. 136). Sage University.

Anderson, T. W., & Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, *20*(1), 46–63.

Andridge, R. R., & Little, R. J. A. (2011). Proxy pattern-mixture analysis for survey nonresponse. *Journal of Official Statistics, 27*(2), 153–180.

Angrist, J. D., Imbens, G. W., & Rubin, D. B. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–455.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: What is it and how does it work? *International Journal of Methods in Psychiatric Research, 20*(1), 40–49.

Basmann, R. L. (1974). Exact finite sample distributions for some econometric estimators and test statistics: A survey and appraisal. In M. D. Intrilligator & D. A. Kendrick (Eds.), *Frontiers of quantitative economics, Vol. 2* (pp. 209–288). North-Holland Publishing Co.

Berk, R. A. (1979). Generalizability of behavioral observations: A clarification of interobserver agreement and interobserver reliability. *American Journal of Mental Deficiency, 83*(5), 460–472.

Bloom, H. S. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, *8*(2), 225–246.

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). Russell Sage Foundation.

Bloom, H., Zhu, P., & Unlu, F. (2010). *Finite sample bias from instrumental variables analysis in randomized trials*. MDRC.

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

Bound, J., Jaeger, D. A., & Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, *90*(430), 443–450.

Briesch, A. M., Hemphill, E. M., Volpe, R. J., & Daniels, B. (2015). An evaluation of observational methods for measuring response to classwide intervention. *School Psychology Quarterly, 30*(1), 37–49. https://eric.ed.gov/?id=EJ1056679

Buse, A. (1992). The bias of instrumental variable estimators. *Econometrica*, *60*(1), 173-180.

Calonico, S., Cattaneo, M. D., & Titiunik, R. (2014). Robust nonparametric confidence intervals for regression discontinuity designs. *Econometrica, 82*(6), 2295–2326.

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

Cragg, J. G., & Donald, S. D. (1993). Testing identifiability and specification in instrumental variables models. *Econometric Theory, 9*(2), 222–240.

Daniel, R., Zhang, J., & Farewell, D. (2020). Making apples from oranges: Comparing noncollapsible effect estimators and their standard errors after adjustment for different covariate sets. *Biometrical Journal*, *63*(3), 528–557.

Deke, J. (2014). *Using the linear probability model to estimate impacts on binary outcomes in randomized controlled trials*. Office of Adolescent Health. https://opa.hhs.gov/sites/default/files/2020-07/lpm-tabrief.pdf

Deke, J., & Chiang, H. (2017). The WWC attrition standard: Sensitivity to assumptions and opportunities for refining and adapting to new contexts. *Evaluation Review, 41*(2), 130–154. https://doi.org/10.1177/0193841X16670047

Efthimiou, O. (2018). Practical guide to the meta-analysis of rare events. *Evidence Based Mental Health*, *21*(2), 72–76.

Fier, D., Lemieux, T., & Marmer, V. (2016). Weak identification in fuzzy regression discontinuity designs. *Journal of Business and Economic Statistics, 34*(2), 185–196.

Gelman, A., & Pardoe, I. (2007). 2. Average predictive comparisons for models with nonlinearity, interactions, and variance components. *Sociological Methodology*, *37*(1), 23–51.

Goldstein, H., Browne, W., & Rasbash, J. (2002). Partitioning variation in multilevel models. *Understanding Statistics*, *1*(4), 223–231.

Gomila, R. (2021). Logistic or linear? Estimating causal effects of experimental treatments on binary outcomes using regression analysis. *Journal of Experimental Psychology: General*, *150*(4), 700–709.

Grames, E., Stillman, A., Tingley, M., & Elphick, C. (2019). An automated approach to identifying search terms for systematic reviews using keyword co-occurrence networks. *Methods in Ecology and Evolution, 10*(10), 1645–1654.

Greene, W. (2000). *Econometric analysis* (4th ed.). Prentice Hall.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, *50*, 1029–1054.

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

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics, 6*(2), 107–128. https://eric.ed.gov/?id=EJ248027

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341-370. https://eric.ed.gov/?id=EJ782531

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60-87. https://eric.ed.gov/?id=EJ782420

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

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

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. A. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3*, 224-239. https://eric.ed.gov/?id=ED543975

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. A. (2013). A standardized mean difference effect size for multiple baseline designs. *Research Synthesis Methods, 4,* 324-341. https://eric.ed.gov/?id=EJ1109014

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

Hellevik, O. (2007). Linear versus logistic regression when the dependent variable is a dichotomy. *Quality & Quantity*, *43*(1), 59-74.

Horner, R., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Expanding analysis and use of single-case research. *Education and Treatment of Children, 35*, 269-290.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica, 62*, 467-475.

Imbens, G. W., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies, 79*(3), 933-959.

Imbens, G. W., & Lemieux. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics, 14(2),* 615-635.

Imbens, G. W., & Rubin, D. B. (1997a). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics*, *25*, 305-327.

Imbens, G. W., & Rubin, D. B. (1997b). Estimating outcome distributions for compliers in instrumental variables models. *Review of Economic Studies*, *64*(4), 555-574.

Jones, M. P. (1996). Indicator and stratification methods for missing explanatory variables in multiple linear regression. *Journal of the American Statistical Association, 91*(433), 222-230.

Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A. M. K., Hammerstrøm, K., & Sathe, N. (2017). *Searching for studies: A guide to information retrieval for Campbell systematic reviews*. The Campbell Collaboration.

Lee, D., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics, 142*(2), 655–674.

Little, R. J., D'Agostino, R., Cohen, M. L., Dickersin, K., Emerson, S. S., Farrar, J. T., Frangakis, C., Hogan, J. W., Molenberghs, G., Murphy, S. A., Neaton, J. D., Rotnitzky, A., Scharfstein, D., Shih, W. J., Siegel, J. P., & Stern, H. (2012). The prevention and treatment of missing data in clinical trials. *The New England Journal of Medicine, 367*(14), 1355–1360.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics, 142*(2), 698–714.

McGowan, J., Sampson, M., Salzwedel, D. M., Cogo, E., Foerster, V., & Lefebvre, C. (2016). PRESS peer review of electronic search strategies: 2015 guideline statement. *Journal of Clinical Epidemiology, 75*, 40–46.

McNeish, D., Stapleton, L. M., & Silverman, R. D. (2017). On the unnecessary ubiquity of hierarchical linear modeling. *Psychological Methods, 22*(1), 114–140.

Morgan, K. L., & Rubin, D. B. (2012). Rerandomization to improve covariate balance in experiments. *The Annals of Statistics, 40*(2), 1263–1282.

Nelson, C., & Startz, R. (1990). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica, 58*(4), 967–976.

Norton, E. C., & Dowd, B. E. (2018). Log odds and the interpretation of logit models. *Health Services Research, 53*(2), 859–878.

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy, 40*(4), 357–367.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2014). Non-overlap analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 127–151). American Psychological Association.

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research, 74*(4), 525–556. https://eric.ed.gov/?id=EJ737272

Polanin, J. R., Tanner-Smith, E. E., & Hennessy, E. A. (2016). Estimating the difference between published and unpublished effect sizes: A meta-review. *Review of Educational Research, 86*(1), 207–236. https://eric.ed.gov/?id=EJ1090515

Puma, M. J., Olsen, R. B., Bell, S. H., & Price, C. (2009). *What to do when data are missing in group randomized controlled trials* (NCEE 2009-0049). National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education. https://eric.ed.gov/?id=ED511781

Pustejovsky, J. E. (2019). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods*, *24*(2), 217–235. https://eric.ed.gov/?id=ED581547

Pustejovsky, J. E., Chen, M., & Hamilton, B. (2021). scdhlm*: A web-based calculator for between-case standardized mean differences* (Version 0.5.2) [Web application]. https://jepusto.shinyapps.io/scdhlm

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. L. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics, 39*, 368–393. https://eric.ed.gov/?id=EJ1041686

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

Reardon, S., & Robinson, J. P. (2012). Regression discontinuity designs with multiple rating-score variables. *Journal of Research on Educational Effectiveness, 5*(1), 83–104.

Rethlefsen, M. L., Kirtley, S., Waffenschmidt, S., Ayala, A. P., Moher, D., Page, M. J., Koffel, J. B., & PRISMA-S Group. (2021). PRISMA-S: An extension to the PRISMA Statement for Reporting Literature Searches in Systematic Reviews**.** *Systematic Reviews, 10*, article 39.

Richardson, D. H. (1968). The exact distribution of a structural coefficient estimator. *Journal of the American Statistical Association*, *63*(324), 1214–1226.

Robinson, L. D., & Jewell, N. P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review / Revue Internationale de Statistique*, *59*(2), 227.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* Wiley.

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.

Sanderson, E., & Windmeijer, F. (2016). A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of Econometrics, 190*(2), 212–221.

Sawa, T. (1969). The exact sampling distribution of ordinary least squares and two-stage least squares estimators. *Journal of the American Statistical Association*, *64*(325), 923–937.

Schafer, J. L. (1999). Multiple imputation: A primer. *Statistical Methods in Medical Research, 8*(1), 3–15.

Schochet P. Z. (2008). Statistical power for random assignment evaluations of education programs. *Journal of Educational and Behavioral Statistics, 33*(1), 62–87. https://eric.ed.gov/?id=EJ788461

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs*. Houghton Mifflin.

Slocum, T. A., Pinkelman, S. E., Joslyn, P. R., & Nichols, B. (2022). Threats to internal validity in multiple-baseline design variations. *Perspectives on Behavior Science*, 1–20.

Staiger, D., & Stock, J. H. (1997). Instrumental variables regression with weak instruments. *Econometrica, 64*(3), 557–586.

Stock, J., Wright, J., & Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *Journal of Business and Economic Statistics, 20*(4), 518–529.

Stock, J., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In J. Stock & D. W. K. Andrews (Eds.), *Identification and inference for econometric models: Essays in Honor of Thomas J. Rothenberg* (pp. 80–108). Cambridge University Press.

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

Thompson, C. G., & Becker, B. J. (2014). The impact of multiple endpoint dependency on *Q* and I^2 in meta-analysis. *Research Synthesis Methods*, 5, 235–253. https://eric.ed.gov/?id=EJ1109039

Weber, F., Knapp, G., Ickstadt, K., Kundt, G., & Glass, Ä. (2020). Zero-cell corrections in random-effects meta-analyses. *Research Synthesis Methods*, *11*(6), 913–919. https://eric.ed.gov/?id=EJ1274816

Wolf, A. Price, C. Miller, H., & Boulay, B. (2017). *Establishing baseline equivalence: A practical guide for evaluators.* U.S. Department of Education, Institute of Education Sciences.

Wong, V., Steiner, P., & Cook, T. (2013). Analyzing regression-discontinuity designs with multiple assignment variables: A comparative study of four estimation methods. *Journal of Educational and Behavioral Statistics, 38*(2), 107–141.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. MIT Press.