What Works Clearinghouse™

Reviewer Guidance for Use With the
Procedures and Standards Handbook (version 3.0)
Updated December 9, 2016

CONTENTS

I WHAT WORKS CLEARINGHOUSE REVIEWER GUIDANCE 1

A. Introduction 2

II GUIDANCE FOR REVIEWS UNDER THE GROUP DESIGN STANDARDS 3

A. Study Definition 4

B. Cluster Designs 6

C. Groups Defined by Cohort 8

D. Compromised Random Assignment 9

E. Missing Data 11

F. Analyses with Potentially Endogenous Covariates 12

G. Non-Replicability of Study-Reported Effect Sizes 13

H. Non-Reporting of Required Adjusted Impacts 14

I. Posttest Reliability 15

J. Multiple Comparisons 16

K. Analyses of Gain Scores 17

L. Pretest Reliability 18

M. Analyses using Propensity Scores 19

N. Confounding Factors 20

III. GUIDANCE FOR REVIEWS UNDER THE SINGLE-CASE DESIGN STANDARDS 23

A. Confounding Factors in Single-Case Designs 24

B. Inter-Assessor Agreement Reporting 26

C. Changing Criterion Designs 24

D. Alternating Treatment Designs 31

E. Reversal-Withdrawal, Multiple Baseline, and Multiple Probe Designs with More Than Two Conditions 36

F. Reversal-Withdrawal, Multiple Baseline, and Multiple Probe Designs with More Than the Minimum Number of Required Phases 39

G. Timing of Sessions in Multiple Baseline and Multiple Probe Designs 43

H. Concurrence in Multiple Baseline and Multiple Probe Designs 46

IV. GUIDANCE FOR REVIEWS OF STUDIES THAT PRESENT A COMPLIER AVERAGE CAUSAL EFFECT 48

I. What Works Clearinghouse Reviewer Guidance

A. Introduction

The *What Works Clearinghouse (WWC) Procedures and Standards Handbook (version 3.0)* provides reviewers with a detailed description of the procedures and standards used by the WWC in the review of studies and production of reports. While the *Handbook* addresses most situations, reviewers occasionally need additional guidance. The WWC has produced a series of guidance documents for reviewers to provide clarification and interpretation of standards and support consistency across reviews. These guidance documents do not articulate any changes to the procedures and standards in the *version 3.0 Handbook*. Rather, the guidance documents clarify how the standards should be implemented in situations where the current *Handbook* is not sufficiently specific to ensure consistent reviews.

These guidance documents are provided here to inform WWC readers on the additional guidance provided to reviewers applying version 3.0 procedures and standards. Guidance can change based on future WWC Statistical, Technical, and Analysis Team work and feedback from the Institute of Education Sciences and others. Revised guidance will be shared with WWC reviewers and will be updated on the WWC website. The guidance in this document was last updated on December 9, 2016. Substantive updates include additional guidance on reviewing studies that use single-case designs.

II. GUIDANCE FOR REVIEWS UNDER THE GROUP DESIGN STANDARDS

A. Study Definition

The number of studies affects the WWC's reporting and summarization of evidence. The level of evidence in practice guides is affected by the number of studies, and the topic area summarization of intervention findings depends on the specific number of studies (for example, for intervention findings to be characterized as *positive* rather than *potentially positive,* two studies must have demonstrated positive findings).

The *WWC Procedures and Standards Handbook (version 3.0)* defines a study as "the examination of the effect of an intervention on a particular sample (e.g., a set of students, schools, or districts) and set of outcomes" (p. 7). A study is not the same as a manuscript, journal article, book chapter, or report. Individual journal articles or reports can include multiple studies (for example, many psychology articles include several experiments). Similarly, a single study can be described in multiple reports (for example, a 5-year study of an intervention will often release interim reports annually).

Although the *Handbook* defines a study, it does not provide guidance on exactly how to identify a study. For example, in a study that analyzed 20 subgroups, each subgroup analyzed could be considered a study according to a literal interpretation of the *Handbook* definition. Similarly, a randomized controlled trial (RCT) that randomly assigned students to condition separately within 10 sites and reported site-specific impacts would be considered 10 studies, even if the design and analysis were conducted by one research team and reported as one impact.

Guidance

For RCTs, to ensure that sampling errors are independent, a study is defined by a randomization conducted by a single research team. So that WWC-reported results reflect study-reported findings, if a single research team reports pooled impacts combining distinct randomized samples, that will be considered one study. The reviewer should classify a sample (and corresponding analysis) as one study if the intervention and comparison groups were formed in either one randomization process or multiple processes reported as pooled findings by one team. The reviewer should classify a sample (and corresponding analysis) as *n* studies if the intervention and comparison groups were formed in *n* randomization processes and not combined (for example, if random assignment occurred across multiple research teams).

For a quasi-experimental design (QED), to ensure that sampling errors are independent or largely independent, a study is defined by unique intervention and comparison groups that do not overlap with other known intervention and comparison groups.[1] So that WWC-reported results reflect study-reported findings, if a single research team reports pooled impacts combining unique QED samples, that will be considered one study. The reviewer should classify an intervention/comparison analysis as a different study from another intervention/comparison analysis only if the samples do not overlap and reported impacts are not combined. For example, if an intervention group from a QED study is compared to a different comparison group, this

[1] If the overlap is trivial, the review team leadership can determine that the comparisons form separate studies.

second contrast will not be considered a new study because the intervention groups in both sets of contrasts overlap with each other.

B. Cluster Designs

The *WWC Procedures and Standards Handbook (version 3.0*) includes standards for designs in which clusters (grouping units such as classrooms or schools) rather than individuals are the unit of assignment. The analytic samples for these studies may include two types of individuals: *stayers,* who were in the sample when the clusters were assigned to condition; and *joiners,* who entered the sample afterwards. For example, schools may be assigned to condition in the fall, and the analytic sample might include new students who transferred into the schools later in the school year. The choice of individuals to include in the analysis sample has implications for the inferences that can be drawn from the findings.

Cluster Randomized Control Trials

When clusters are randomly assigned, the WWC considers the initial design an RCT regardless of whether individuals are randomly assigned to clusters.

An analysis of *stayers*, focusing on the effect on individuals in clusters at the time of assignment, can *Meet WWC Group Design Standards Without Reservations*. For these analyses, attrition is assessed at both the cluster level and at the individual level (using only clusters remaining in the sample).

For an analysis of *stayers* and *joiners*, the rating depends on the type of inference described in the study, particularly in the discussion of findings. To understand the inferences the authors are making, read how they discuss the findings. If they are talking about improvement in school scores, then it is a cluster-level inference. If they are talking about increases in student scores, then it is an individual-level inference.

- If the study is making cluster-level inferences, the study can *Meet WWC Group Design Standards Without Reservations*. The impact estimates reflect the combination of the effect on individuals who were in the clusters at the time of assignment and the effect on the composition of individuals within the clusters over time, which is noted in our reporting. For cluster-level inferences, attrition is assessed only at the cluster level.

- If the study is making individual-level inferences, the study can only *Meet WWC Group Design Standards With Reservations.* The inclusion of *joiners* not in the clusters at the time of random assignment means that the study must demonstrate the equivalence of the analytic sample (including both *stayers* and *joiners*) at baseline.

If the study is not clear or consistent about the level of inference from the analysis, then assume the study is attempting to make an individual-level inference.

Cluster Designs That Need to Demonstrate Equivalence (High-Attrition RCTs and QEDs)

If the study is making individual-level inferences, equivalence must be assessed for the analytic sample using individual-level standard deviations (not cluster-level).

If the study is making cluster-level inferences, equivalence can be assessed using either cluster-level or individual-level standard deviations. Additionally, the study can demonstrate equivalence using either: (1) the same cohort from the previous year or (2) an earlier, adjacent cohort measured at the same grade as the cohort used in the impact analysis.

C. Groups Defined by Cohort

A design in which groups are defined by cohort is often labeled a *successive-cohort design* or *cohort design*. As an example, the comparison group consists of a cohort of third graders in year Y, and the intervention group consists of another cohort of third graders in year Y+1. Usually both cohorts are observed in one school or the same set of schools.

WWC standards indicate that a confounding factor exists when "a component of the study design or the circumstances under which the intervention was implemented are perfectly aligned… with either the intervention or comparison group" (*WWC Procedures and Standards Handbook version 3.0*, p. 19). The confounding factor must plausibly influence outcomes. 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 occur between those time periods. These changes (e.g., new district policies, new personnel, or new state tests) could plausibly affect outcomes. Because many of the changes that occur over time are likely to be unobserved or not reported, the WWC cannot assess how problematic the potential changes are in individual studies.

Guidance

In group design studies, outcomes must be measured at about the same time for intervention and comparison units. When intervention and comparison groups are composed of different cohorts, the reviewer should classify time as a confounding factor and rate the study *Does Not Meet WWC Group Design Standards*.

D. Compromised Random Assignment

The *WWC Procedures and Standards Handbook (version 3.0)* notes that "the distinguishing characteristic of a randomized controlled trial is that study participants are assigned randomly to form two or more groups that are differentiated by whether or not they receive the intervention under study. Thus, at the time the sample is identified (and before the intervention), the groups should be similar, on average, on both observable and unobservable characteristics." (p. 9).

The internal validity of a RCT can be threatened by events that undermine the similarity of the groups, such as loss of sample after random assignment (attrition) and the non-random movement of sample members after random assignment that changes how they are classified in the analytic sample. Both threats have the potential to lead to bias when measuring the effect of an intervention, because they can lead to differences in outcomes between the intervention and comparison groups that are not caused by the intervention.

The *Handbook* provides extensive detail on attrition, including quantitative values to determine whether a study has high or low attrition. If attrition is high, the study can no longer receive the highest rating, and must establish equivalence to receive a rating of *Meets WWC Group Design Standards With Reservations.* This section provides guidance for determining whether an RCT has been compromised, and if so, how to deal with it.

Guidance

Random assignment can be compromised when one or both of the following occur: (1) group status is changed for the analysis or (2) investigator manipulation of the analytic sample. If random assignment is compromised due to any individuals affected by either of these reasons, the study must demonstrate equivalence of the analytic sample to receive a rating of *Meets WWC Group Design Standards With Reservations.*

Changes in group status

When one or more units are analyzed based on the condition that they ultimately received, rather than the groups to which they were originally assigned, the random assignment is compromised. For example, a study might include individuals assigned to the intervention condition who did not actually receive the intervention. This RCT would be compromised if the study includes these individuals as members of the comparison group. However, if they were analyzed based on their original assignment (often called an intent-to-treat estimate), the integrity of random assignment would be maintained.

In addition, if outcomes were missing for all contaminated and crossover units, the random assignment would not be compromised; this issue would be addressed through the attrition standard.

Investigator manipulation

Random assignment is also compromised when an investigator deliberately includes or excludes units based on a measure that was determined after randomization that could have been affected by the intervention. For example, random assignment would be compromised if an

investigator excluded students from the analytic sample who missed a certain number of days of school during the intervention period. Rather than being considered as attrition, deliberate manipulation of the analytic sample in this way compromises the random assignment.

E. Missing Data

The *WWC Standards and Procedures Handbook (version 3.0)* provides a list of acceptable statistical methods to account for missing data, focusing primarily on missing outcomes in low-attrition RCTs (pp. 18–19). While these methods can be used for missing baseline variables, equivalence cannot be demonstrated using imputed data. This section provides more guidance on missing data approaches for low-attrition RCTs and other study designs.

Guidance

Clarifications on acceptable methods of addressing missing data

Nonresponse weights can be used to address missing outcome data, but not to address missing covariate data.

Imputing a constant value to a covariate and including an indicator to signify the missing covariate is acceptable for covariates but not for outcomes.

Missing or imputed data in low-attrition RCTs

Because low-attrition RCTs do not need to demonstrate equivalence to meet standards, missing or imputed data for baseline measures or outcomes do not affect the rating, so long as the researchers used acceptable methods to address the missing data.

Establishing equivalence when some data are missing or imputed

Because imputation can be a valid method of accounting for missing data, a study that needs to demonstrate equivalence can still *Meet WWC Group Design Standards With Reservations* even when the study only presents findings that include some imputed data. In these cases, if the study mentions an unimputed analysis (for example, "impacts are the same when estimated using unimputed data"), the reviewer should conduct an author query, requesting impact estimates using actual data and baseline means/standard deviations (to establish equivalence) for only those observations with actual outcomes and baseline statistics. If the study does not mention an unimputed analysis, the reviewer should conduct an author query, asking if an analysis had been conducted using actual (non-imputed) data. If the authors did not conduct an analysis using actual data, then the query ends (and the study is rated *Does Not WWC Meet Group Design Standards*). If the authors did conduct analyses with actual data, then the query should request impact estimates using actual data and baseline means/standard deviations for only those observations with actual outcomes and baseline data.

F. Analyses with Potentially Endogenous Covariates

When reviewing a study for the WWC, reviewers should examine model specifications and descriptions of analytic procedures to ensure that the estimates of intervention effects from study-reported analyses are credible, given the proposed analytic procedure. In some impact evaluations, researchers will estimate regression models (e.g., Ordinary Least-Squares or Hierarchical Linear Modeling) where the outcome of interest is regressed on an indicator for the intervention and a series of covariates.

Guidance

WWC reviewers should carefully examine whether studies include covariates in their impact analyses that were assessed or obtained after baseline. In this situation, if these variables were influenced by the intervention, their inclusion in an impact analysis as covariates will produce a biased test of the effect of an intervention. Note: If a covariate is obtained after baseline, and the variable is generally considered time invariant (e.g., demographics such as gender, race, etc.), then there is not a concern that the variable has been influenced by the intervention.

For example, a study that examines the impact of an intervention on student achievement outcomes may collect data on factors such as (1) student attendance during the intervention, or (2) quality of student–teacher interactions. These variables associated with intervention dose, quality, or fidelity may have been affected by the intervention, which means that there would be a correlation between these variables and the intervention indicator that is caused by the intervention. If the impact analysis of this study examines student achievement as the dependent variable in a regression equation, and includes either (1) student attendance during the intervention, and/or (2) quality of student–teacher interactions is included in the regression model along with the intervention indicator variable, then the correlation between the intervention indicator and these variables will produce bias in the magnitude and standard error of the intervention beta coefficient. The WWC cannot use the results of the regression model as a credible source of information about program effects. The WWC can use alternate model specifications reported in the study that do not include these endogenous covariates, or can request unadjusted means and standard deviations from the authors. If the only analysis available to the reviewer is not credible, then the study *Does Not Meet WWC Group Design Standards* because there is no finding that meets standards.

If reviewers have questions about whether a variable included in a regression model is endogenous, they should bring their questions to the review team leadership. In addition, a lead methodologist or content expert can indicate when they believe that a covariate is unlikely to be influenced by an intervention, and that the impact estimate from a regression model can be used by the WWC as a credible test of a program's effectiveness.

G. Non-Replicability of Study-Reported Effect Sizes

The WWC reports the magnitude of study findings using Hedges' *g* and the improvement index to characterize the magnitude of the effect at a point in time. Most studies do not provide these measures, and the WWC's Study Review Guide (SRG) calculates them using the appropriate formula provided in the *WWC Procedures and Standards Handbook (version 3.0)* (Appendix F). When the study reports an effect size, the reviewer should list the reported effect size in the SRG Data tab. The SRG will also automatically calculate Hedges' *g* and the improvement index using the other information provided by the reviewer. However, the author-reported and SRG-calculated effect sizes might differ for several reasons.

Guidance

When the author-reported and SRG-calculated effect sizes differ, the reviewer should attempt to identify the source of the difference.

- If the author-reported effect size does not adjust for baseline differences, and the SRG-calculated effect size incorporates the difference-in-differences adjustment,[2] the reviewer should use the SRG-calculated effect size, because the WWC prefers adjusted findings (*Handbook*, Appendix F). The reviewer should explain the source of the discrepancy in the Main sheet in Row 50 (Did the authors present effect sizes? If so, how were they computed?).
- If the author-reported effect size uses a different formula (for example, the comparison standard deviation rather than the pooled standard deviation), then the reviewer should use the SRG-calculated effect size.
- If the author-reported and SRG-calculated effect sizes are both based on adjusted means (or, in the case of an RCT with low-attrition that does not report pretests, both effect sizes are based on unadjusted means), the reviewer should re-enter the appropriate statistics from the study in a new SRG to ensure that there is no problem with data entry and the reviewer has not inadvertently changed the SRG formulas.
- If the SRG-calculated effect size is based on the correct data and calculated correctly, the reviewer should report the SRG-calculated effect size. The reviewer should also explain the source of the discrepancy in the Main sheet in Row 50.

[2] If the study reports adjusted means or an adjusted impact estimate (β or γ), the SRG will calculate Hedges' *g* using the adjusted means or estimate, and the SRG should not perform a difference-in-differences adjustment.

H. Non-Reporting of Required Adjusted Impacts

The *WWC Procedures and Standards Handbook (version 3.0)* specifies (p. 15) that "For differences in baseline characteristics that are between 0.05 and 0.25 standard deviations, the analysis must include a statistical adjustment for the baseline characteristics to meet the baseline equivalence requirement." (p. 15). If the analysis for an outcome does not include the required covariate, the analysis for that outcome *Does Not Meet WWC Group Design Standards.* This section provides guidance on how to deal with non-reported impact estimates in studies that require statistical adjustments.

Guidance

When a study requires an adjustment, and the authors do not report means or an impact that captures this adjustment (for example, adjusted means or a coefficient from a regression model that adjusts for the baseline variables), the reviewer should conduct an author query and request the adjusted means or impacts.

If the authors do not provide the adjusted means or impacts, but the study text reports the direction and appropriate statistical significance of the impact (after adjusting for the baseline difference), then the reviewer should rate the study *Meets WWC Group Design Standards With Reservations* and report the statistical significance of the effect, but not the magnitude. If the authors do not provide the adjusted means or impacts, and the study does not report the direction and appropriate statistical significance of the impact (after adjusting for the baseline difference), then the reviewer should rate the study *Does Not Meet WWC Group Design Standards* because there is no finding that meets standards. Review team leadership should talk to the product lead about how to document the rating in the product.

I. Posttest Reliability

According to the *WWC Procedures and Standards Handbook (version 3.0)* (pp. 16–17), outcomes must demonstrate reliability "by meeting the following minimum standards: (a) internal consistency (such as Cronbach's alpha) of 0.50 or higher; (b) temporal stability/test-retest reliability of 0.40 or higher; or (c) inter-rater reliability (such as percentage agreement, correlation, or kappa) of 0.50 or higher." The *Handbook* notes that *standardized tests* do not need to explicitly meet reliability requirements—they are assumed to be reliable—and that the review protocol can require higher standards for assessing reliability and/or "stipulate how to deal with outcomes related to achievement that are unlikely to provide reliability information."

The reliability requirements aim to set standards for maximum allowable random measurement error (higher reliability indicates lower measurement error). Although this random error does not create bias, the error reduces precision and the likelihood of detecting an impact if one actually exists.

The different measures of reliability capture measurement error from different sources. Internal consistency and test-retest reliability can capture measurement error that results from poor question wording, for example, while inter-rater reliability can capture measure error that results from coder judgment. In addition, a measure's reliability can vary depending on the sample used to estimate reliability.

Guidance

An outcome's reliability can be demonstrated either by (a) meeting any of the *Handbook* or review protocol minimum standards for internal consistency, temporal stability/test-retest reliability, or inter-rater reliability; or (b) having the content expert or lead methodologist determine that that responses can be scored by a single coder with low error. For example, without quantitatively meeting one of the three reliability standards listed in the *Handbook,* an outcome may still be deemed reliable if the content expert or lead methodologist for a review determine that that responses can be scored by a single coder with low error (e.g., a multiple choice test or counts of words spelled correctly). These reliability standards do not apply to outcomes that are a behavior measured administratively (for example, graduation, enrollment in school, or grade retention). Established subscales from standardized tests also do not need to demonstrate reliability, but a non-established subscale composed of items from a standardized test must.

When a study does not report reliability statistics for an outcome, the reviewer should search the internet to see if reliability information is available; if reliability statistics can be obtained, contact review team leadership and ask if out-of-sample reliability statistics are appropriate. If appropriate reliability statistics are not available, the reviewer should conduct an author query to request reliability statistics from the study authors, or check with review team leadership about whether responses can be scored by a single coder with a high degree of reliability.

J. Multiple Comparisons

As described in the *Procedures and Standards Handbook (version 3.0)*, pp. 25–26 and Appendix G, the WWC applies a multiple comparison (MC) adjustment "to account for multiple comparisons or 'multiplicity,' which can lead to inflated estimates of the statistical significance of findings." (p. 25). A WWC review may report on multiple outcomes in a single domain for a number of reasons, as explained below. The MC adjustment that the WWC applies is to correct for the lens used in the WWC review of particular contrasts in a study, which is not necessarily all of the contrasts presented in a study. The only time that the WWC needs to apply an MC adjustment is when there are a number of contrasts, subgroups, and/or outcomes within a domain that are reported in a study that meet standards and are included in the WWC's presentation of the findings.

The appropriate application of the MC adjustment requires an understanding of what the WWC review team and the study authors consider primary and secondary contrasts of interest.

Primary contrasts are typically the most informative (or best test) of the intervention for WWC reports. These contrasts contribute to the evidence rating for a particular intervention. Typically, primary contrasts of interest have the following features:

- Contrasts estimated using the full study sample, not subgroups.
- Contrasts estimated on composite (full-scale) outcomes, not subscales.
- In the case of multi-arm studies, contrasts that provide (best) tests of the effectiveness of the intervention.

Secondary contrasts are additional comparisons presented in the study that are useful results; however, these results are not the main focus of the WWC review. That is, these results are typically included in WWC reports for transparency, but they do not contribute to the evidence rating. Typically, secondary contrasts will have the following features:

- Contrasts estimated on subgroups.
- Contrasts estimated on subscales.
- In the case of multi-arm studies, contrasts that provide tests of alternate, but related, versions of the intervention (those that are not primary/best test[s] of the intervention).

The WWC does not penalize studies that present secondary contrasts in an MC adjustment that examines impacts on primary contrasts. As a result, MC adjustments should only be applied within primary or secondary contrasts. If it is unclear whether a contrast should be considered primary or secondary, reviewers should raise the issue to be discussed during reconciliation.

Contrasts do not meet WWC group design standards are not included in the WWC's operationalization of the MC adjustment procedure. Consult with a review's lead (or deputy) methodologist for additional guidance regarding any specific issues encountered during a review or reconciliation.

K. Analyses of Gain Scores

As noted on page F.7 of the *WWC Procedures and Standards Handbook (version 3.0)*, some study authors will calculate gain scores by subtracting a pretest from the posttest and using the resulting difference as the dependent variable in an impact analysis. Such gain score or differenced analyses are eligible to meet WWC group design standards, but reviewers should address two special considerations in reviews of these studies.

1. **Analyses of gain scores do not typically establish equivalence, nor do they satisfy the requirement of a statistical adjustment.** First, nothing in the creation of a gain score ensures that the pretests are equivalent, although pretest means and standard deviations can be used to establish equivalence for an analysis of gain scores. Second, though a gain score is created using a pretest, the analysis of a gain score does not usually include the pretest separately because construction of the gain score effectively imposes a coefficient of 1 on the pretest. This approach fails to account for the correlation between scores on the pretest and posttest and therefore, could lead to over- or underestimating the adjusted group mean difference. If the evaluation of equivalence requires the analysis to statistically adjust for the pretest, the pretest must be included in the analysis separately, which would be unusual for a gain score analysis.

2. **Effect sizes from gain score analyses cannot be used in WWC reporting.** Effect sizes used in WWC reporting must be based on the posttest standard deviation, not the gain score standard deviation. Effect sizes calculated using means and standard deviations of gain scores, or from a regression coefficient in which the dependent variable was a gain score, are not comparable to effect sizes calculated using the posttest as a dependent variable (either with or without the pretest as a covariate). An effect size from gain scores will generally be larger than an effect size based on the posttest, because the standard deviation of gain scores is typically smaller than the standard deviation of posttest scores. For WWC reports that potentially use information from multiple outcomes or studies, such as intervention reports, the effect size based on the gain score cannot be included, and the reviewer should request the unadjusted posttest standard deviations from the study author(s)—gain score means can be used in the calculation. However, a study can *Meet WWC Group Design Standards With or Without Reservations* even if no effect sizes can be reported.

L. Pretest Reliability

RCTs with high attrition and QEDs can demonstrate equivalence using a pretest measure that is related, but different, from the outcome measure. For example, a study might examine impacts on a state-administered assessment provided at the end of third grade, but use a researcher-developed assessment that covers similar content for the demonstration of equivalence in the fall. In these cases, the reliability of the pretest measure is likely different from the reliability of the outcome measure.

The reliability of a pretest measure is a concern for establishing equivalence, with lower reliability leading to an underestimation of baseline differences. The true amount of dispersion for a test reflects how much student performance differs on the underlying skills measured by the test if the skills were measured without any error. Test scores measured with error have observed standard deviations that exceed the true amount of dispersion in student performance, and the standard deviation increases with lower reliability.

For example, suppose two tests measure the same true dispersion in student performance and observe the same difference in test score units (e.g., three points) between the two groups. The standardized mean difference divides the observed test score difference by the standard deviation. The test with the lower reliability will have a larger standard deviation, resulting in a *smaller* standardized mean difference. As reliability falls, there is a higher likelihood of incorrectly concluding that the groups are equivalent. This section provides guidance on assessing equivalence using a pretest that differs from the posttest.

Guidance

In group design studies, a pretest measure used to establish equivalence must satisfy the same reliability criteria specified for outcomes in the *WWC Procedures and Standards Handbook (version 3.0)*. If a measure of reliability is required for a pretest measure, but is not reported or obtained through an author query, or is below the required threshold, the measure cannot be used to assess or establish equivalence.

If a pretest is not required to establish equivalence, as with a low-attrition randomized controlled trial, the inclusion of the pretest in the analysis does not affect the study rating, even if it has low reliability.

M. Analyses using Propensity Scores

A propensity score is the probability that an observation would appear in the intervention group given its baseline and demographic characteristics. The scores are often used to identify a comparison group that is similar to an intervention group by finding individuals from a pool of potential comparison group members with scores that are similar to those of intervention group members. Alternatively, the scores can be used as weights in a regression analysis in which observations in one group that are similar to those in the other group receive more weight. This section provides guidance on how the WWC should review studies using propensity scores.

Guidance

When a study employs propensity scoring approaches, the WWC will review the study using the same framework as any other QED study, with a required demonstration of baseline equivalence of the analytic sample. However, for a propensity score analysis to credibly demonstrate equivalence of the analytic sample, there are three key considerations that WWC reviewers must assess.

1. **Only exogenous covariates have been used to create the propensity scores.** If endogenous covariates or outcomes are used in the creation of the propensity scores, the scores may ultimately lead to biased impact estimates. For more details, see the WWC guidance for reviewing analyses that include potentially endogenous covariates (p. 13, above).
2. **Sample sizes have not been artificially inflated through matching with replacement.** The WWC standards indicate that propensity score analyses that use either weighting or matching techniques are acceptable, including matching with replacement. However, if the study used matching with replacement in a manner that leads to larger sample sizes following the matching procedure, reviewers should examine whether the study authors took reasonable precautions in the calculation of standard errors to ensure that the additional records do not contribute to "artificially" precise estimates. For example, a study might appropriately address this concern by applying a clustering correction to account for repeated observations of units in the analysis.
3. **Consistency in the analytic approach used to demonstrate equivalence and the approach used to estimate impacts.** When assessing equivalence for an analysis that uses propensity scores, reviewers should examine whether baseline means and standard deviations were calculated in a manner consistent with how the impact of the intervention was estimated. As with any analysis that needs to demonstrate equivalence, reviewers should assess equivalence based on the data that were included in the analytic sample. If the study used propensity score weights, the baseline means should also be calculated using the same weights. Equivalence must be demonstrated on the variables specified in the review protocol; it is not sufficient to establish equivalence on the propensity scores.

N. Confounding Factors

The *WWC Procedures and Standards Handbook (version 3.0)* defines a confounding factor as "a component of the study design or the circumstances under which the intervention was implemented are perfectly aligned … with either the intervention or comparison group" (p. 19). That is, a confounding factor is one that is always present for members of one group and never present for members in the other group. As a result, it is not appropriate to report that the observed impact was caused solely by an intervention. This section provides guidance on key features that should be examined when assessing confounding factors in group design studies.

Examples of Confounding Factors

Three specific types of confounding factors are described on pages 19–20 in the *WWC Procedures and Standards Handbook (version 3.0)*. These are:

1. The intervention or comparison group contains a single study unit (n = 1).

Examples:

- Two schools are randomly assigned, one to each condition.
- A study has two intervention classrooms and two comparison classrooms, but both intervention classrooms had the same teacher, who was different from the teachers in the comparison group and had no interaction with the comparison group.

Examples of similar circumstances that are <u>not</u> confounding factors:

- Students are randomly assigned to condition and are all taught by the same teacher in the same school. This is not a confounding factor because the same teacher taught both conditions.
- Schools from three school districts are randomly assigned to condition. Two of the districts have schools that are represented in both conditions, but all schools in the third district were assigned to a single group. This example does not have a confounding factor because both groups contain schools from at least two districts.
- The intervention condition includes a single school, and the comparison group includes multiple schools. The intervention of interest is attending the school itself. This is not a confounding factor because in this instance, the school and the intervention are one and the same.

2. The characteristics of the units in the intervention or comparison group differ systematically in ways that are associated with the outcomes with no overlap.

Example:

- Comparison students were fifth-grade students enrolled in a school during the 2013–14 school year, and intervention students were fifth-grade students enrolled in the same school during the 2014–15 school year. (Also see *Guidance: Groups Defined by Cohort*.)

- Students in the comparison condition are all in grade 7, and students in the intervention condition are all in grade 8.

Examples of similar circumstances that are not confounding factors:

- Students volunteer to enroll in two different types of mathematics courses: one uses a novel group-based approach (the intervention condition), and one uses a more traditional teacher-directed style (the comparison condition). Some characteristics of students who volunteered for the intervention condition may differ from those who volunteered for the comparison condition (for example, more extroverted students select the group-based program, and more introverted students select the teacher-directed style). This type of student self-selection into condition is not a confounding factor because the selection procedure does not create groups that differ systematically with no overlap. However, differences in the composition of the samples by condition that arise due to selection could be identified in the equivalence assessment.
- Classrooms in the intervention condition have much lower rates of students who are eligible for free or reduced-price lunch, compared to those in the comparison condition. This is not a confounding factor because there is some overlap in the characteristic between the groups. However, under some review protocols, this difference could be considered in an assessment of equivalence.
- The intervention group includes students who received the intervention in 2 different school years. The comparison group also includes students in both school years, but in a different proportion (i.e., there are more intervention group students in the Year 1 sample, and more comparison group students in the Year 2 sample). This is not a confounding factor because there is some overlap in the characteristic between the groups. However, under some review protocols, this difference could be considered in an assessment of equivalence.

3. The intervention is always offered in combination with a second intervention (and the combined intervention is not of interest for the review according to the review protocol).

Examples:

- The focus of the review is a specific software program. Students in the intervention condition were exposed to two software programs (the specific software program that is the focus of the review, plus an additional program), but students in the comparison condition were not exposed to any software programs.
- A researcher conducted a previous study in the same classrooms as the current study, and students in the intervention condition received a different intervention during the first study that is not the focus of the review. The previous intervention is a confounding factor if the pretest used in the current study was given before or during the earlier study, such that students received both interventions between the pretest and the posttest.

Example of a similar circumstance that is not a confounding factor:

- A researcher conducted a previous study in the same classrooms as the current study. The pretest used to establish equivalence in the current study was given after the completion of the earlier study. Although the current study occurs in the same classrooms as the earlier study, this is not a confounding factor because the experiences of students prior to the pretest are not relevant to WWC reviews.

III. GUIDANCE FOR REVIEWS UNDER THE SINGLE-CASE DESIGN STANDARDS

A. Confounding Factors in Single-Case Designs

In single-case designs, teachers, parents, or peers (collectively labeled *interventionists*) can administer the intervention to study participants. When study participants experience a different interventionist across baseline and intervention phases of the study, the study has a potential confounding factor. This section provides additional guidance for the identification of confounding factors in single-case designs.

Guidance

As it can sometimes be difficult to determine whether something is a confounding factor, the examples below describe situations for which the interventionist is and is not a confounding factor.

- Examples of confounding factors: participants have a different interventionist across the baseline and intervention phases, noted by underline below.
    - One teacher teaches all cases in the baseline phase and a different teacher teaches all cases in the intervention phase.

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

    - One teacher teaches all cases in the baseline phase, and that same teacher and another teacher (or trainer) teach all cases in the intervention phase.

	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

- Examples of similar circumstances that are not confounding factors
    - 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

    - Multiple teachers teach different cases; teachers do or do not teach different phases.

	Baseline	*Intervention*	*Baseline*	*Intervention*
Case 1	Teacher 1	Teacher 3	Teacher 1	Teacher 1

Case 2	Teacher 2	Teacher 4	OR	Teacher 2	Teacher 2
Case 3	Teacher 2	Teacher 4		Teacher 3	Teacher 3

If a confounding factor is identified, then the study *Does Not Meet WWC Pilot Single-Case Design Standards* because measures of effectiveness cannot be attributed solely to the intervention.

B. Inter-Assessor Agreement Reporting

Single-case design studies reviewed by the WWC require a demonstration of sufficient outcome reliability. Appendix E of the *WWC Procedures and Standards Handbook (version 3.0)* states: "For each case, the outcome variable must be measured systematically over time by more than one assessor. The design needs to collect inter-assessor agreement [IAA] in each phase and at least 20% of the data points in each condition (e.g., baseline, intervention) and the inter-assessor agreement must meet minimal thresholds." This section provides additional guidance for evaluating the inter-assessor agreement in a study.

IAA assessed in each condition

A footnote to the second sentence listed above states that "Study designs where 20% of the total data points include IAA data, but where it is not clear from the study text that 20% of the data points in each condition include IAA data, are determined to meet this design criterion, although the lack of full information will be documented." The *Handbook* does not indicate how IAA information should be communicated in WWC products, or whether an author query should request this information when it is not provided.

Guidance

When a study does not report the percentage of sessions *in each condition* that are included in the IAA data—but the study mentions that at least 20% of the total sessions are checked for IAA, and IAA is checked at least once in each phase—reviewers should document the lack of information on IAA by condition. In Appendix B of either an intervention or single study report, the description of the outcome should include the following text: "The authors collected inter-assessor agreement (IAA) data in each phase and on at least 20% of all sessions, but it is unknown if IAA data were collected during 20% of the data points in each condition."

Provided that the authors report that at least 20% of the total sessions are checked for IAA and that IAA is checked at least once in each phase, an author query should not be conducted for whether IAA was measured in at least 20% of the sessions in each condition. Author queries should be conducted only if the authors do not report (1) the total percentage of sessions checked for IAA, (2) whether IAA was checked at least once in each phase for each participant, or (3) the IAA statistic (for example, percentage agreement) used to demonstrate reliability.

If study authors do not report that at least 20% of the total sessions were checked for IAA and/or that IAA was checked at least once in each phase, the study *Does Not Meet WWC Pilot Single-Case Design Standards* because the eligible outcomes do not meet WWC requirements; more specifically, the outcomes do not meet minimum IAA requirements.

IAA assessed in each phase for each case

The *Standards Handbook* states that each outcome must be measured over time by more than one assessor, with inter-assessor agreement collected in each phase. However, the *Standards Handbook* does not indicate whether an author query should be conducted if there is uncertainty about whether the study collected IAA data during *each phase and for each case*.

Guidance

An author query should be conducted if the authors do not specify that IAA data were collected during *each phase and for each case* for an outcome (in other words, IAA data must be collected at least once for each phase/case combination).

- If a study with more than one case uses a statement such as "IAA data were obtained for this outcome for approximately 25% of sessions, across each phase," an author query should be conducted to verify that IAA data were collected in each phase *for each case*.
- If a study uses a statement such as "IAA data were obtained for this outcome for all cases, across each condition," and there were multiple phases within conditions (for example, in a reversal-withdrawal design), an author query should be conducted to verify that IAA data were collected *during each phase* for each case.
- If the authors randomly chose the sessions during which IAA data were collected, an author query should be conducted if the study does not make clear that IAA data were collected during each phase and for each case.
- If a study uses a statement such as, "IAA data were collected for this outcome across all phases and participants," reviewers can give the study the benefit of the doubt and assume that IAA data were collected during each phase, for each case for the outcome.

If study authors do not report that IAA data were collected at least once for each phase/case combination, the study *Does Not Meet WWC Pilot Single-Case Design Standards* because the eligible outcomes do not meet WWC requirements; more specifically, the outcomes do not meet minimum IAA requirements.

IAA minimum thresholds

The existing standards do not provide minimum thresholds for specific IAA metrics. The *Handbook* states "Inter-assessor agreement (commonly called inter-observer agreement) must be documented on the basis of a statistical measure of assessor consistency. Although there are more than 20 statistical measures to represent inter-assessor agreement (e.g., Berk, 1979; Suen & Ary, 1989), commonly used measures include percentage agreement (or proportional agreement) and Cohen's kappa coefficient (Hartmann, Barrios, & Wood, 2004). According to Hartmann et al. (2004), minimum acceptable values of inter-assessor agreement range from 0.80 to 0.90 (on average) if measured by percentage agreement and at least 0.60 if measured by Cohen's kappa." (p. E-2). The *Handbook* also does not specify whether inter-assessor agreement must meet minimal thresholds for each outcome *across all cases in the study,* or for each outcome *separately for each case and/or phase*.

Guidance

The minimum for percentage agreement—regardless of whether the metric is exact agreement or agreement within one—is 80% (or 0.80). The minimum kappa or correlation is 0.60. IAA needs to meet these minimum values for each outcome *across all phases/cases*, but not separately for each case or phase. If study does not meet these minimum values for each outcome *across all phases/cases*, the study *Does Not Meet WWC Pilot Single-Case Design Standards* because the

eligible outcomes do not meet WWC requirements; more specifically, the outcomes do not meet minimum IAA thresholds.

C. Changing Criterion Designs

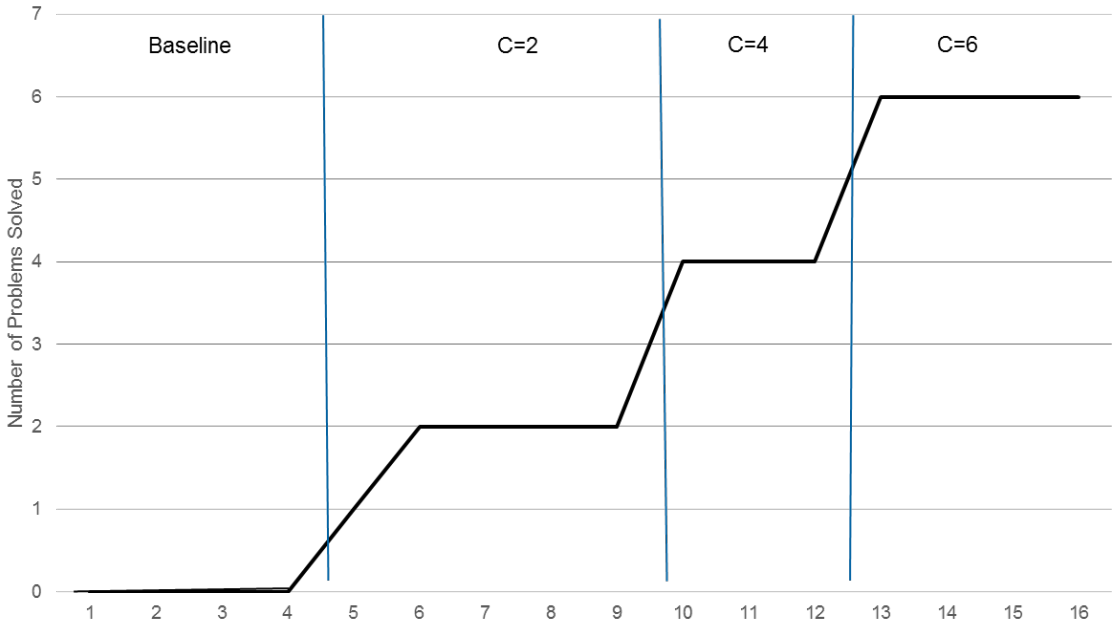
The *WWC Procedures and Standards Handbook (version 2.1)* noted that the changing criterion design is a variant of a reversal-withdrawal (or ABAB) design, "In this design the researcher examines the outcome measure to determine if it covaries with changing criteria that are scheduled in a series of predetermined steps within the experiment. An A phase is followed by a series of B phases (e.g., B1, B2, B3…BT), with the Bs implemented with criterion levels set for specified changes. Changes/differences in the outcome measure(s) are assessed by comparing the series associated with the changing criteria." (pp. 65–66). This section provides additional guidance for evaluating changing criterion designs under the version 3.0 Pilot Single-Case Design Standards described in Appendix E of the *WWC Procedures and Standards Handbook (version 3.0)*.

Guidance

The reversal-withdrawal design standards and visual analysis approach described in Appendix E of the *Handbook* should be applied to changing criterion designs. Each baseline/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 using this 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.

Figure C.1 provides an example of a changing criterion design experiment. The example displays the number of math problems correctly solved during baseline and intervention phases. After a stable baseline of 0 problems solved was established, a criterion of 2 was established and 10 minutes of free choice time was made contingent on meeting criterion. Once the child met this criterion for several consecutive sessions, the criterion was raised to 4. Once the child met this performance, the criterion was increased to 6.

Figure C.1. Example of a Changing Criterion Design Experiment



D. Alternating Treatment Designs

Alternating treatment (AT) designs rapidly alternate between two or more interventions to examine how outcomes change. Appendix E of the *WWC Procedures and Standards Handbook (version 3.0)* states that AT designs must have:

- A minimum of five data points per condition to *Meet WWC Pilot Single-Case Design Standards Without Reservations*.
- A minimum of four data points per condition to *Meet WWC Pilot Single-Case Design Standards With Reservations*.

Only phases with at most two data points are considered, because a phase with more than two data points does not constitute a rapid alternation.

This section provides additional guidance for reviews of AT designs, including the potential for residual treatment effects, characterizing the level of evidence for a causal relationship, and determining a baseline pattern of responding.

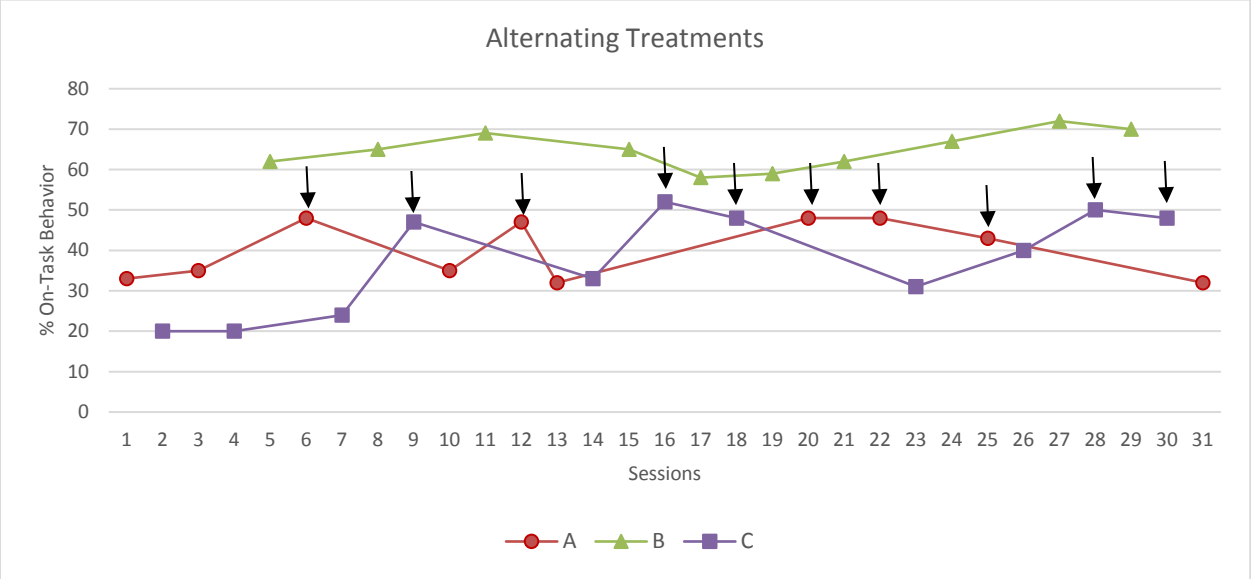
Guidance

Residual treatment effects

The *Handbook* states "when designs include multiple intervention comparisons (e.g., A versus B, A versus C, C versus B), each intervention comparison is rated separately," but also gives methodologists and context experts discretion in determining if the "design is appropriate for evaluating an intervention." This discretion is needed in AT designs because of the potential for *residual treatment effects*—responses within phases/conditions that are caused by interventions in previous phases/conditions (sometimes called *multiple treatment interference* [Kazdin, 2011]). It is not possible to isolate each intervention for separate comparison as required by the *Handbook* when residual treatment effects are present.

For example, consider an experiment in which (1) interventions A, B, and C are all behavior modification interventions that aim to impact the percentage of on-task behavior, (2) interventions A and C do not have residual treatment effects, and (3) intervention B is an effective intervention that causes students to engage in more on-task behavior for the next several hours, including sessions during which other interventions are implemented. In this example, average on-task behavior for interventions A and C will be higher on average when the intervention session follows B than when B follows A and C (see arrows in Figure D.1 for a graphical representation). In this example, the comparison of A vs. C depends, in part, on which condition follows a B session.

Figure D.1. Example of Residual Treatment Effects



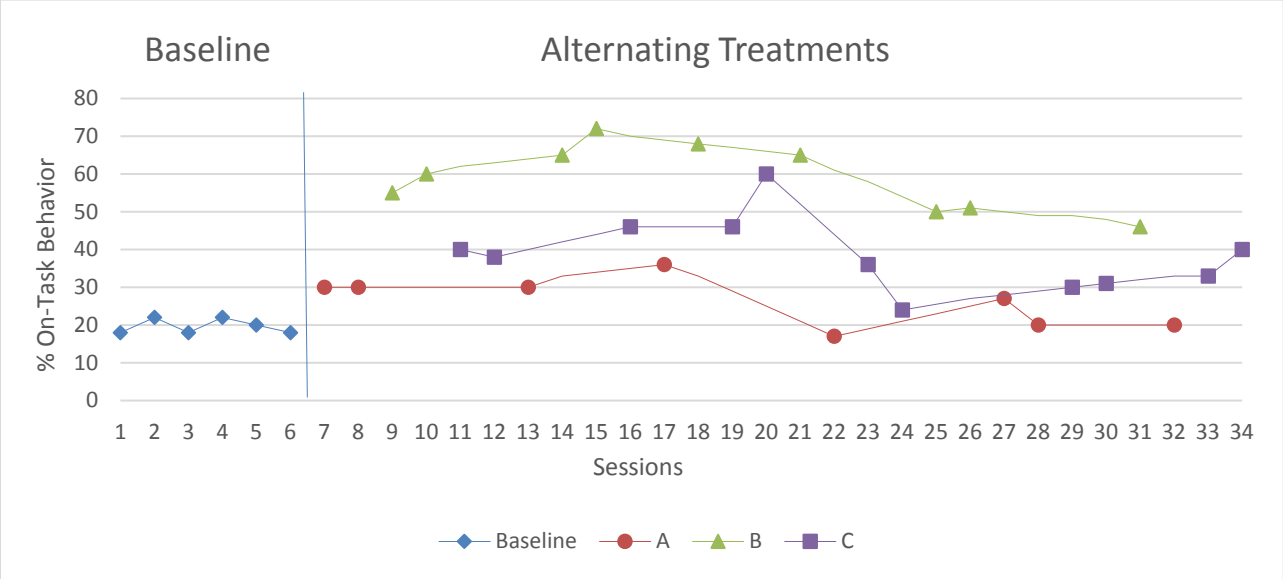
When a review team identifies an eligible AT-design experiment that uses three or more interventions, the team should ask the content expert to determine whether residual treatment effects are likely given the specific interventions and outcomes in the experiment (the review team can rely on previous approval of similar conditions and outcomes from the content expert). The review team should 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.

If the content expert and reviewer both agree that there are likely to be residual treatment effects, then the study *Does Not Meet WWC Pilot Single-Case Design 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 with the content expert. 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.

No residual treatment effects

When completing reviews with no residual treatment effects, the reviewer should focus only on the intervention(s) under review and relevant comparison condition(s) when assigning a study rating or conducting the visual analysis to assess the level of evidence. For example, the comparison of interest in a study with three intervention conditions may be A and B, while C is not of interest. In this case, when one or more sessions of intervention C occur between interventions A and B, the reviewer should ignore the C session(s), and compare only the A and B sessions (for example, in Figure D.2, reviewers would ignore data points from Sessions 11 and 12 [C] when comparing Session 10 [B] to Session 13 [A]). Also, reviewers should only compare the A-B points that are closest together (for example, reviewers should compare Points 8 and 9 in Figure D.2, for the first A-B comparison, and Points 10 and 13 for the next comparison of these two conditions).

Figure D.2. Example AT Design with Baseline Phase



Assessing the level of evidence

The *Handbook* characterizes an experiment as having *Strong Evidence* of a causal relation when it has "at least three demonstrations of the treatment effect along with no noneffects" and *Moderate Evidence* of a causal relation when it has "three demonstrations of an effect and also includ[ing] at least one demonstration of a noneffect."

Reviewers might find it more difficult to identify effects and noneffects in AT designs than in multiple baseline designs or reversal-withdrawal designs where a pattern of data points emerges over the course of a phase. Specifically, an effect or noneffect in a reversal-withdrawal design might be clearly demonstrated by similar or dissimilar data trends in adjacent phases of two different conditions. In contrast, it may be difficult to identify effects or noneffects in AT designs when the comparison from one phase to the next is based on only one or two data points in each condition.

Additionally, multiple baseline and reversal-withdrawal designs (e.g., ABAB) often only have three or four phase changes, so a noneffect in one phase change may clearly illustrate a lack of impact for the entire design. However, in an AT design with numerous phase changes, it may be difficult to characterize the evidence through visual analysis if there are many clear effects and one ambiguous effect, or if there are three clear effects, but the overall mean of the two conditions is the same.

Finally, in a multiple baseline experiment, a noneffect indicates that the intervention did not have an impact on a particular case or cases. However, in AT designs, several noneffects occurring in a short period of time within the same case raise concerns about whether any effects are valid (i.e., when there are three demonstrations of an effect and several noneffects within one case, it is plausible that any observed "effects" are random variation or noise). Comparing the overall means across conditions helps verify that the intervention has an actual effect.

For AT design experiments, the requirements for a *Strong Evidence* or *Moderate Evidence* characterization (or visual analysis rating) when comparing two conditions are:

1. At least three demonstrations of an effect in the same direction, based on a comparison of levels in phases closest together (ignoring any intervening interventions)
2. No clear effects in the opposite direction
3. The overall mean levels for the intervention and comparison conditions clearly demonstrate a visual effect. The overall means should include all points, including outliers, for each condition to provide evidence that random variation (noise) is implausible as an explanation for effects [the standards indicate that the "observed and projected patterns of the outcome variable between phases… demonstrates evidence of a causal relation"].

Consistent with the *Handbook*, if the data meet the three parameters above, and there are no noneffects, there is *Strong Evidence* of a causal relationship. If the three parameters above are met, but there is at least one demonstration of a noneffect, there is *Moderate Evidence* of a causal relationship. If the data do not meet all three parameters above, or there are not three demonstrations of an effect, there is *No Evidence* of a causal relationship.

Finally, *ambiguous effects* can play a role in the characterization of the evidence of a causal relationship, but *trends* in the effects do not.

Ambiguous effects. When conducting visual analysis, reviewers may encounter an ambiguous (or very small) effect (e.g., the comparison of Points 20 and 21 in Figure D.2). Ambiguous effects should never be counted towards the three demonstrations required for a *Strong Evidence* or *Moderate Evidence* characterization. Additionally, the ambiguous effects should only be treated as noneffects when they are in the *opposite direction* from the other effects counted towards the three demonstrations. Only then can ambiguous effects lead to a *Moderate Evidence* characterization.

For example,

- In the B-C comparison in Figure D.2, there are: (1) at least three demonstrations of a positive effect; (2) no clear effects in the opposite direction; (3) no clear noneffects (we do not count the ambiguous effect in the same direction—from Sessions 20 to 21 as a noneffect), and (4) clear differences in the overall mean level of the two conditions. The reviewer should assign a *Strong Evidence* rating.
- In the A-C comparison in Figure D.2, there are: (1) at least three demonstrations of a positive effect, (2) no clear effects in the opposite direction (we do not count the ambiguous effect in the opposite direction—from Sessions 24 to 27 as an effect in the opposite direction), (3) one noneffect (we do count the ambiguous effect in the opposite direction—from Sessions 24 to 27 as a noneffect) and (4) clear differences in the overall mean level of the two conditions. The reviewer should assign a *Moderate Evidence* rating.

Trends. Data trends are not considered when using visual analysis to characterize the evidence rating of an AT design experiment. For example, if an experiment demonstrates at least three effects early on, but then the data patterns merge towards the same point, this design can still be characterized as providing *Moderate Evidence* of an effect as long as the overall condition means are different and there are no clear effects in the opposite direction.

Baseline sessions (establishing the concern)

The *Handbook* states that "the first step in the visual analysis is to determine whether the data in the Baseline 1 (first A) phase document that the proposed concern/problem is demonstrated (e.g., tantrums occur too frequently)." However, some AT designs do not include a baseline phase, and when a baseline phase is included, it typically does not reflect a counterfactual—in an AT design, the alternating treatment(s) that is not the intervention of interest serves as the counterfactual. For example, in Figure 2, the initial baseline data points (during Sessions 1–6) will not serve as the counterfactual when conducting the visual analysis. An exception is AT designs that include "baseline" as one of the alternating treatments; in these designs, baseline data points that occur during the AT phase do serve as a counterfactual.

The visual analysis for AT designs for which the counterfactual is not represented by the baseline period does not need to determine if there is sufficient demonstration of a clearly defined baseline pattern of responding that can be used to assess the effect of an intervention. Instead, reviewers should use the appropriate guidance below, depending on the data presented in the study.

AT designs with baseline sessions. As part of Step 1 of the visual analysis, reviewers should determine if the proposed concern (e.g., lack of on-task behavior) is demonstrated. Using Figure D.2 as an example, a reviewer would evaluate the initial baseline data points (Sessions 1–6) to assess whether the proposed concern is demonstrated. In this example, the low rate of on-task behavior indicates a concern.

AT designs that do not have initial baseline sessions. The proposed concern can be demonstrated through one or more of the following three sources of evidence: (1) a business-as-usual condition—some AT designs include business-as-usual or baseline as one of the alternating treatments (e.g., in Figure D.1, if C was business-as-usual, the first three points demonstrate the concern), (2) a description of the problem in the text, or (3) a determination from the lead methodologist that the experiment does not need to demonstrate the concern because of ethical concerns. The third source is only a possibility when the review protocol indicates that the lead methodologist has discretion to require fewer than three data points in a phase for ethical reasons. If none of these conditions is met, the concern is not demonstrated, and the report should characterize the experiment as providing *No Evidence* because it does not demonstrate the proposed concern.

E. Reversal-Withdrawal, Multiple Baseline, and Multiple Probe Designs with More Than Two Conditions

Some reversal-withdrawal, multiple baseline, or multiple probe single-case design experiments have more than two conditions (e.g., ABCABC reversal-withdrawal design). The *WWC Procedures and Standards Handbook (version 3.0)* does not provide specific standards for reviews of studies that use these designs. This section provides guidance on the contrasts of interest and the potential for intervening intervention effects in these studies.

Guidance

Contrasts of interest

When there are multiple possible comparison conditions, there can be multiple possible research questions (e.g., Is A more effective than B? Is A more effective than C? Is A more effective than B or C?). The research question and focal comparison condition can influence the direction and magnitude of any effects. The *Handbook* does not specify whether an effect must be demonstrated with only one comparison condition at a time (e.g., A vs. B and A vs. C) or whether an intervention can be simultaneously compared to combinations of two or more conditions (e.g., A vs. B or C).

To be consistent with the typical WWC study research question, the three-demonstrations-of-an-effect requirement should only refer to contrasts between a single intervention condition and a single comparison condition. Comparisons between the intervention and multiple conditions (e.g., A vs. B or C) are not eligible for review.

For some single-case design experiments, the reviewer, after discussion with review team leadership, might decide that two conditions are effectively identical and should be reviewed as one phase. This decision should be based on: (1) study descriptions that indicate the two conditions are similar, and (2) outcome data that indicate the two conditions are similar. For example, in an ABCABC design, where B and C are slight variations of the same intervention, the content expert can use the text descriptions to determine that B and C are effectively the same condition and can be treated as a single "B" phase. If the data are consistent with this determination (see Figure E.1), this experiment can then be reviewed as an ABAB design. In this case, the reviewer should proceed with the review, treating the design as an ABAB design. Reviewers should always discuss cases like this with review team leadership before completing the review.

Residual treatment effects

The WWC standards require three attempts to demonstrate an effect for each comparison, but intervening third or fourth conditions can hinder direct comparisons. For example, in an ABCABC design, there are only two adjacent A-B comparisons, and no direct attempt to demonstrate a reversal effect from B to A. Ignoring the intervening C condition would allow an assessment of a reversal effect. However, this is only justified when *residual treatment effects*—responses within phases or conditions that are caused by interventions in previous phases or conditions (sometimes called *multiple treatment interference* [Kazdin, 2011])—are not present. Specifically the assessment of the reversal from B to A could be confounded by a persistent effect of condition C.

For reversal-withdrawal and multiple baseline/probe designs, additional conditions that occur after the relevant intervention and comparison condition (e.g., the C condition in an ABABC reversal-withdrawal design or ABC/ABC/ABC multiple baseline design) cannot create residual treatment effects. Accordingly, the reviewer should only evaluate the experiment with the first two conditions (e.g., A vs B). The WWC prioritizes the review of the first two conditions because that comparison does not require any assumptions about residual treatment effects—this experiment provides a stronger design.

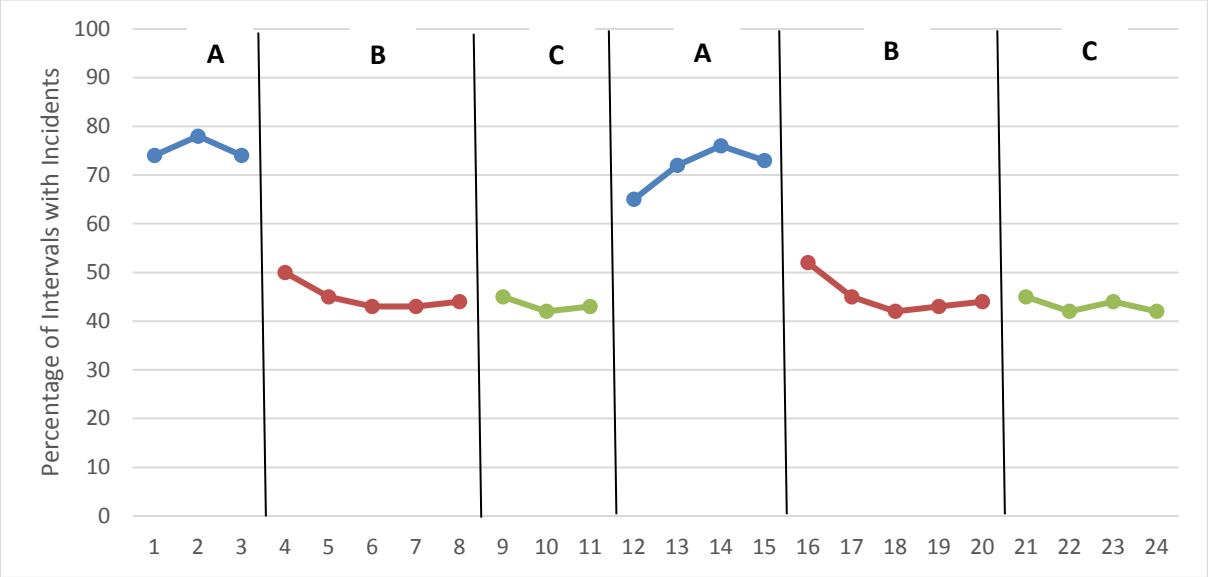
For multiple baseline/probe designs with a third condition, reviewers should review the first two phases. Review team leadership can consider an experiment including the third phase, if (1) the first two phases do not form an experiment that meet standards with or without reservations, or (2) a comparison with the third phase is most relevant. Similarly, review team leadership can decide to review a reversal-withdrawal design with an intervening third condition if an alternative design is not available or useful.

When reviewing a design with an intervening third condition, the review team needs to first determine whether there are likely to be residual treatment effects, following the same steps described in the guidance for alternating treatment designs. If residual effects are likely, the comparison with an intervening condition should be rated *Does Not Meet WWC Pilot Single-Case Design Standards* because the measures of effectiveness cannot be attributed solely to the intervention.

For reversal-withdrawal designs, if residual effects are unlikely, then the reviewer(s) should work with the review team leadership and content experts to identify appropriate standards for the review, focusing only on the intervention under review and the relevant comparison condition when assigning a study rating or conducting the visual analysis (i.e., ignoring any third or fourth interventions). The alternating treatment design guidance can be used as a foundation. However, reversal-withdrawal, multiple baseline, and multiple probe designs generally have longer phases than alternating treatment designs, which means more time will pass between the non-contiguous phases that will be compared (e.g., between the first B and second A in an ABCAB reversal-withdrawal design). This additional time could make it difficult to determine the immediacy of an effect and allows more threats to causal validity, such as history or maturation. These threats will need to be considered as part of the review, and if the threats are determined to be large, the team can determine that the design *Does Not Meet WWC Pilot Single-Case Design Standards* because the measures of effectiveness cannot be attributed solely to the intervention.

Reviewers should document the phases used in the review and the reasons why some may have been excluded from the review. This information will also be documented in WWC products that cite the study.

Figure E.1. Effectively identical B and C conditions.



Note. In this figure, Conditions B and C are effectively identical based on descriptions in the study.

F. Reversal-Withdrawal, Multiple Baseline, and Multiple Probe Designs with More Than the Minimum Number of Required Phases

Appendix E of the *WWC Procedures and Standards Handbook (version 3.0)* requires at least three attempts to demonstrate an effect at three different points in time in a single-case design experiment. To do so, the design must include a minimum number of phases. Specifically, reversal-withdrawal (ABAB) designs must have a minimum of four phases per case, and multiple baseline and multiple probe designs must have at least six phases (two phases for at least three cases or subjects). The *Handbook* requires that phases must have at least three data points to qualify as an attempt to demonstrate an effect, unless there is an exception noted in the protocol.

Some experiments have more than the minimum required number of phases. For example, a reversal-withdrawal design with six phases (ABABAB), or a multiple baseline design with four cases where each case has two phases.

For studies with more than the minimum required number of phases, the WWC study rating and evidence rating might depend on whether the reviewer evaluates all phases or only a subset of phases. This section provides guidance for studies that have more phases than the minimum required to meet standards. This guidance applies only to experiments with two conditions (see the separate guidance for designs with more than two conditions [e.g., ABCABC]). For guidance on alternating treatment designs with more than the minimum number of phases, see the Alternating Treatment Design guidance.

Guidance

The reviewer should first conduct the review considering all phases/cases (i.e., review the experiment as conducted and reported). If the experiment *Meets WWC Pilot Single-Case Design Standards With or Without Reservations* when considering all phases/cases, the reviewer should complete the review without separately considering subsets of phases. Phases that are not primarily aimed at measuring the effectiveness of the intervention of interest, such as those related to diagnostic assessment or generalization, should always be excluded from the review.

If the experiment *Does Not Meet WWC Pilot Single-Case Design Standards* when considering all relevant phases (e.g., because some phases do not have at least three data points), the reviewer should conduct the review considering the subset of consecutive phases with enough points and determine if the subset can meet standards.

When selecting a subset of phases to review, the ultimate choice should be discussed with review team leadership. Reviewers should document the phases and cases used in the review and the reasons why some may have been excluded from the review. This information will also be documented in WWC products that cite the study.

The following examples illustrate this approach in practice.

Example 1: A reversal-withdrawal design has six phases (Figure F.1). The first two phases have only two data points each, and do not fulfill the criteria required to *Meet WWC Pilot Single-Case Design Standards*. However, the last four phases form a subset that would *Meet WWC Pilot Single-Case Design Standards Without Reservations* because there are at least five

data points in each phase. The third phase of the original design (first phase of the reviewed design) serves as a baseline to establish the problem and provide a counterfactual for the first reviewed B phase.

Example 2: A multiple baseline design has four cases (Figure F.2). The baseline phase for the first case has only two data points and does not fulfill the criteria required to meet WWC standards. However, focusing on just the last three cases, the experiment would *Meet WWC Pilot Single-Case Design Standards Without Reservations* because there are at least five data points in each phase and there are three attempts to demonstrate an effect.

Example 3: A multiple baseline design has four cases. All of the data points for the third and fourth cases completely overlap; thus, these two cases do not allow an effect to be demonstrated at different points in time. In this example, the reviewer should focus the review on the first three cases, which do allow for an effect to be demonstrated at three different points in time.

Example 4: The design is AAA|BBB|AA|BBB|AAA. Excluding the third phase (AA) would result in just two attempts to demonstrate an effect. Such an experiment *Does Not Meet WWC Pilot Single-Case Design Standards*.

All phases should be included in the review unless inclusion would cause the experiment to be rated *Does Not Meet WWC Pilot Single-Case Design Standards*. The following two examples illustrate this approach.

Example 5: The first five phases of an ABABAB design each have five points, and the sixth phase has four points. The review should include all six phases even though the highest rating the experiment can receive is *Meets WWC Pilot Single-Case Design Standards With Reservations* (instead of *Meets WWC Pilot Single-Case Design Standards Without Reservations* if only the first five phases are reviewed).

Example 6: The first five phases of an ABABAB design each have three points, and the sixth phase has two points. The review should include the first five phases even though the first four would form a design that could meet standards; the exception would be if including the first or fifth phase caused the study to be rated *Does Not Meet WWC Pilot Single-Case Design Standards* (for example, by resulting in IAA being assessed on less than 20% of sessions).

Finally, there may be multiple rigorous subsets of phases. Reviewers should select the subset aimed at measuring the effectiveness of the intervention of interest and the ultimate choice should be discussed with review team leadership. The following example illustrates this point.

Example 7: The design is AAAA|BBBB|AAAA|BBBB|AA|BBBB|AAAA|BBBB|AAAA|. Excluding the fifth phase (AA) would result in two separate designs, each which *Meets WWC Pilot Single-Case Design Standards With Reservations.* A close look at the article suggests that the intervention of interest (B) was altered by the teacher in phases six and eight—one component was not fully implemented—so the review should focus on the first four phases.

Figure F.1. Reversal-withdrawal design with six phases.

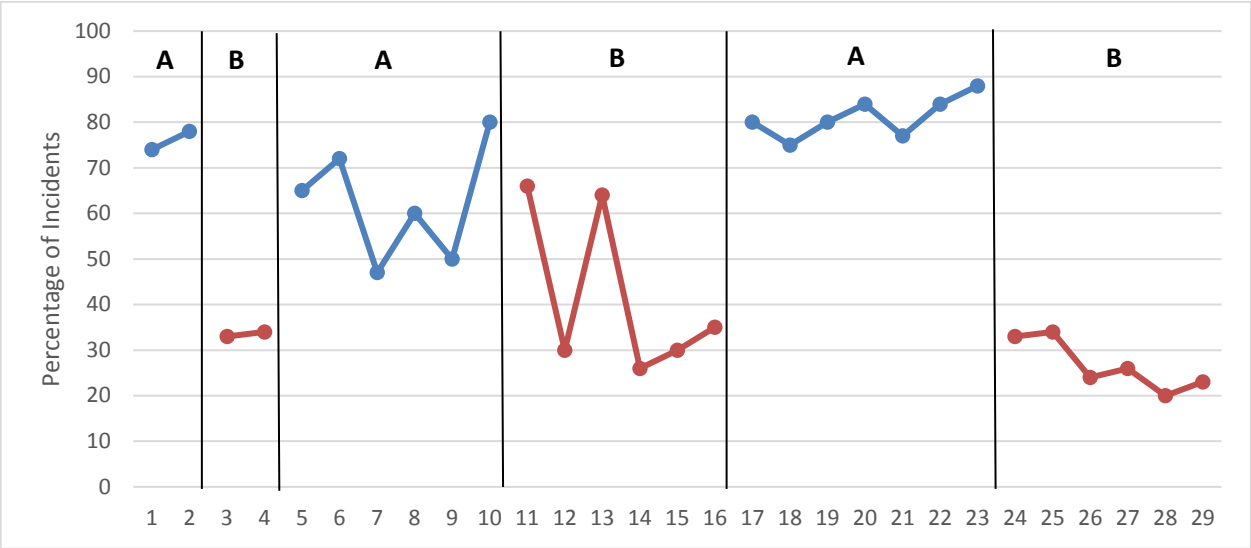
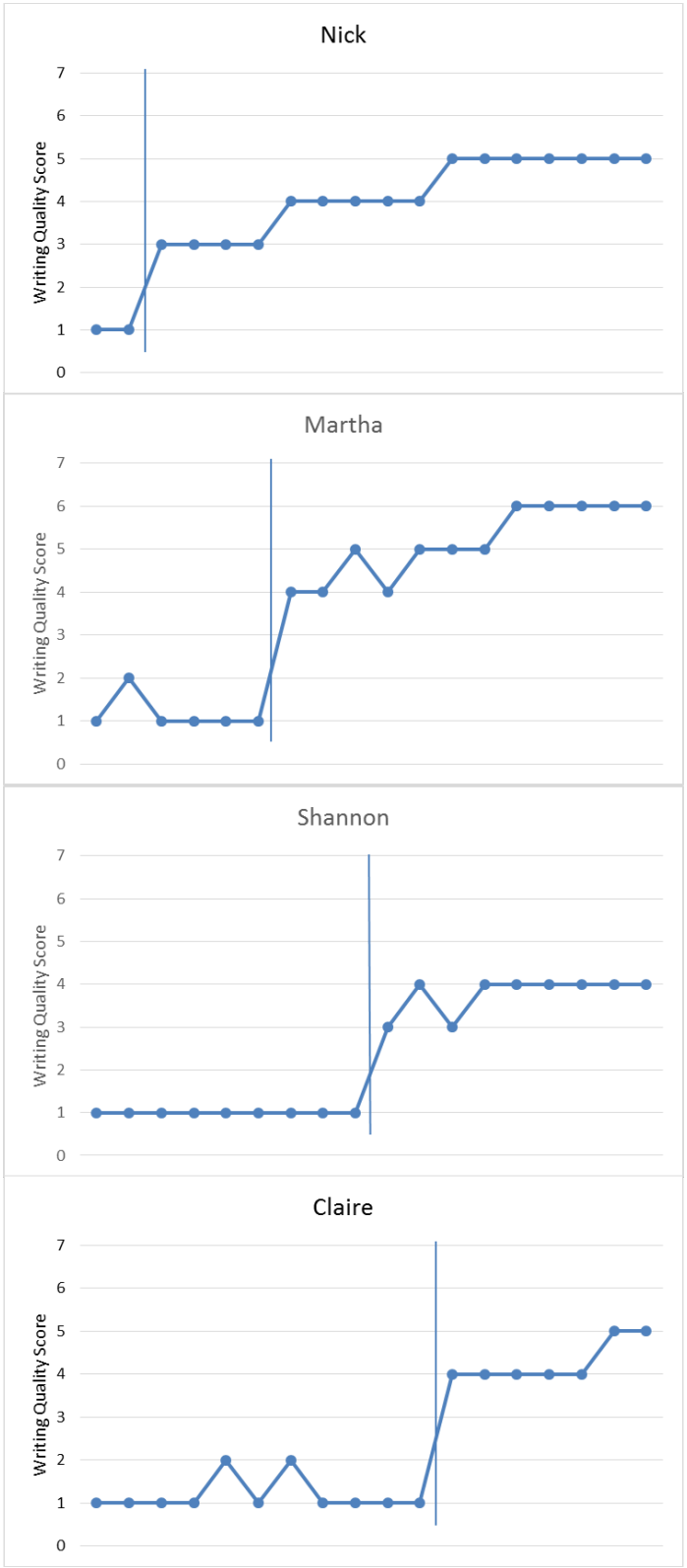


Figure F.2. Multiple baseline design with four cases.



G. Timing of Sessions in Multiple Baseline and Multiple Probe Designs

Single-case design experiments collect outcome data regularly, but the frequency of data-collection sessions can vary across or within studies (e.g., daily on school days, twice a week, or multiple sessions each day). The *WWC Procedures and Standards Handbook (version 3.0)* does not specify standards for how regularly sessions must occur, but consistent displays of time within an experiment are critical to implementing the standards in reviews of multiple baseline and multiple probe designs.

Multiple baseline and multiple probe designs involve concurrent and repeated outcome measures across all cases. (Multiple baseline experiments can also occur for one case across multiple school subjects or settings, with each school subject or setting presented as a different tier in the design. For simplicity, the examples in this guidance will refer to multiple baseline experiments across cases.) In multiple baseline designs, data collection occurs for each case during each session. In a multiple probe design, baseline data collection is intermittent (data are not collected for some cases during some sessions), generally because continuous collection of baseline data is unnecessary, impractical, or would be intrusive to participants (e.g., outcomes that require skills that the participant clearly does not possess).

The data for all cases in multiple baseline and multiple probe designs are usually presented in one figure, often with the same x-axis (measuring time). Although this type of display seems to imply that the same numbered session for each case occurred at a similar time, rather than days apart, this may not always be the case. For example, the left side of Figure G.1 appears to show Session 1 occurring at the same time for all three cases, whereas the right side of Figure G.1 clarifies that the timing was actually different for each case. Study authors may not always clearly indicate whether sessions occurred at the same time for each case.

Reviews of studies with these designs must examine graphical presentations of data that use a consistent display of time across cases for three reasons: to classify the design, to implement the standards, and to assess the timing of phase changes.

1. *Classify the design.* When time is not consistently displayed, a multiple probe experiment might be incorrectly classified as a multiple baseline. For example, the experiment represented in Figure G.1 uses a multiple probe design, which can be clearly identified when time is consistently displayed as shown in the right panel. However, when time is displayed inconsistently across the cases, as shown on the left, the experiment appears to use a multiple baseline design.

2. *Implement the standards.* In multiple probe designs, the "initial pre-intervention sessions must overlap vertically." This means that the initial sessions must occur on the same days. A graphical display that labels initial sessions with the same session numbers despite occurring on different days does not satisfy this requirement. For example, when time is displayed inconsistently across cases as shown on the left side of Figure G.1, Sarah appears to have two overlapping baseline points with the other two cases in the first three sessions. The right side of Figure G.1 clarifies that the overlap did not actually occur.

3. *Assess the timing of phase changes*. Reviewers must assess threats to internal validity related to timing of sessions including history, maturation, and testing. To do so, reviewers must verify that phase changes occur at different points in time. This is only possible when the graphical presentation of data for multiple baseline and multiple probe designs uses a consistent display of time across cases. For example, when time is displayed inconsistently as shown on the left side of Figure G.1, the intervention appears to have been implemented for Matt and Sarah at the same time (between Sessions 4 and 5). The right side of Figure G.1 clarifies that Matt's sessions actually occurred before Sarah's.

Guidance

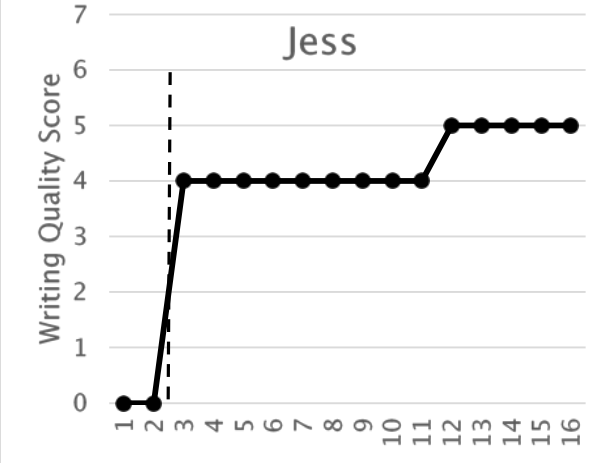
Reviewers should assess whether the display of time is consistent across all cases or tiers in the experiment and clearly identify any gaps in data collection. In particular, any time period with a data point for one case should either have a data point for other cases (if those cases had a session) or clearly identify the data as missing (if other cases did not have a session). For example, Figure G.1 (right side) shows a multiple probe design that clearly demonstrates intermittent baseline data collection.

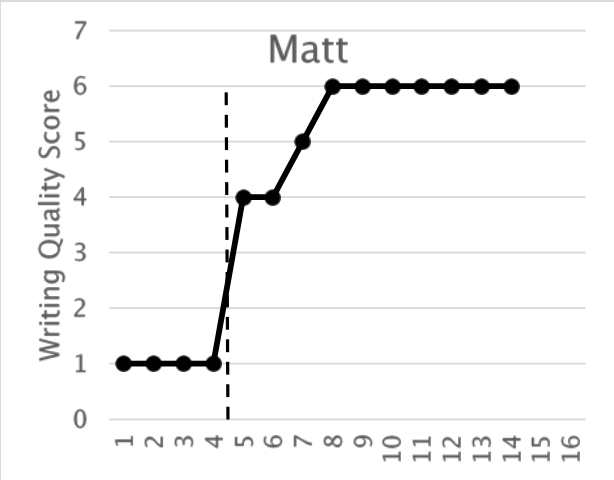
If the graph appears to present time consistently across cases and tiers, and there is nothing in the study text that suggests otherwise, reviewers should assume that the display of time is consistent across all cases and tiers and complete the review. If the display of time is not consistent, or the reviewer has concerns based on the study text, he or she should raise those concerns with the review team leadership. An author query may be needed to obtain a consistent display. Studies that do not present time consistently across all cases should be rated *Does Not Meet WWC Pilot Single-Case Design Standards* because there are insufficient data to evaluate the attempts to demonstrate an intervention effect.

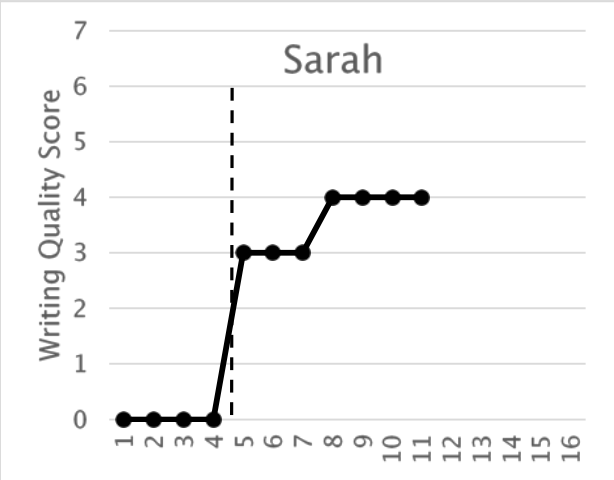
Note that some experiments intentionally do not collect data during a training phase for either the teacher or student; this guidance does not apply to those designs. Please see the guidance on experiments in which no outcome data were collected during the training phase.

Figure G.1. Inconsistent and consistent graphical presentations of time.

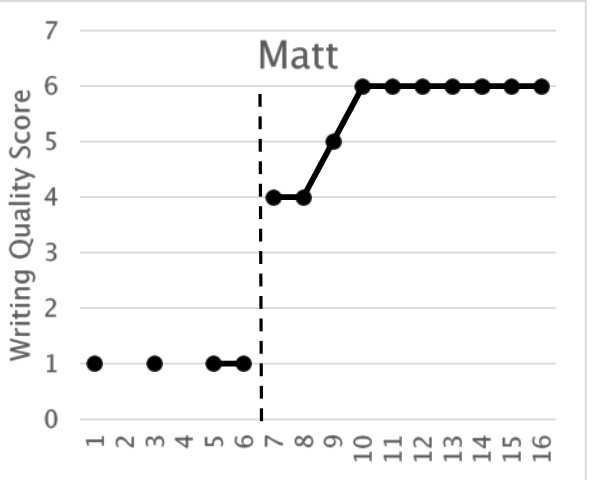
Graphical Presentation with Time Presented Inconsistently Across Cases

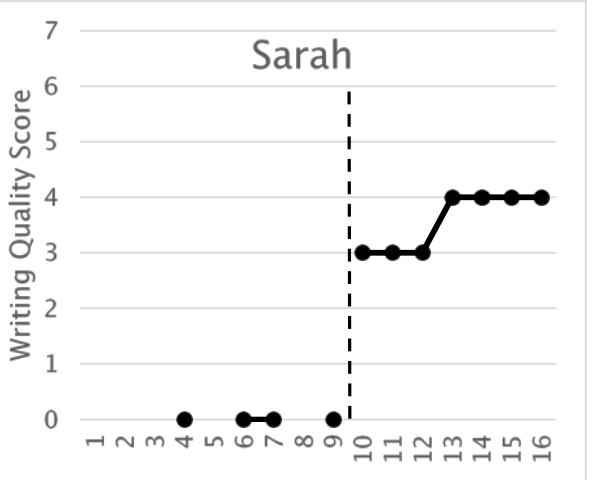






Graphical Presentation with Time Presented Consistently Across Cases





H. Concurrence in Multiple Baseline and Multiple Probe Designs

To meet standards, multiple baseline and multiple probe designs reviewed by the WWC must have some overlap in timing (concurrence) of data across cases or conditions. The WWC requires concurrence for these designs to make threats to internal validity less plausible. Appendix E of the *WWC Procedures and Standards Handbook (version 3.0)* states that multiple baseline and multiple probe designs "implicitly require some degree of concurrence in the timing of their implementation across cases when the intervention is being introduced." The *Handbook* does not specify the exact requirements for "some degree of concurrence." This section provides guidance on that requirement.

Guidance

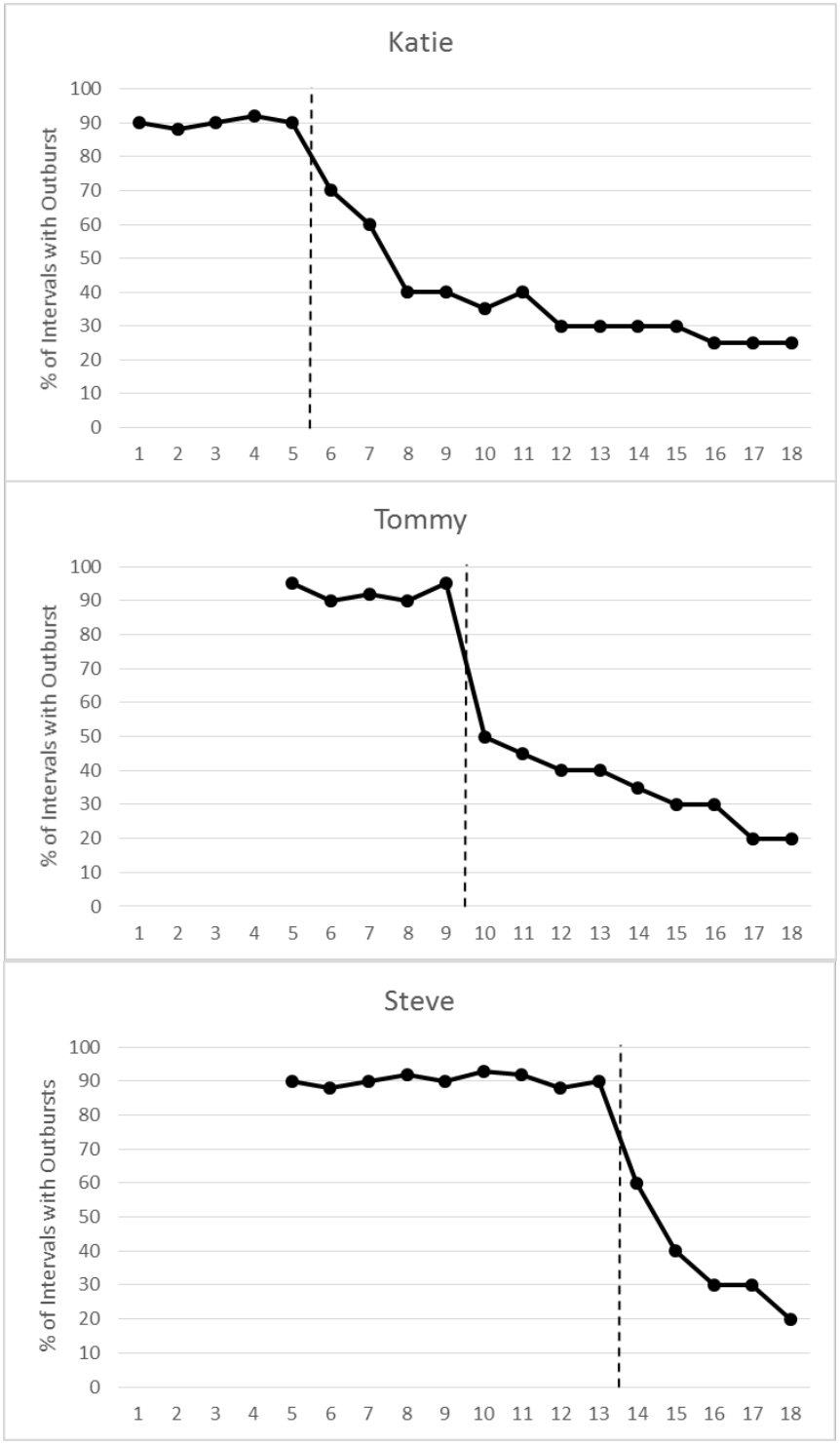
For studies relying on multiple baseline and multiple probe designs, reviewers should examine cases that have not yet received the intervention and determine whether these cases have baseline data before the intervention is administered to the first case (i.e., overlapping baselines) to meet the concurrence requirements.

For example, consider a hypothetical multiple baseline single-case design with three cases (see Figure H.1 below). Katie starts to receive the intervention in Session 6, Tommy starts the intervention in Session 10, and Steve starts the intervention in Session 14. For this multiple baseline design to have adequate concurrence, data collection for all three cases must begin before Session 6. In the figure below, this requirement is met. However, if such concurrent data were not collected, the design could not exclude threats to internal validity and would receive a rating of *Does Not Meet WWC Pilot Single-Case Design Standards* because there are insufficient data to evaluate the attempts to demonstrate an intervention effect.

The *Handbook* describes additional requirements for multiple probe designs, requiring that each case that has not yet received the intervention has outcome data collected in a session where another case either (a) first receives the intervention or (b) reaches the pre-specified intervention criterion.

For both multiple baseline and multiple probe designs, the level and trend of data points do not affect the study rating, but instead are used to assign an evidence rating based on the visual analysis.

Figure H.1. Demonstration of concurrence in a multiple baseline design.



IV. Guidance for Reviews of Studies that Present a Complier Average Causal Effect

In randomized controlled trials (RCTs), study participants are randomly assigned to groups that differ in eligibility for an intervention. However, study participants do not always comply with their assigned conditions. In the assigned intervention group—the group whose assignment makes them eligible for the intervention—some study participants might choose not to take up (receive) intervention services. In the assigned comparison group—the group whose assignment makes them ineligible for the intervention—some study participants might nevertheless take up the intervention.

In the presence of noncompliance, RCT studies have typically estimated either or both of two impacts. First, to estimate the effect of being assigned to the intervention, known as the intent-to-treat (ITT) effect, the mean difference in outcomes between the *entire* assigned intervention group and the *entire* assigned comparison group is calculated (possibly adjusting for unequal probabilities of assignment across different sample members).

Second, to estimate the effects of actually taking up the intervention, one common approach is to estimate the complier average causal effect (CACE).[3] The CACE is the average effect of taking up the intervention among *compliers*—those who would take up the intervention if assigned to the intervention group and who would not take up the intervention if assigned to the comparison group.

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 (IV) 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. An IV estimator starts from an assumption, known as the exclusion restriction, that neither the outcomes of always-takers nor the outcomes of never-takers differ between the intervention and comparison groups (because assignment to those groups cannot influence their take-up status). Any difference between the intervention and comparison groups must therefore be attributable to compliers. Likewise, the difference in take-up rates between the two groups reveals the fraction of study sample members who are compliers. Conceptually (and, in certain scenarios, mathematically), an IV estimator therefore estimates the effect of the intervention on compliers by dividing the difference in outcomes between the intervention and comparison groups by the difference in take-up rates. As discussed later, conventional statistical tests based on IV estimators perform well only if sample members' randomly assigned status has a strong association with take-up.

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

Prior to the approval of this guidance, WWC design standards for RCTs have focused only on ITT estimates, and CACE estimates have not been reviewed. The guidance described in this document is intended to specify the scenarios under which CACE estimates from RCTs are eligible for review and subsequently eligible to *Meet WWC Group Design Standards Without Reservations* or *Meet WWC Group Design Standards With Reservations*.

I. CRITERIA FOR WHETHER RCT STUDIES ARE ELIGIBLE FOR REVIEW UNDER CACE STANDARDS

To be eligible for review, a CACE estimate from an RCT must meet several technical criteria. To specify these technical criteria, it is necessary to define some key terms, as discussed next.

A. Key Terms

We refer to the following commonly accepted terms from the econometric literature on instrumental variables:

- **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 study participants' 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 induce variation in the endogenous independent variable 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 group to which study participants 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 ought to 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.

B. Technical Eligibility Criteria

To be eligible for review under the CACE guidance, a CACE estimate from an RCT must be based on statistical methods that meet all of the conditions below.

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 of the possibility 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 must have exactly one endogenous independent variable. With multiple endogenous independent variables, criteria for evaluating instrument strength (see Stock and Yogo, 2005) would require matrix algebraic quantities that are rarely reported in education evaluations, and the WWC does not request authors to conduct new analyses.[4]

The instrumental variables must be binary indicators for the groups (intervention and comparison groups) to which study participants 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, a study could construct site-specific indicators for being assigned to the intervention group (formed, for example, by interacting a single intervention assignment indicator with site indicators), which then serve as the instruments. The use of site-specific intervention assignment indicators allows the first-stage equation to model variation across sites in the extent to which assignment to the intervention group influences take-up.[5] 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 group that cannot receive the intervention—to

[4] 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 and Donald, 1993; Stock and Yogo, 2005; Sanderson and Windmeijer, forthcoming). 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, the WWC would not be able to calculate it unless the WWC had the individual-level data used for the evaluation.

[5] When site-specific intervention assignment indicators serve as the instrumental variables, both the first-stage and structural equations must control for site indicators so that intervention and comparison group members are being compared only within the same site. 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.

which each study participant could be randomly assigned. In this scenario, the instrumental variables are binary indicators for all but one of the assignment groups.[6]

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, 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 (TSLS) or a method that produces the same estimate as TSLS. In TSLS, 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 (OLS), 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 OLS. From this second stage, the estimated coefficient on the predicted take-up variable is equivalent to the TSLS estimate of the CACE (and the standard error of the coefficient must be adjusted to account for the first-stage prediction, as discussed later).

When there is only one instrument, the TSLS estimate is the same as a ratio in which the numerator is the ITT 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 ITT estimate divided by the take-up rate in the intervention group. It is equivalent to the TSLS estimator when (1) there is no take-up in the comparison group, and (2) 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.[7]

Although TSLS is the most widely used approach to CACE estimation, other methods exist. Alternative methods include (1) limited information maximum likelihood (Anderson and Rubin, 1949); (2) generalized method of moments (Hansen, 1982); and (3) missing-data methods based on Bayesian procedures or the EM algorithm (Imbens and Rubin, 1997a). Because these methods have not been used frequently in education evaluations, we have not proposed standards that apply to these methods.

[6] In all of 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.

[7] 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.

II. REPORTING OF CACE ESTIMATES IN WWC PRODUCTS

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

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

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

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

III. OVERVIEW OF PROCEDURES FOR RATING CACE ESTIMATES

A CACE estimate from an RCT is evaluated on a different set of criteria, depending on whether the RCT has low or high attrition:

- **A CACE estimate from an RCT with low attrition** *meets WWC group design standards without reservations* if it satisfies two conditions: (1) no clear violations of the exclusion restriction, and (2) sufficient instrument strength.[8] It *does not meet WWC group design standards* if at least one of those conditions is not satisfied.

- **A CACE estimate from an RCT with high attrition** *meets WWC group design standards with reservations* if it satisfies three conditions: (1) no clear violations of the exclusion restriction, (2) sufficient instrument strength, and (3) demonstration of baseline equivalence. It *does not meet WWC group design standards* if at least one of those conditions is not satisfied.

The following sections provide details on the procedures for assigning ratings to CACE estimates. Section IV describes the method for determining whether an RCT has low or high attrition when rating CACE estimates. Sections V and VI then describe the procedures for rating CACE estimates from RCTs with low and high attrition, respectively.

IV. CALCULATING ATTRITION WHEN RATING CACE ESTIMATES

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, Section B of the *Procedures and Standards Handbook*). In particular, both overall and differential attrition must be calculated. Table III.1 of the *Handbook* will then 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 than the method used when rating ITT estimates. When rating ITT 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 ITT estimates, because those estimates are intended to represent how assignment to the intervention would, on average, affect *all* study participants.

[8] 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. 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 study participant's outcome is unaffected by the treatment status of other participants. Therefore, these standards assume that monotonicity is satisfied.

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 earlier), the attrition rates for compliers likewise cannot be directly calculated. Instead, the attrition rates must be estimated on the basis of specific assumptions, discussed below.

For the usual scenario in which there are two assigned groups—the intervention group (denoted by $Z = 1$) and the comparison group (denoted by $Z = 0$)—the differential attrition rate for compliers, $\Delta^{complier}$, will be estimated as

$$(1) \qquad \hat{\Delta}^{complier} = \frac{\bar{A}_{1,ran} - \bar{A}_{0,ran}}{\bar{D}_{1,ran} - \bar{D}_{0,ran}}$$

where $\bar{A}_{z,ran}$ 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 (1) is the differential attrition rate that the WWC calculates when rating ITT estimates, and the denominator is the difference in take-up rates between assigned groups. Equation (1) 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 (1) 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 non-zero) 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.[9]

To calculate the overall attrition rate for compliers, we will calculate the attrition rate for compliers in the intervention and comparison groups separately and then take 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, $R_0^{complier}$, will be estimated as[10]

[9] 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, the WWC will not have all of the information needed for calculating the denominator of equation (1). In this case, we assume 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 (1) and therefore, leads to an upper bound for the differential attrition rate.

[10] The intuition behind equation (2) 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 [2]), we 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 [2]). The resulting difference is an estimate of the attrition rate for comparison-group compliers.

(2) $$\hat{R}_0^{complier} = \frac{(1-\bar{D}_{0,ran})\bar{A}_{00}-(1-\bar{D}_{1,ran})\bar{A}_{10}}{\bar{D}_{1,ran}-\bar{D}_{0,ran}}.$$

The attrition rate for compliers in the intervention group, $\hat{R}_1^{complier}$, will then be estimated as

(3) $$\hat{R}_1^{complier} = \hat{R}_0^{complier} + \hat{\Delta}^{complier}.$$

The overall attrition rate, $\hat{R}_{overall}^{complier}$, will then be calculated as

(4) $$\hat{R}_{overall}^{complier} = \frac{\hat{R}_1^{complier} N_1 + \hat{R}_0^{complier} N_0}{N_1 + N_0}$$

where N_1 and N_0 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 estimating a TSLS 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 (rather than the comparison group) serves as the instrumental variable. The estimated coefficient on the take-up indicator is equivalent to the differential attrition rate shown in equation (1).

If there are three or more groups to which each sample member could be randomly assigned, the procedure we will follow is likewise equivalent to estimating a TSLS regression in which attrition is the outcome, a take-up indicator is the endogenous independent variable, and a set of assigned group indicators (one for each group except an omitted reference group) constitute the instrumental variables. In this procedure, we will first order the assigned groups from lowest to highest take-up rate. For each comparison between consecutively ordered groups, we will apply equations (1) through (4) to obtain differential and overall attrition rates for compliers relevant to that comparison—that is, for study participants who are induced to take up the intervention by being assigned to the higher-ordered group instead of the lower-ordered group. We will then take a weighted average of both the overall and differential attrition rate across those different comparisons, with weights specified in Imbens and Angrist (1994). Appendix B provides formulas for those weights.

V. PROCEDURES FOR RATING CACE ESTIMATES WHEN ATTRITION IS LOW

A CACE estimate from a low-attrition RCT *meets WWC group design standards without reservations* if it satisfies two criteria: (1) no clear violations of the exclusion restriction, and (2) sufficient instrument strength. If at least one of those criteria is not met, the CACE estimate *does not meet WWC group design standards.* Next, we describe the two criteria in detail.

A. Criterion 1: No Clear Violations of the Exclusion Restriction

Conceptual background. 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"—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, the implied violation of the exclusion restriction is also a confounding factor that, under current WWC group design standards, would cause a study to be rated *Does Not Meet WWC Group Design Standards*.

One scenario that does not represent a confounding factor in ITT studies would be a violation of the exclusion restriction. 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 were defined as enrolling in the intervention being studied (such as an intensive after-school program), whereas take-up in the assigned comparison group were 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 study participants 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 study participants to search

for a job to avoid having to participate in the intervention services. In this case, the intervention assignment might have effects on 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.

Criterion for the WWC. For a CACE estimate to have no clear violations of the exclusion restriction, 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 study participants even if they do not take up the intervention.

B. Criterion 2: Sufficient Instrument Strength

Conceptual background. The condition of sufficient instrument strength requires that the group assignment indicators (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 TSLS estimators to have low type I (false positive) error rates.

The need for sufficient instrument strength stems from the statistical properties of TSLS estimators. An extensive statistical literature has demonstrated that, in finite samples, TSLS, estimators of CACE impacts include part of the bias of OLS estimates (Richardson, 1968; Sawa 1969; Basmann, 1974; Nelson and Startz, 1990; Buse, 1992; Bound et al., 1995; Bloom et al., 2010).[11] Moreover, in finite samples, TSLS estimators do not have a normal distribution—the distribution typically used to construct confidence intervals. For these reasons, conventional statistical tests—such as *t*-tests and *F*-tests—based on TSLS estimators in finite samples have actual type I error rates that generally are higher than the assumed type I error rates (Stock and 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.

The bias issue with TSLS 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 (1) the association between the instrument and endogenous independent variable is larger, or (2) the association is more precisely estimated. In the context of

[11] As discussed by Bloom et al. (2010), the finite-sample bias of IV 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.

estimating CACE effects, group status is a stronger instrument when group take-up rates differ more and when sample sizes are larger.

Instruments also must be strong enough for statistical tests of TSLS estimators to have "acceptably" low type I error rates. As instruments become stronger, the probability distributions of TSLS estimators converge to normal distributions centered on the true CACE impact. Type I error rates follow suit and converge to their assumed levels. We put "acceptably" in quotes because defining what is acceptable requires its own standard, which we explain below.

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 deviations. We use this precedent to set acceptable type I error rates. In Appendix A, we present 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 0.05, ensuring a bias of less than 0.05 standard deviations implies actual type I error rates of less than 0.10. Thus, the guidelines for instrument strength specified here are based on an upper limit of 0.10 for the type I error rate.

Criterion for the WWC. 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 instruments in the first-stage equation—at least as large as the minimum required level shown in Table 1. 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 0.10 for a *t*-test whose assumed type I error rate is 0.05.[12]

When baseline covariates are included in the 2SLS 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.

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 one instrumental variable (that distinguishes a single intervention group and single comparison group) and no clustering, the WWC can obtain a conservative (lower bound) value for the first-stage *F*-statistic if information is available on the take-up rate for analysis sample members in the intervention group ($\overline{D}_{1,an}$), the take-up rate for analysis sample members in the comparison group ($\overline{D}_{0,an}$), the number of analysis sample members in the intervention group ($N_{1,an}$), and the number of analysis sample members in the comparison group ($N_{0,an}$). The first-stage *F*-statistic is represented as

[12] 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 0.10. See Stock and Yogo (2005) for details.

$$\frac{(\overline{D}_{1,an} - \overline{D}_{0,an})^2}{\frac{\overline{D}_{1,an}(1-\overline{D}_{1,an})}{N_{1,an}} + \frac{\overline{D}_{0,an}(1-\overline{D}_{0,an})}{N_{0,an}}}$$

which is a lower bound because it does not take into account precision gains from controlling for other covariates in the first-stage equation.

Table 1. First-Stage F-Statistic Thresholds for Satisfying the Criterion of Sufficient Instrument Strength

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

Source: Stock and Yogo (2005).

If a CACE estimate does not have an associated first-stage *F*-statistic reported in the study, then the WWC will attempt to obtain it through an author query. If the authors do not provide this statistic upon being queried, then the WWC will try to calculate the first-stage *F*-statistic using the formula above, provided that 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.

VI. PROCEDURES FOR RATING CACE ESTIMATES WHEN ATTRITION IS HIGH

A CACE estimate from a high-attrition RCT *meets WWC group design standards with reservations* if it satisfies three criteria: (1) no clear violations of the exclusion restriction, (2) sufficient instrument strength, and (3) demonstration of baseline equivalence. If at least one of those criteria is not satisfied, the CACE estimate *does not meet WWC group design standards.*

The first two criteria are identical to those discussed in Section V for RCTs with low attrition. The remainder of this section describes the third criterion, demonstration of baseline equivalence.

The baseline equivalence standard for CACE estimates in high-attrition RCTs follows the basic elements of the usual baseline equivalence standard. For each baseline characteristic specified in the review protocol, we will calculate a difference between intervention and comparison group members in the analytic sample. If the reported difference is greater than 0.25 standard deviations in absolute value, then baseline equivalence is not demonstrated. If the difference is between 0.05 and 0.25 standard deviations, the analysis must control for the baseline characteristic in the TSLS regression. Differences of less than or equal to 0.05 require no statistical adjustment (see Table III.2 of the *Handbook*).

However, the specific method for calculating a baseline difference when rating CACE estimates is different than the usual method used when rating quasi-experimental designs or ITT estimates from high-attrition RCTs. The usual method assesses the degree of imbalance between groups in the *entire* analytic sample. However, for the purpose of rating CACE estimates, it is necessary to assess the degree of imbalance between groups *only among compliers* in the analytic sample.

For each characteristic *X* specified in the review protocol, we will use the following approach to calculate the baseline difference between compliers in the intervention and comparison groups within the analytic sample. Let $\bar{X}_{z,an}$ be the mean of the characteristic for members of the analytic sample with assigned status $Z = z$, and let $\bar{D}_{z,an}$ be the take-up rate among analytic sample members with assigned status $Z = z$. We will estimate the baseline difference among compliers as

$$(5) \qquad \hat{B}^{complier} = (\bar{X}_{1,an} - \bar{X}_{0,an})/(\bar{D}_{1,an} - \bar{D}_{0,an}),$$

and then express this difference in standard deviation units (with standard deviations calculated in the usual way, based on the pooled analytic sample).

The numerator of equation (5) is the baseline difference that the WWC calculates when rating ITT estimates from high-attrition RCTs, and the denominator is the difference in take-up rates between the intervention and comparison groups in the analytic sample. This equation is justified by the same type of assumption that underlies the differential attrition rate calculation in equation (1). Specifically, equation (5) provides a conservative (upper-bound) estimate of the baseline difference for compliers under the assumption that baseline differences for always-takers and never-takers (if non-zero) have the same sign as the baseline difference for compliers.

In fact, because attrition is the key source of bias that can lead to baseline differences in RCTs, assumptions about attrition behavior (from Section IV) shape what types of assumptions about baseline differences are reasonable. Baseline differences emerge when intervention group members who leave the study are different than 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 Section IV, 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, justifying the use of equation (5).

If there are three or more groups to which each sample member could be randomly assigned, we will first order the assigned groups from lowest to highest take-up rate, calculate baseline differences in the analytic sample between compliers of consecutively ordered groups, and take a weighted average of those baseline differences (Imbens and Angrist, 1994). See Appendix B for details.

VII. REPORTING REQUIREMENTS FOR ESTIMATED VARIANCES OF CACE ESTIMATES

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

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

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

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

APPENDIX A: LINKING ESTIMATION BIAS WITH TYPE I ERROR RATES

Here we provide a statistical framework for deriving the relationship between the bias of an impact estimator and the estimator's type I error rate. We focus on a conventional *t*-test. In this framework, 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, we cannot use the distribution of the TSLS estimator, 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, we consider a generic estimator expressed in effect size units, $\hat{\beta}_1^{biased}$. It is distributed normally with expected value equal to $b > 0$ standard deviations when the true impact is zero. The probability of a type I error using a 5-percent significance test is

$$\begin{aligned}\text{Type I Error Rate} &= \Pr\left(\frac{\hat{\beta}_1^{biased}}{SE(\hat{\beta}_1^{biased})} > z_{0.975}\right) + \Pr\left(\frac{\hat{\beta}_1^{biased}}{SE(\hat{\beta}_1^{biased})} < z_{0.025}\right) \\ &= \Pr\left(\frac{\hat{\beta}_1^{biased} - b}{SE(\hat{\beta}_1^{biased})} > z_{0.975} - \frac{b}{SE(\hat{\beta}_1^{biased})}\right) + \Pr\left(\frac{\hat{\beta}_1^{biased} - b}{SE(\hat{\beta}_1^{biased})} < z_{0.025} - \frac{b}{SE(\hat{\beta}_1^{biased})}\right) \\ &= 1 - \Phi\left(z_{0.975} - \frac{b}{SE(\hat{\beta}_1^{biased})}\right) + \Phi\left(z_{0.025} - \frac{b}{SE(\hat{\beta}_1^{biased})}\right)\end{aligned} \quad \text{(A.1)}$$

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 (A.1) 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, we must pick a value for the standard error. 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 (A.1).

As the benchmark, we assume a level of precision corresponding to a study for which the minimum detectable effect size (MDES) is 0.25 standard deviations, the WWC threshold for substantively important effects. A value for MDES, 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 (SE), as follows (see Bloom, 2004):

$$MDES = \left[\Phi^{-1}(1-0.05/2) + \Phi^{-1}(0.8)\right] \times SE = 2.802 \times SE \quad \text{(A.2)}$$

Using equation (A.2), a study designed to have an MDES of 0.25 is expected to have a standard error of 0.09 standard deviations (= 0.25 / 2.802).

By substituting the benchmark standard error, 0.09 standard deviations, for $SE(\hat{\beta}_1^{biased})$ in equation (A.1), we completely specify the relationship between the type I error rate and the amount of bias. Equation (A.1) becomes

$$\text{Type I Error Rate} = 1 - \Phi(z_{0.975} - b/0.09) + \Phi(z_{0.025} - b/0.09) \quad \text{(A.1')}$$

The final step is to substitute into equation (A.1') a value for *b* that represents the maximum tolerable bias. As discussed earlier in the document, the maximum value for *b* that is acceptable to the WWC is 0.05 standard deviations. Setting *b* = 0.05 in equation (A.1'), we obtain a maximum tolerable type I error rate equal to

$$\text{Maximum Tolerable Type I Error Rate} = 1 - \Phi(z_{0.975} - 0.05/0.09) + \Phi(z_{0.025} - 0.05/0.09) = 0.086.$$

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) calculate several different values for the minimum required first-stage *F*-statistic, depending on whether the maximum tolerable type I error rate is 0.10, 0.15, 0.20, or 0.25. For setting the WWC standard, our preceding calculations yield a maximum tolerable type I error rate of 0.086, which we round to 0.10, the closest value addressed by Stock and Yogo (2005). We then use this value to produce values for the minimum required first-stage *F*-statistic based on Stock and Yogo's (2005) calculations.

APPENDIX B: CALCULATING ATTRITION AND BASELINE DIFFERENCES WHEN THERE ARE THREE OR MORE GROUPS TO WHICH EACH SAMPLE MEMBER COULD BE RANDOMLY ASSIGNED

A. Calculating Attrition

Section IV of these standards provided formulas for calculating the overall and differential attrition rate for compliers when there are two assigned groups (the intervention group and comparison group). Here, we consider the scenario in which there are three or more groups to which each sample member could be randomly assigned (for instance, 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, there is still only a single intervention being studied, so there is still only a single measure of take-up—a binary variable for taking up any portion of the intervention.

First, we order the assigned groups with the index $k = 0, 1, 2,\ldots K$ from lowest to highest take-up rate. We also make a monotonicity assumption (Imbens and Angrist, 1994): any sample member who would take up the intervention if assigned to group k would also take up the intervention if assigned to a group ordered after k. For each comparison between group $(k-1)$ and group k, compliers are defined as those who would take up the intervention if assigned to group k but not if assigned to group $(k-1)$. The TSLS estimator of the CACE is a weighted average of complier impacts across these comparisons, with weights given by Imbens and Angrist (1994). Therefore, our method for calculating attrition follows the same approach: we calculate attrition (both overall and differential) for each comparison between consecutively ordered groups and then take a weighted average across those comparisons, using the same weights as those in the TSLS estimator.

Specifically, let $\hat{\Delta}_{k,k-1}^{complier}$ be the differential attrition rate for compliers pertaining to the comparison between groups $(k-1)$ and k, based on applying equation (1). The final differential attrition rate for all compliers, $\hat{\Delta}_{final}^{complier}$, is calculated as

$$\text{(B.1)} \qquad \hat{\Delta}_{final}^{complier} = \frac{\sum_{k=1}^{K} \lambda_k \hat{\Delta}_{k,k-1}^{complier}}{\sum_{k=1}^{K} \lambda_k}$$

where λ_k is the weight on the comparison between groups $(k-1)$ and k. Imbens and Angrist (1994) derive the weight to be

$$\text{(B.2)} \qquad \lambda_k = (\overline{D}_{k,ran} - \overline{D}_{k-1,ran}) \sum_{l=k}^{K} \frac{N_l}{N} (\overline{D}_{l,ran} - \overline{D}_{ran}).$$

where $\overline{D}_{k,ran}$ is the take-up rate for sample members assigned to group k, $\overline{D}_{ran}$ is the take-up rate in the entire randomly assigned sample, N_k is the number of sample members assigned to group k, and N 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.

B. Calculating Baseline Differences

The final calculation of a baseline difference (on a characteristic specified in the protocol) follows a similar approach as that used for calculating attrition. For each comparison between groups ($k - 1$) and k, we use equation (5) to calculate the baseline difference for compliers in the analytic sample. We then take a weighted average of those baseline differences. The weight on each comparison is again specified by equation (B.2), except that all sample sizes and take-up rates are calculated from the analytic sample, not the original randomly assigned sample.

REFERENCES

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.

Angrist, J., Imbens, G., & Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, *91*, 444–472.

Basmann, R. (1974). Exact finite sample distributions for some econometric estimators and test statistics: A survey and appraisal. In M. Intrilligator and D. Kendrick (Eds.), *Frontiers of Quantitative Economics, vol. 2* (pp. 209–288). Amsterdam: 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*, 460–472.

Bloom, H. (1984). Accounting for no-shows in experimental evaluation designs. *Evaluation Review*, *8*(2), 225–246.

Bloom, H. (2004). *Randomizing groups to evaluate place-based programs*. New York: MDRC.

Bloom, H., Zhu, P., & Unlu, F. (2010). *Finite sample bias from instrumental variables analysis in randomized trials*. New York: MDRC.

Bound, J., Jaeger, D., & Baker, R. (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*, 443–450.

Buse, A. (1992). The bias of instrumental variable estimators. *Econometrica*, *60*, 173–180.

Cragg, J., & Donald, S. (1993). Testing identifiability and specification in instrumental variables models. *Econometric Theory*, *9*(2), 222–240.

Greene, W. (2000). *Econometric analysis* (4th ed.). Upper Saddle River, NJ: 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, behavioral assessment* (Vol. 3, pp. 108–127). New York, NY: John Wiley.

Hoover, T. M., Kubina Jr., R. M., & Mason, L. H. (2012). Effects of self-regulated strategy development for POW+TREE on high school students with learning disabilities. *Exceptionality, 20*(1), 20-38.

Imbens, G. W., & Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica, 62*, 467–475.

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.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings*. Oxford University Press.

Nelson, C., & Startz, R. (1990). Some further results on the exact small sample properties of the instrumental variable estimator. *Econometrica*, *58*(4), 967–976.

Richardson, D. H. (1968). The exact distribution of a structural coefficient estimator. *Journal of the American Statistical Association*, *63*, 1214–1226.

Saddler, B. (2006). Increasing story-writing ability through self-regulated strategy development: Effects on young writers with learning disabilities. *Learning Disability Quarterly, 29*(4), 291-305.

Sanderson, E., & Windmeijer, F. (forthcoming). A weak instrument F-test in linear IV models with multiple endogenous variables. *Journal of Econometrics.*

Sawa, T. (1969). The exact sampling distribution of ordinary least squares and two-stage least squares estimators. *Journal of the American Statistical Association*, *64*, 923–937.

Stock, J., & Yogo, M. (2005). Testing for weak instruments in linear IV regression. In J. Stock and D. W. K. Andrews (Eds.), *Identification and inference for econometric models: Essays in honor of Thomas J. Rothenberg* (pp. 80–108). Cambridge: Cambridge University Press.

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.

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