

What Works Clearinghouse Procedures and Standards Handbook, Version 5.0

WWC 2022008
U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Evaluation at IES



WHAT WORKS CLEARINGHOUSE PROCEDURES AND STANDARDS HANDBOOK, VERSION 5.0

March 2022

Statistical, technical, and analysis team (STAT) members

Jack Buckley

AMERICAN INSTITUTES FOR RESEARCH

Terri Pigott

GEORGIA STATE UNIVERSITY

John Ferron

UNIVERSITY OF SOUTH FLORIDA

Allan Porowski

ABT ASSOCIATES

Michael S. Garet

AMERICAN INSTITUTES FOR RESEARCH

James Pustejovsky

UNIVERSITY OF WISCONSIN-MADISON

Russell Gersten

INSTRUCTIONAL RESEARCH GROUP

David Rindskopf

CITY UNIVERSITY OF NEW YORK

Ben B. Hansen

UNIVERSITY OF MICHIGAN

Jessaca Spybrook

WESTERN MICHIGAN UNIVERSITY

Fran Harmon

DEVELOPMENT SERVICES GROUP

Emily Tanner-Smith

UNIVERSITY OF OREGON

Larry V. Hedges

NORTHWESTERN UNIVERSITY

Elizabeth Tipton

NORTHWESTERN UNIVERSITY

Wendy Machalicek

UNIVERSITY OF OREGON

Jeffrey Valentine

UNIVERSITY OF LOUISVILLE

Rebecca Maynard

UNIVERSITY OF PENNSYLVANIA

Elias Walsh

MATHEMATICA

Hiren Nisar

2M RESEARCH

Vivian Wong

UNIVERSITY OF VIRGINIA

Staff

Molly Cain

Sarah Caverly

Alicia Garcia

Natalya Gnedko-Berry

Daniel Hubbard

David Miller

Joshua Polanin

Jordan Rickles

Sarah Sahn

Lisa Shimmel

Joe Taylor

Ryan Williams

AMERICAN INSTITUTES FOR RESEARCH

Danny Swan

Emily Tanner-Smith

UNIVERSITY OF OREGON

Jeffrey Valentine

UNIVERSITY OF LOUISVILLE

Project OFFICERS

Erin Pollard

Betsy Wolf

Jonathan Jacobson

INSTITUTE OF EDUCATION SCIENCE

WWC 2022008

U.S. DEPARTMENT OF EDUCATION

U.S. Department of Education

Miguel A. Cardona
Secretary

Institute of Education Sciences

Mark Schneider
Director

National Center for Education Evaluation and Regional Assistance

Matthew Soldner
Commissioner

Elizabeth Eisner
Associate Commissioner of Knowledge Use

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

This report is in the public domain. Although permission to reprint this publication is not necessary, the citation should be as follows:

What Works Clearinghouse (2022). *What Works Clearinghouse Procedures and Standards Handbook, Version 5.0*. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance (NCEE). This report is available on the What Works Clearinghouse website at <https://ies.ed.gov/ncee/wwc/Handbooks>.

Alternate Formats

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

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

CHAPTER VI. REVIEWING FINDINGS FROM SINGLE-CASE DESIGN STUDIES

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

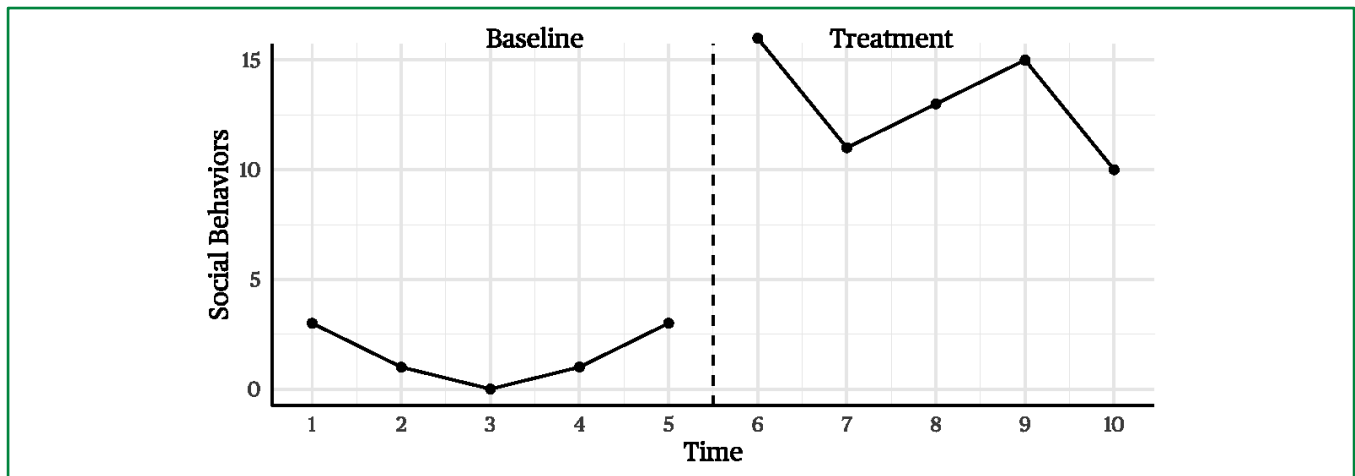
Additional eligibility requirements for SCDs

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

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

[Figure 15](#) displays the simplest form of an SCD, a single individual (case) with one baseline phase and one treatment phase. This simple design is sometimes referred to as an AB design. In SCDs, a phase typically refers to a set of data points from the same condition, observed across time without the interruption of data points from a different condition. When phases are referred to using a string of capital letters, each letter represents a phase from a different condition. For instance, an ABC design would be a design with three phases. The A phase would typically be the baseline phase, the B phase would be an intervention phase, and the C phase would be another intervention phase, either in the form of a modified intervention, an alternative intervention condition, or a maintenance phase. An ABCABC design would be a six-phase design with three conditions that would begin like the ABC design described above, but in the fourth phase it would return to the original baseline A phase, then transition back to another B phase, and finally transition once again to another C phase. Some designs deviate from this typical structure and rapidly alternate interventions within the same experimental phase. See the alternating treatment design section for more information about these designs.

Figure 15. Basic single-case design



In the example in [figure 15](#), the effect of the intervention can be conceptualized as the change in the outcome between the baseline phase (the A phase) and the treatment phase (the B phase). However, most SCD experts consider this simple form of the SCD to have weak internal validity because the effect of the intervention could be due to some other change that co-occurred with the intervention, such as developmental changes for the participant or changes that took place in the classroom that were unrelated to the intervention. Ensuring that SCDs guard against these threats is an important component of the WWC’s SCD standards.

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

Reviewing findings from SCDs according to WWC standards

The process for reviewing SCD studies that are found eligible for review is presented in [figure 16](#). After a study is found eligible for a WWC review as an SCD, the next step is the same as in other designs and includes reviewing the study’s outcome measures and checking for confounding factors. If none of the outcome measures are consistent with the WWC’s requirements or if the study contains a confounding factor, the study will receive a research rating of *Does Not Meet WWC Standards* and the review will stop.

Outcome measure requirements

The WWC’s outcome measure requirements for SCDs are similar to those for randomized controlled trials (RCTs) and quasi-experimental designs (QEDs) described in [Chapter III, Outcome measures](#), including (1) face validity, (2) reliability, (3) not overaligned with the intervention, (4) consistent data collection procedures, and in some instances (5) the independence of the outcome measure. Differences in these requirements for SCDs are described next.

Requirement 1: Face validity

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

Requirement 2: Reliability

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

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

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

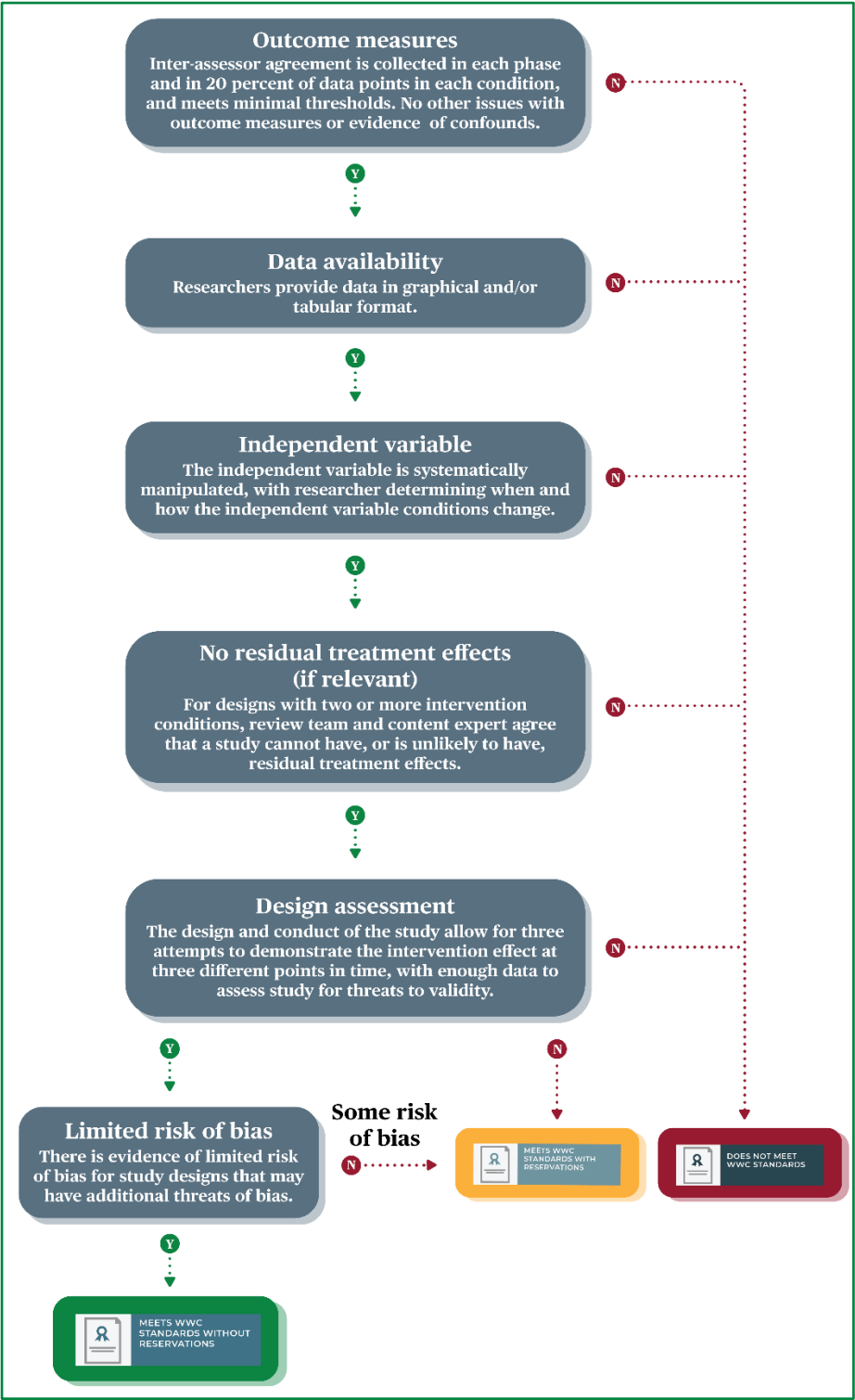
If a study contains measures that are not direct observations of behavior, such as a test of an academic outcome, then the reliability requirements for these measures will follow the guidelines in [Chapter III, Outcome measures](#).

Requirement 3: Not overaligned

Overalignment occurs when an outcome measure contains content or materials provided to the cases in one condition but not another. This rule does not apply when material covered by an outcome measure must be explicitly taught, or when an outcome measure is broadly educationally relevant. Content experts can provide advice on whether an outcome has broad educational relevance. These two caveats to the overalignment requirement are particularly important to SCDs, which frequently focus on narrow, specific outcomes that may require explicit teaching, or on daily-living outcomes with educational relevance. The functional skills domain from the [Study Review Protocol](#) contains examples such as dressing, preparing and eating food, or hygiene,

where the researcher might teach the participant a checklist or a set of steps that need to be repeated and the outcome might be some measure of success at repeating the checklist or steps that were taught to the participant.

Figure 16. Single-case design review process for eligible study findings



Requirement 4: Consistent data collection procedures

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

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

Additional consideration: Independence of outcome measure

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

Confounding factors

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

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

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

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

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

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

	Baseline	Intervention
Case 1	Teacher 1	Teacher 1 + <u>Teacher 2</u>
Case 2	Teacher 1	Teacher 1 + <u>Teacher 2</u>
Case 3	Teacher 1	Teacher 1 + <u>Teacher 2</u>

There are similar-appearing circumstances that are not confounding factors.

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

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

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

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

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

Evidence review process for eligible findings from SCDs

To be considered potential evidence of an intervention's effectiveness by the WWC, an SCD must meet four standards: data availability, researcher-manipulated independent variable, no residual treatment effects, and design assessment. To receive a research rating of *Meets WWC Standards Without Reservations*, findings based on some designs must meet requirements for limiting sources of bias. These standards are summarized in [figure 16](#) and detailed in the following sections.

Data availability

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

Data sharing in the form of plots is standard practice in SCD research. Sharing data allows other researchers to perform their own reanalysis in the form of visual analysis, and access to the raw data allows the WWC to assess whether the study meets WWC standards of internal validity for SCDs, as well as allow for effect size estimation when appropriate. If the data are not available in graphical or tabular format, then the study will receive a research rating of *Does Not Meet WWC Standards*.²⁹

Researcher-manipulated independent variable

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

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

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

Residual treatment effects

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

Residual treatment effect

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

When a review team identifies an eligible alternating treatment design experiment that uses two or more interventions, the review team should ask the content expert to assess whether residual treatment effects are likely given the specific interventions or experimental conditions, the timing of the and length of observation sessions, the order of the interventions or experimental conditions, and the outcomes in the experiment. The review team can rely on previous approval of similar conditions and outcomes from the content expert, but the plausibility of residual effects should not be solely informed by the data reported in a given study. The review team will then assign the study for review and pass along the content expert determination to the reviewers.

²⁹ When data are available only in graphical format, the WWC will extract tabular data from the plots. The WWC intends to explore ways of making the extracted tabular data available for researchers interested in performing confirmatory analyses or syntheses of findings from studies reviewed by the WWC.

Reviewers should raise any additional concerns they have about residual treatment effects as part of their reviews. Reviewers should focus on the plausibility of residual treatment effects based on theoretical and contextual considerations given the research design and intervention characteristics but should not raise concerns based on data reported in the study.

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

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

Research design requirements

The primary goal of the SCD research design requirements is to ensure that the study was designed in a way that allows for at least three demonstrations of an intervention effect at three different points in time with reasonable certainty that the observed data are sufficient to capture important information about the pattern of responding. The pattern of responding includes information such as the within-phase mean, the within-phase variability, and any increasing or decreasing trend that might be present.

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

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

When there are a sufficient number of opportunities to demonstrate an intervention effect, but a limited number of data points are available, a study will receive a lower rating. However, in some cases there are a small number

of data points but it is still possible to be reasonably certain that the observed data points capture all the necessary information about the phase.

Any phase with three or more data points and no within-phase variability represents enough data that reviewers can be reasonably certain that additional data points would not provide additional information. [Figure 17](#) provides some example baseline phases with both zero and low variability. Although data points at the scale minimum or scale maximum are not the only contexts in which there might be zero within-phase variability, these are likely the most common scenarios where zero within-phase variability will be observed.

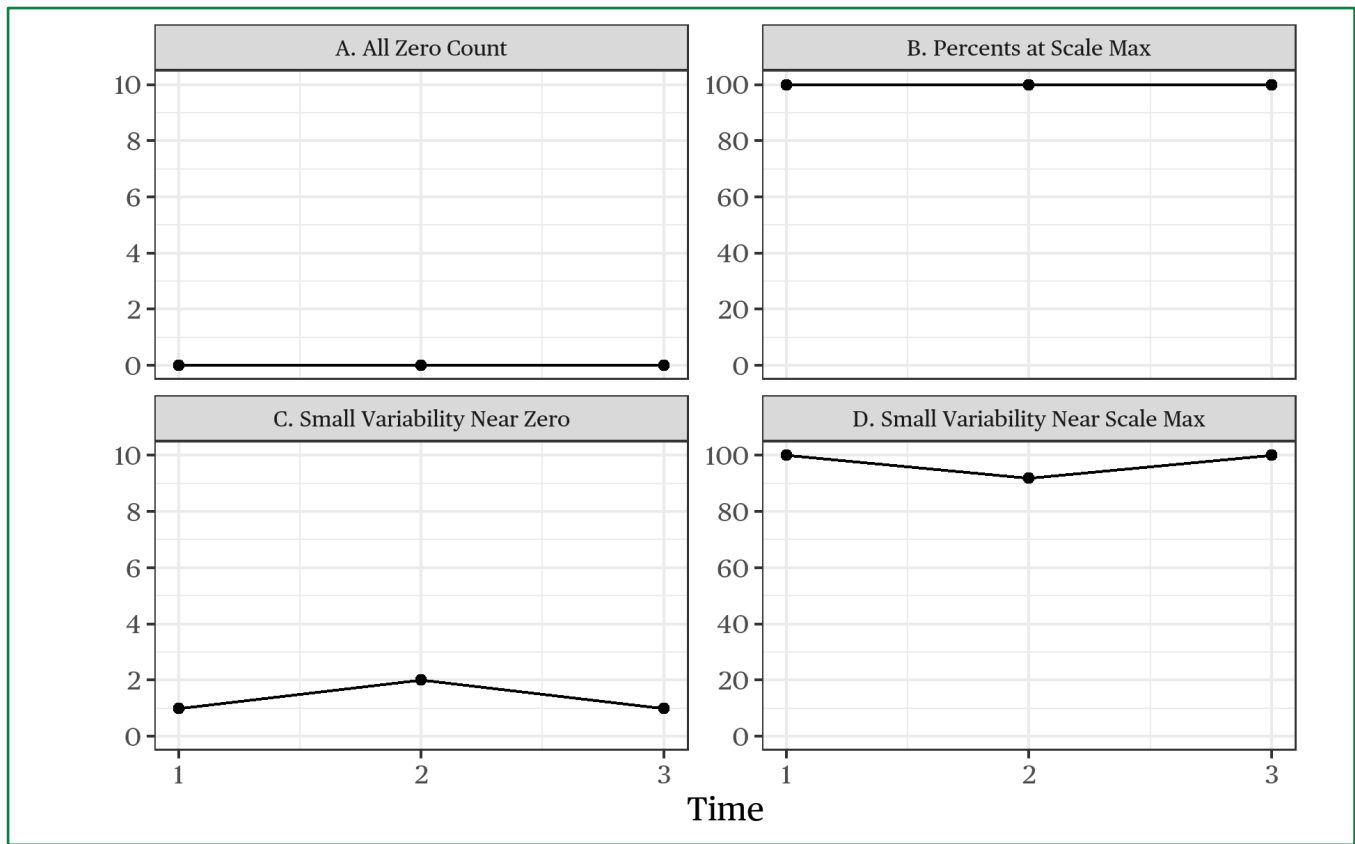
Panels A and B show plots with three data points at the scale minimum and scale maximum, respectively. There is no variability within each phase, and reviewers can be reasonably certain that if the researcher had gathered another data point, it would contain no additional information about the average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

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

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

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

Figure 17. Zero- and low-variability baseline examples



Reversal/withdrawal designs (AB^K)

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

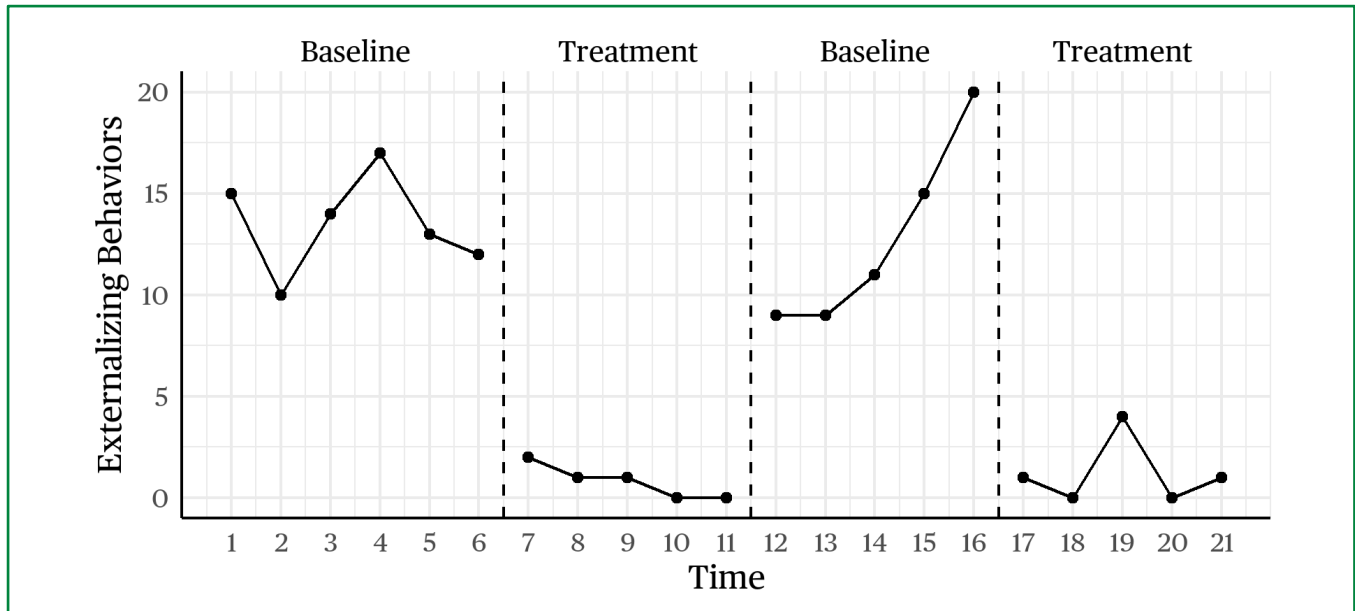
Data points per phase. For treatment reversal/withdrawal design findings to be eligible to be rated *Meets WWC Standards Without Reservations*, the first baseline phase must have at least six data points,³⁰ and at least two phases per condition must have five or more data points per phase. In the simplest design that compares a baseline and intervention condition, this corresponds to an initial baseline phase with six or more observations, a treatment phase with five or more observations, a return-to-baseline or withdrawal phase with five observations, and a second treatment phase with five or more observations. Additionally, any phase with three or more data points and zero within-phase variability also will count toward the required phases for a finding to be eligible to be rated *Meets WWC Standards Without Reservations*, including initial baseline phase.

³⁰ A small group of applied and methodological SCD experts agreed that requiring more data in the baseline phase to ensure that reviewers could make accurate judgments was a reasonable requirement.

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

[Figure 18](#) provides a simple reversal/withdrawal design example eligible to receive a research rating of *Meets WWC Standards Without Reservations* due to meeting the number of phases and number of data points per phase requirements for the reversal/withdrawal design standards. Each participant- or unit-outcome combination in a reversal/withdrawal design is a single experiment, and therefore it should receive separate ratings.

Figure 18. *Reversal/withdrawal design example*



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

Multiple baseline/multiple probe designs

Phases. Multiple baseline designs must have a minimum of six phases split into two conditions for their findings to be rated *Meets WWC Standards Without Reservations* or *Meets WWC Standards With Reservations*. The simplest example of this design has three tiers stacked vertically. Each tier is made up of a baseline and treatment phase (in other words, an A to B comparison). Additionally, transitions from the baseline phase to the intervention phase must have at least three unique timings to ensure that there are three opportunities to demonstrate the

intervention effect at three different points in time. Findings with fewer than six phases will be rated *Does Not Meet WWC Standards*.

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

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

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

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

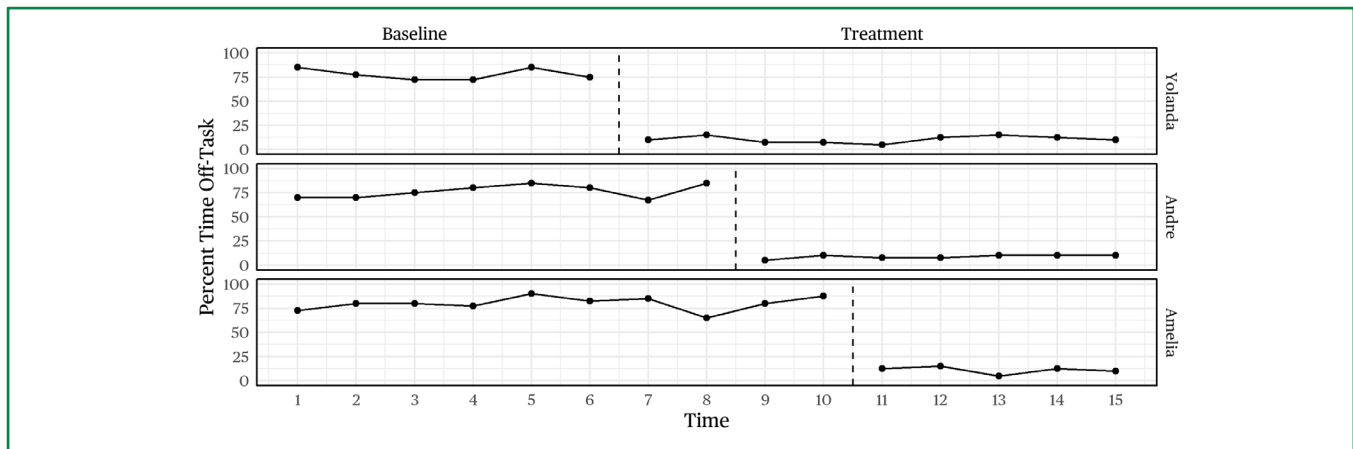
Tiers in multiple baseline designs

A tier refers to a single row in a set of stacked rows in a multiple baseline design. In [figure 19](#), tiers are cases. In that example, the comparisons between baseline and intervention cases can be made within and across cases. Tiers might also be outcomes or contexts within a single individual or case, and effect replication takes place purely within a single case.

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

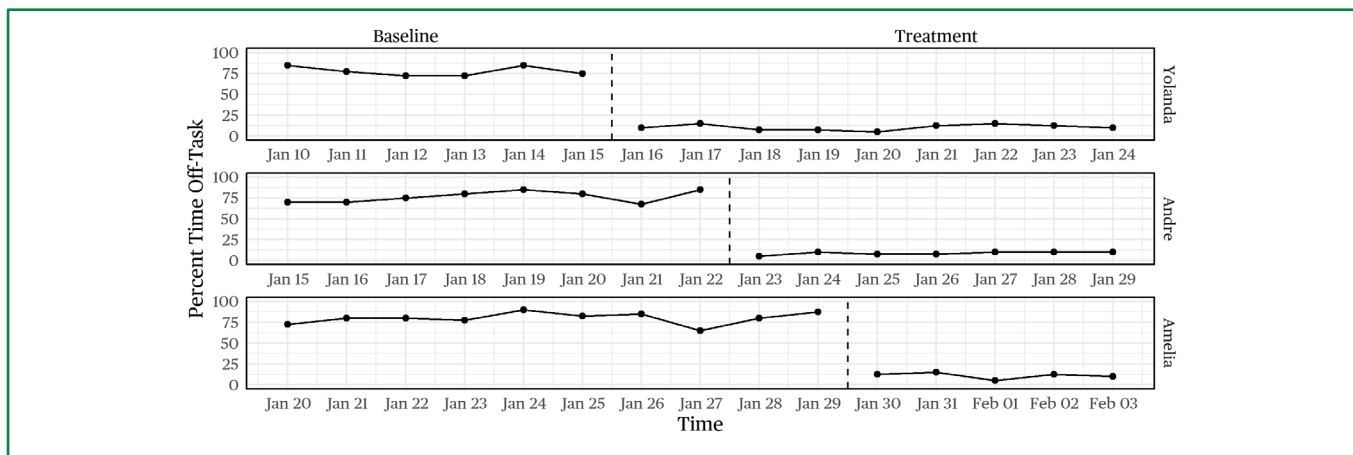
[Figure 19](#) provides an example of a multiple baseline design that is eligible to receive a research rating of *Meets WWC Standards Without Reservations* due to meeting the number of phases, number of data points per phase, and concurrence requirements for multiple baseline designs. Although the y-axis of the plot does not specify the exact date or time of the observations, the WWC generally assumes that authors have aligned their displays in a way that allows for vertical comparison, absent any evidence to the contrary. [Figure 20](#) provides an example with evidence to the contrary. In multiple baseline and multiple probe designs, each stacked plot generally represents a single experiment. Each combination of stacked plot and outcome should receive a separate rating.

Figure 19. Multiple baseline design example



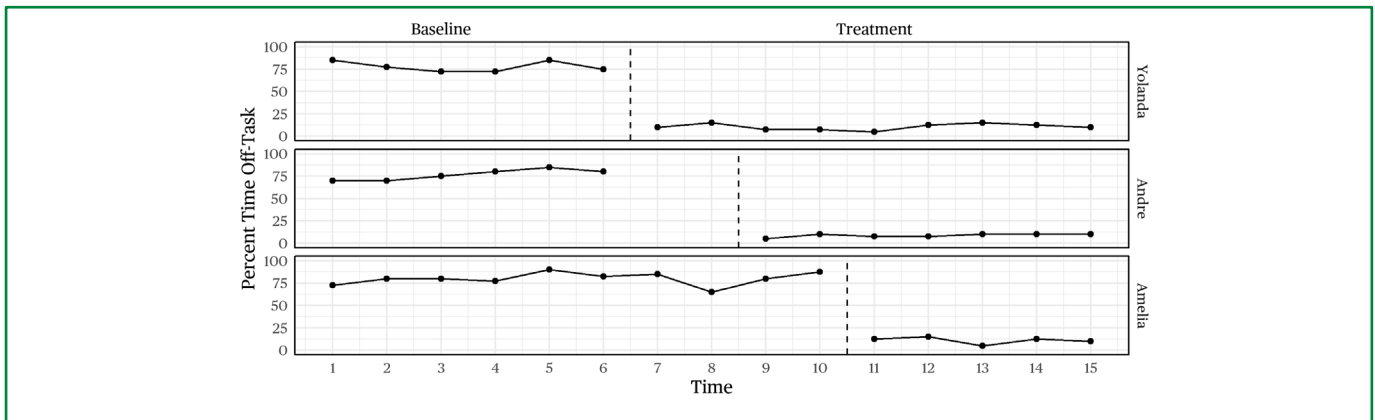
[Figure 20](#) displays an example of a multiple baseline design that does not allow for vertical comparison and fails the requirement that all cases must have data in the baseline phase prior to the introduction of the intervention to any case. Although the first data points for Yolanda, Andre, and Amelia are all arranged in a stacked fashion, the actual timing of those data points are in five-day intervals from each other. Additionally, data points for Amelia do not begin until halfway through Yolanda's treatment phase. This example would be rated *Does Not Meet WWC Standards*.

Figure 20. Example violations of first and second concurrence requirements



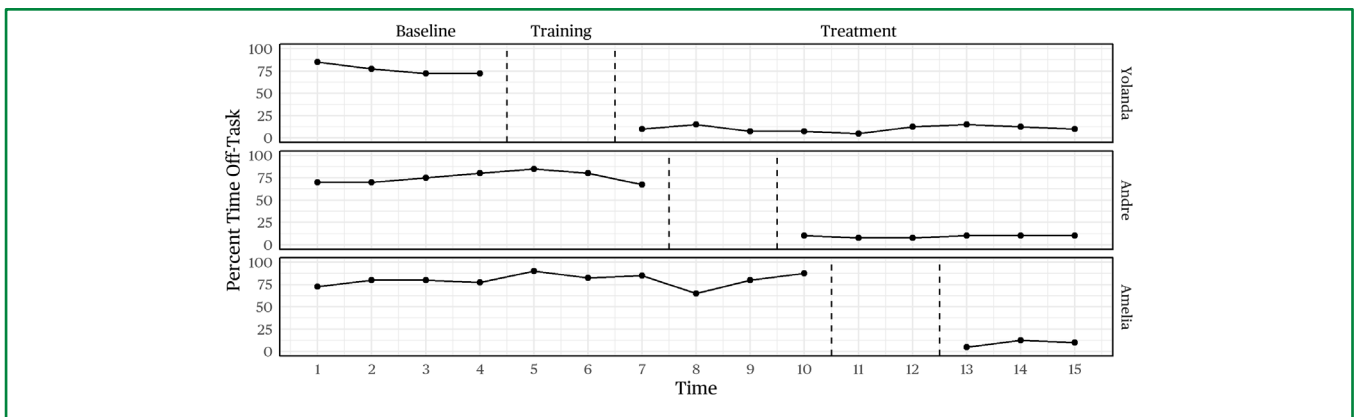
[Figure 21](#) displays an example of multiple baseline design that would not meet the third concurrence requirement. In this example, Andre’s baseline ends after time 6, prior to the onset of the intervention for Yolanda. No data allow for a vertical comparison to ensure that there is no change in Andre’s responses prior to the onset of intervention. The final data points in the baseline phase prior to the onset of the intervention are important for judging any change in the trajectory of the outcome data points, including the WWC’s baseline trend requirement. This example would be rated *Does Not Meet WWC Standards*.

Figure 21. Example violation of the third concurrence requirement



[Figure 22](#) displays an example of a multiple baseline design with empty training phases. These empty training phases are not appropriate for interventions where the training is quick and impact is expected to be immediate. For those types of interventions, the training data would represent the beginning of the treatment phase from the WWC’s perspective and therefore are important to include as a part of the impact estimate. In those cases, this design would be rated *Does Not Meet WWC Standards*. For interventions that require a longer training period to have an impact, empty training phases are acceptable. The requirements for concurrence should ignore the training phase and focus on overlap between the baseline and treatment phases. In cases where the longer training is appropriate, this design would be eligible to receive a rating of *Meet WWC Standards With Reservations*. However, in cases where the training phases last as long or longer than the longest treatment phase in a design, the reviewer should consult a content expert to ensure that the training phases do not constitute an extra condition where data should be available.

Figure 22. Example of empty training phases



Multiple probe design requirements. These designs are a special case of multiple baseline designs. *Planned* missing data is a key element of the multiple probe design and is the major difference between a multiple baseline design and a multiple probe design. Multiple probe designs must meet all the multiple baseline design requirements and additional criteria because baseline data points are intentionally missing.³¹ For multiple probe designs to meet WWC standards, the following must be true:

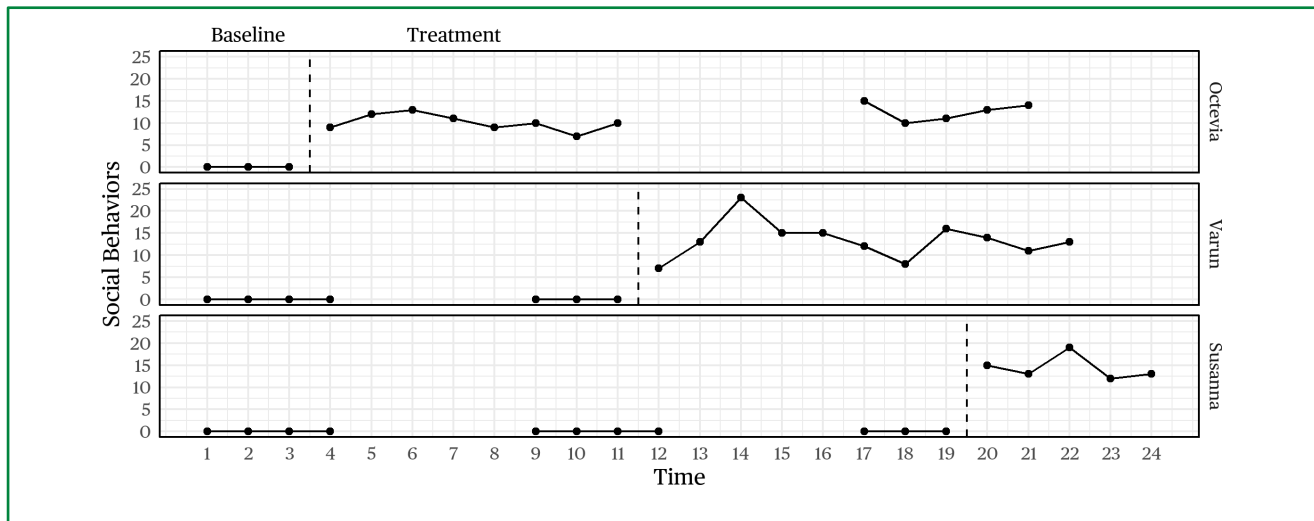
- **Initial preintervention data collection sessions must overlap.** For findings to receive a research rating of *Meets WWC Standards Without Reservations*, each tier must have three data points in the first three sessions. For findings to receive a research rating of *Meets WWC Standards With Reservations*,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

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

[Figure 23](#) displays an example of a multiple probe design that would be eligible to be rated *Meets WWC Standards Without Reservations*. The initial three data collection sessions have data for all cases. All cases have at least three data points directly before intervention begins for any other case. Each case still in the baseline has a data point when the other case first receives the intervention. The baseline phases are all phases with zero variability, and as a consequence three data points are enough data to be rated *Meets WWC Standards Without Reservations*.

³¹ Multiple baseline designs with unintentional missing data should not be reviewed under the multiple probe requirements. Reviewers should note any unplanned missing data in the study review guide.

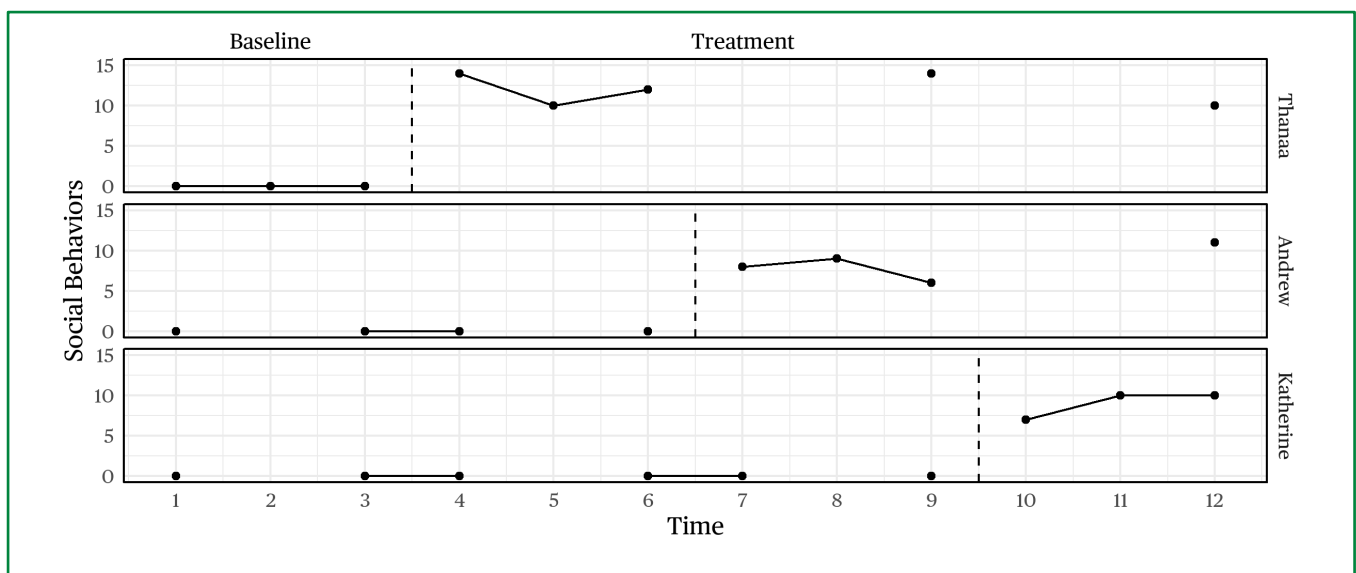
Figure 23. Multiple probe design, example 1



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

[Figure 24](#) displays an example of a multiple probe design that would be potentially eligible for a research rating of *Meets WWC Standards With Reservations*. Each case has a single overlapping data point at time 1 and at time 3, but the study does not have the data point at time 2 for Andrew or Katherine. Each case has at least one data point in the three sessions prior to Thanaa receiving the intervention. Andrew and Katherine each have at least one data point in the three sessions prior to Andrew receiving the intervention. Katherine has at least one data point in the three sessions prior to receiving the intervention. Each case still in the baseline phase has a data point when another case enters the intervention.

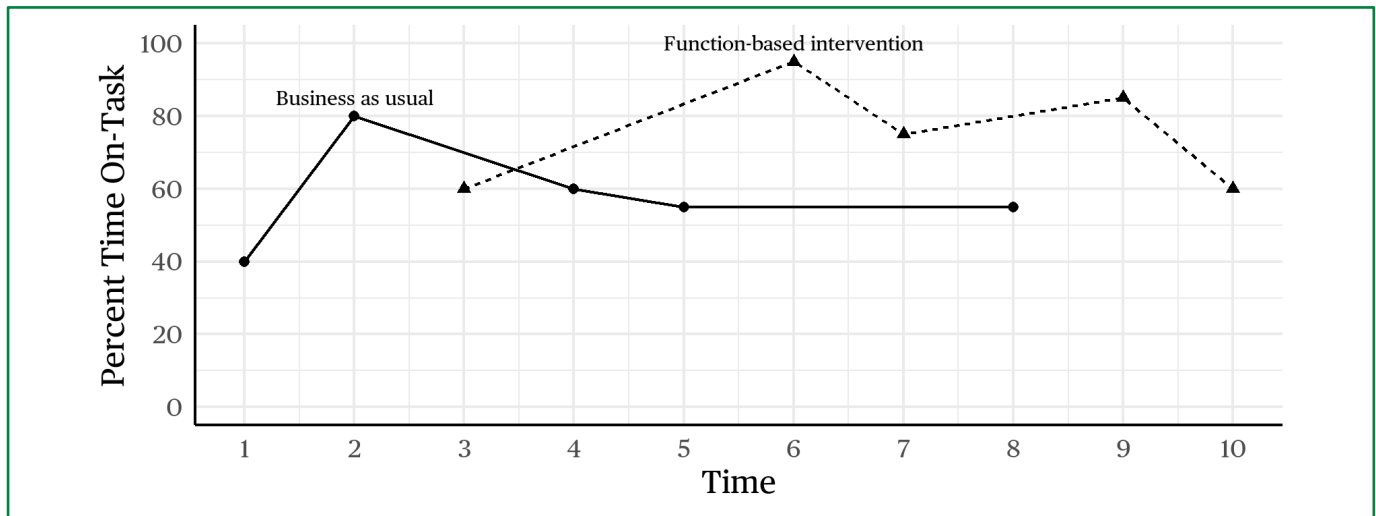
Figure 24. Multiple probe design, example 2



Alternating treatment designs

[Figure 25](#) displays an alternating treatment design example that would be eligible to be rated *Meets WWC Standards Without Reservations* due to meeting the data points per condition and contiguous data points requirement for alternating treatment designs.

Figure 25. Alternating treatment design example



Some alternating treatment designs will contain both a baseline phase and a phase that rapidly alternates between two or more conditions, and other designs will contain only a set of rapidly alternating conditions as seen in [Figure 25](#). Reviews should ignore initial baseline phases included prior to the introduction of the experimental phase in alternating treatment designs, as a comparison between an initial baseline phase and one experimental condition from the rapidly alternating phase will not allow for three demonstrations of the intervention effect at three different points in time.

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

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

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

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

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

Other SCD designs

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

Limited risk of bias

This section is relevant for designs where the primary comparison is for the pattern of responding between separate phases, such as baseline and intervention phases. Of the designs for which the WWC has explicit standards, this includes treatment reversal/withdrawal designs, changing criterion designs, multiple baseline designs, and multiple probe designs. The only design with explicit standards that is not subject to the assessment of limited risk of bias is the alternating treatment design; the contrast of interest is within a rapidly alternating intervention phase rather than between separate phases. Any other designs not explicitly listed in the design standards but reviewed under a set of published professional conventions will be subject to an assessment of limited risk of bias if the design's primary contrast of interest is a comparison between separate phases.

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

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

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

Nonoverlap measures

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

While nonoverlap measures have traditionally been used to describe the effects of interventions, they also can be used to describe the similarity between two sets of data points. For instance, if there is no therapeutic trend in a baseline phase, then the data points at the end of the phase should be similar to the data points at the beginning of the phase. If a behavior is reversible, the pattern of responding in any withdrawal or return to baseline phases in a treatment reversal/withdrawal design should be similar to the initial baseline phase. Higher levels of nonoverlap are associated with less consistency between the data points at the beginning of the phase and the data points at the end of the phase, implying the presence of a trend in the data or incomplete reversibility.

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

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

Minimal therapeutic baseline trend

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

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

Evidence of reversibility

For research designs with return-to-baseline or withdrawal phases and findings that are eligible to receive a rating of *Meets WWC Standards Without Reservations*, reviewers should assess any return-to-baseline or withdrawal phases compared with the initial baseline phase to ensure that minimal reversibility was achieved. Simple multiple baseline designs or multiple probe designs without embedded reversals are not subject to this requirement, nor are alternating treatment designs.

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

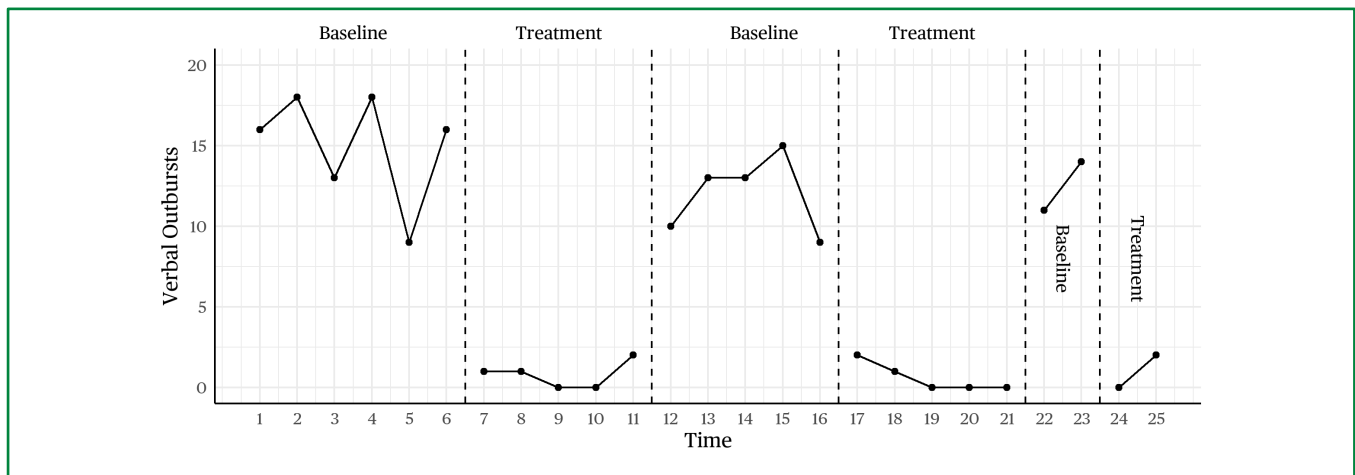
Designs with extra cases/phases or combination designs, or designs with extra conditions

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

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

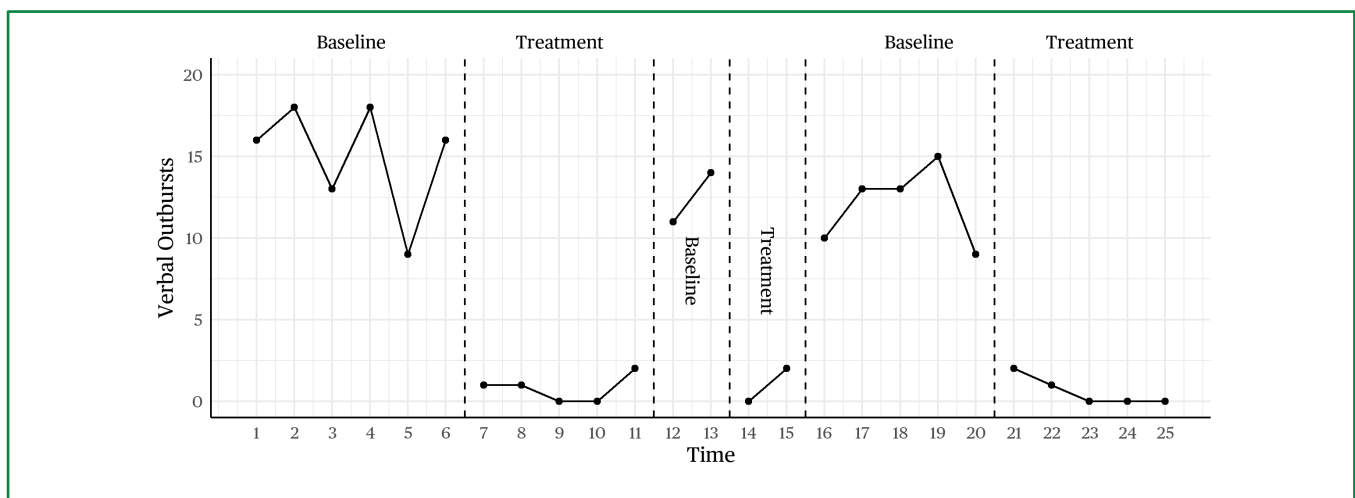
[Figure 26](#) displays an ABABAB reversal/withdrawal example with more than the minimum number of phases to meet standards. The first A phase contains six data points and the subsequent three phases contain five data points each, but the final two phases (the last AB pair) contain only two data points each. This study finding would still be eligible to receive a research rating of *Meets WWC Standards Without Reservations* because it contains enough information to potentially demonstrate the intervention effect at three different points in time within the first four phases.

Figure 26. Treatment reversal design with extra phases that is rated Meets WWC Standards Without Reservations



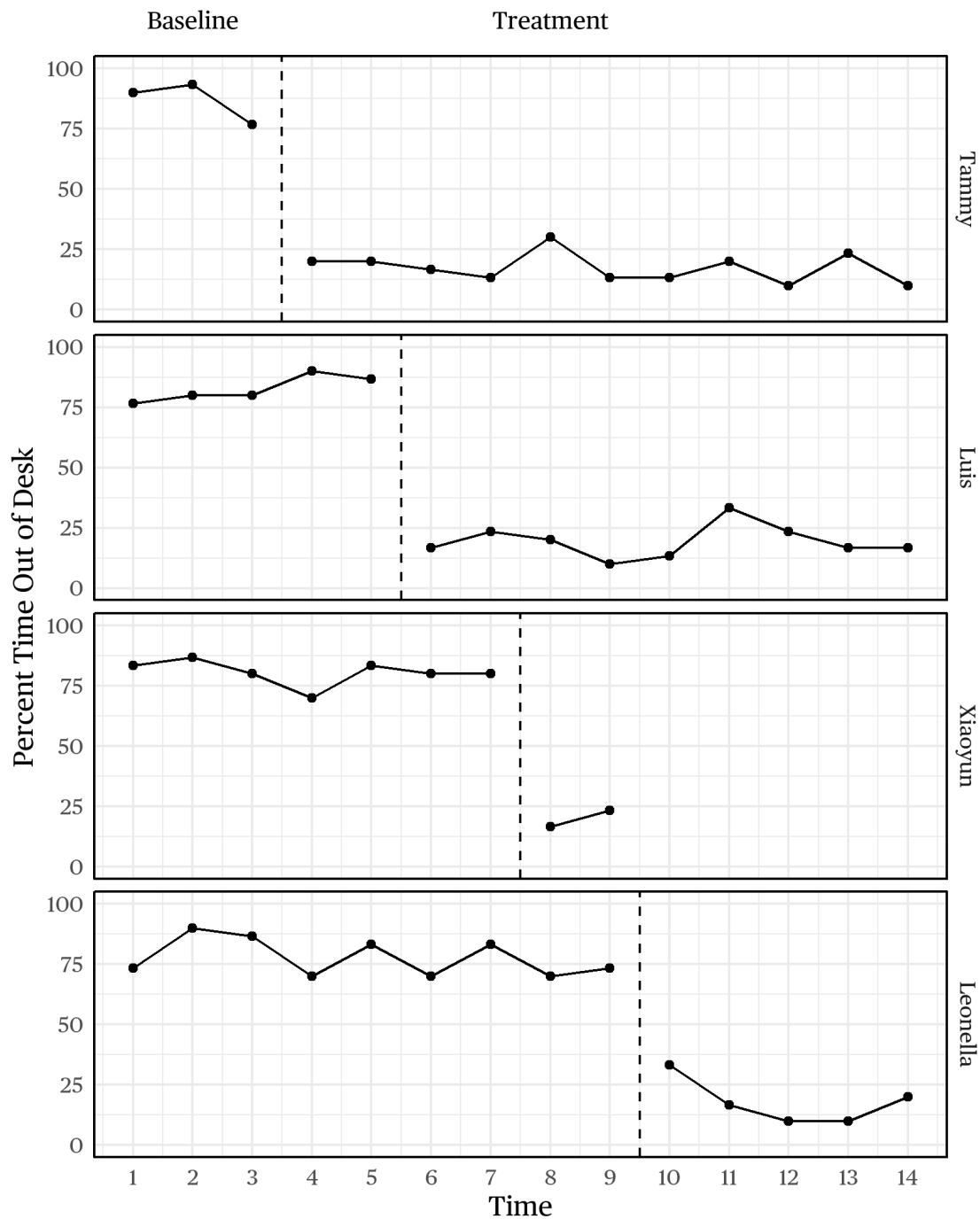
[Figure 27](#) displays another example treatment reversal design with six phases. In contrast to the previous example, this finding would be rated *Does Not Meet WWC Standards*. Although it has four phases that meet the data point requirements for a rating of *Meets WWC Standards Without Reservations*, the two phases in the middle cannot be used as part of a phase transition. Nonsequential phases cannot serve as demonstrations of an intervention's effect. Therefore, only two potential demonstrations of the intervention effect have a sufficient number of data points: one between the first and second phases and one between the fifth and sixth phases.

Figure 27. Treatment reversal design with extra phases that is rated Does Not Meet WWC Standards



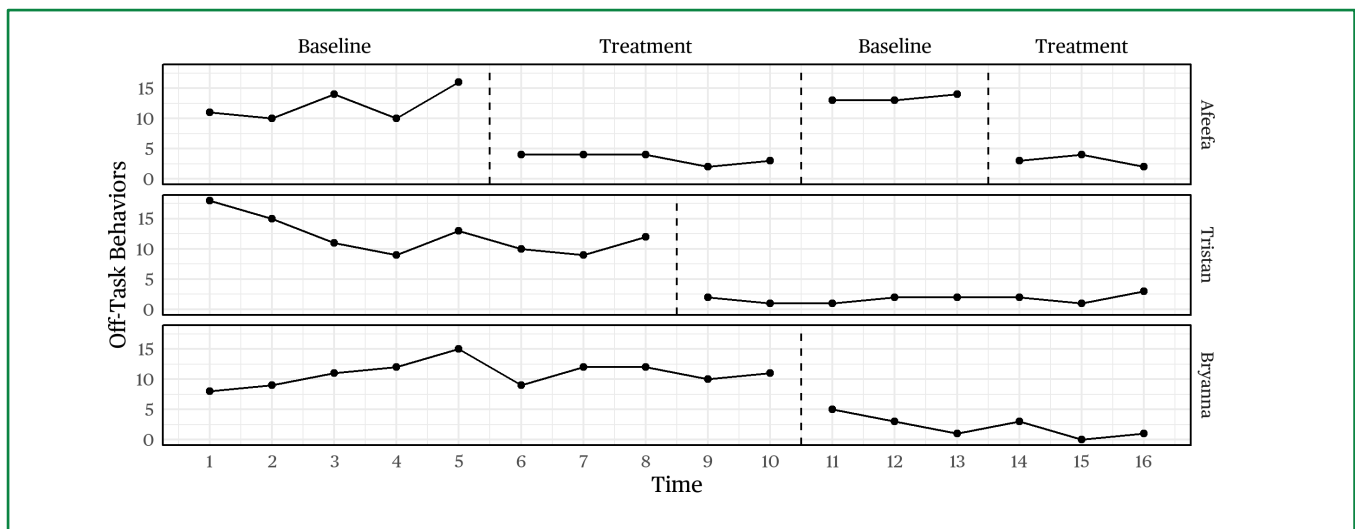
[Figure 28](#) displays a multiple baseline design with four cases that all receive an intervention at staggered times. The first, second, and fourth case all have a sufficient number of data points to be rated *Meets WWC Standards With Reservations*. However, the third case dropped out after only two intervention data points were gathered. This finding would still be eligible to be rated *Meets WWC Standards With Reservations* because it contains enough information to demonstrate the intervention effect at three different points in time.

Figure 28. Multiple baseline design with four cases rated Meets WWC Standards Without Reservations



Combination designs. Findings from combination designs (such as a multiple baseline with embedded reversals) should receive the highest possible rating that any subset of their design is eligible for. [Figure 29](#) displays a multiple baseline design with three cases. The second and third cases only contain the traditional baseline and treatment phase, but the first case also contains an embedded reversal/withdrawal design. The second pair of phases in the first case are brief. These two phases contain only three data points each and therefore are at best eligible to receive a research rating of *Meets WWC Standards With Reservations* under the reversal/withdrawal design requirements. However, given that the subset of initial baseline and treatment phases for all three cases, when considered as a multiple baseline design, would be eligible to receive a research rating of *Meets WWC Standards Without Reservations*, the finding from that combination design should be rated *Meets WWC Standards Without Reservations*.

Figure 29. Combination multiple baseline design with reversals that are rated *Meets WWC Standards Without Reservations*



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