# What Works Clearinghouse™

Reviewer Guidance for Use with the
*Procedures Handbook (version 4.0)* and
*Standards Handbook (version 4.0)*
Updated October 30, 2017

# Contents

# I. Introduction to the What Works Clearinghouse Reviewer Guidance

The *What Works Clearinghouse (WWC) Procedures Handbook (version 4.0) and Standards Handbook (version 4.0)* provide 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 *Handbooks* address 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 Handbook (version 4.0)* and *Standards Handbook* (*version 4.0)*. Rather, the guidance documents clarify how reviews are conducted and how the standards should be implemented in situations where the current *Handbooks* are 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 4.0 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 November 2, 2017. Substantive updates include removing guidance that is no longer needed because of the release of the version 4.0 *Handbooks*, and inserting additional guidance on reviewing studies that use single-case designs.

# II. 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 A of the *WWC Standards Handbook (version 4.0)* 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 *Standards 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 *Standards 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. A-3). The *Standards 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

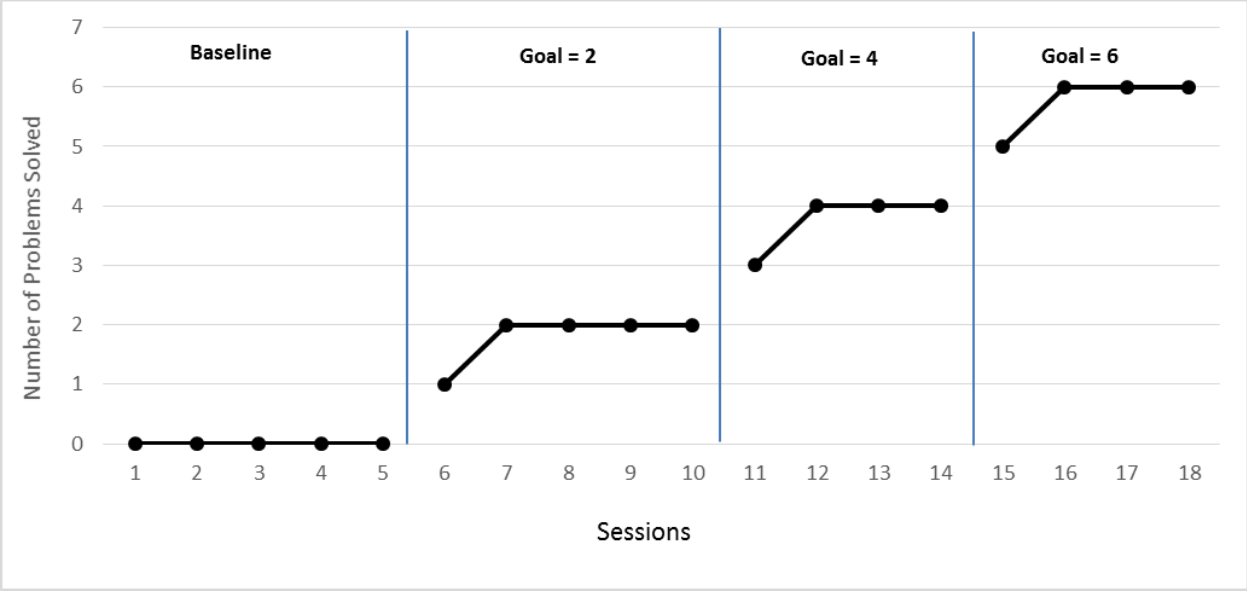
The *WWC Procedures and Standards Handbook (version 2.1)* noted that the changing criterion design is a variant of a reversal-withdrawal (or ABAB) design, "In this design the researcher examines the outcome measure to determine if it covaries with changing criteria that are scheduled in a series of predetermined 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 Pilot Single-Case Design Standards described in Appendix A of the *WWC Standards Handbook (version 4.0)*.

### Guidance

The reversal-withdrawal design standards and visual analysis approach described in Appendix A of the *Standards 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 A of the *WWC Standards Handbook (version 4.0)* 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.

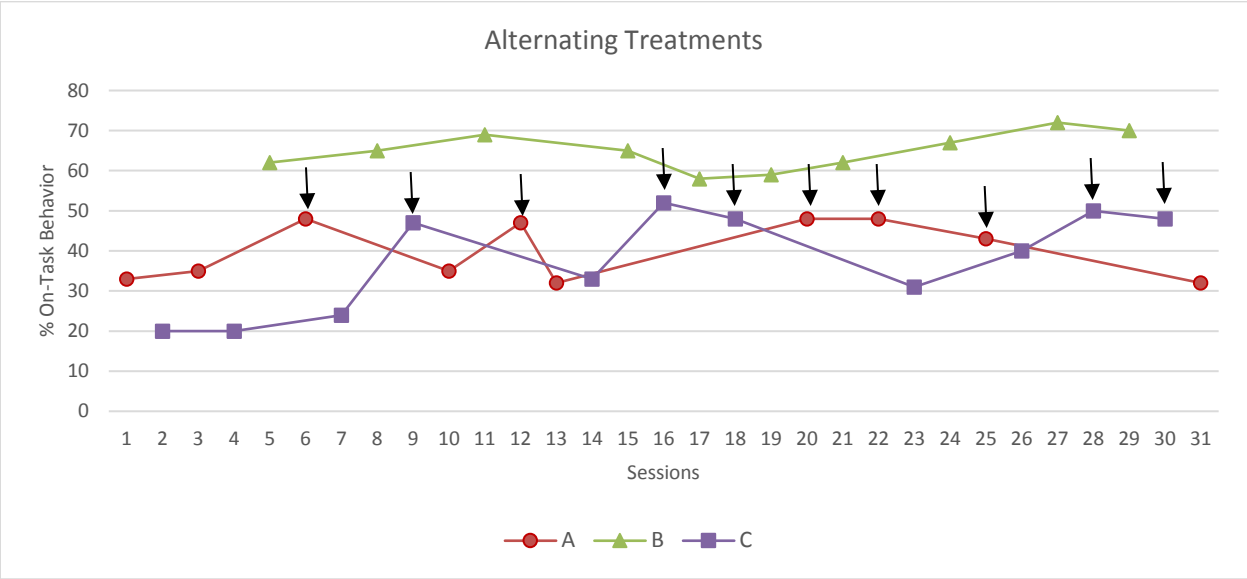
### Guidance

*Residual Treatment Effects*

The *Standards 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 *Standards 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**



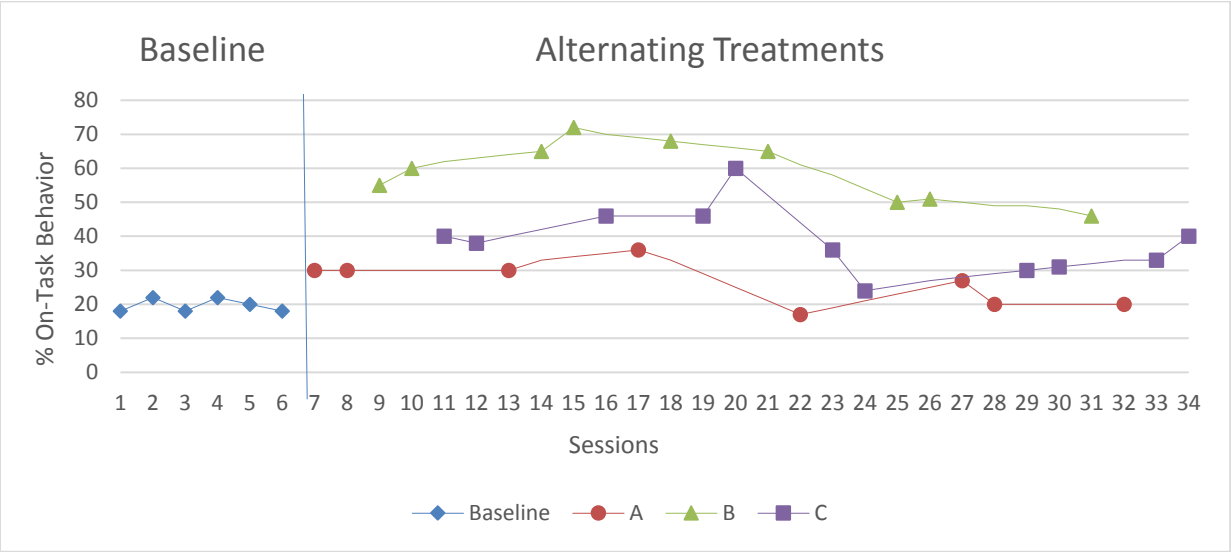

When a review team identifies an eligible AT-design experiment that uses three or more interventions, the review team should ask the content expert to determine whether residual treatment effects are likely given the specific interventions and outcomes in the experiment (the review team can rely on previous 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 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 *Standards 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 *Standards 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 *Standards 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 Standards Handbook (version 4.0)* 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 *Standards 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.

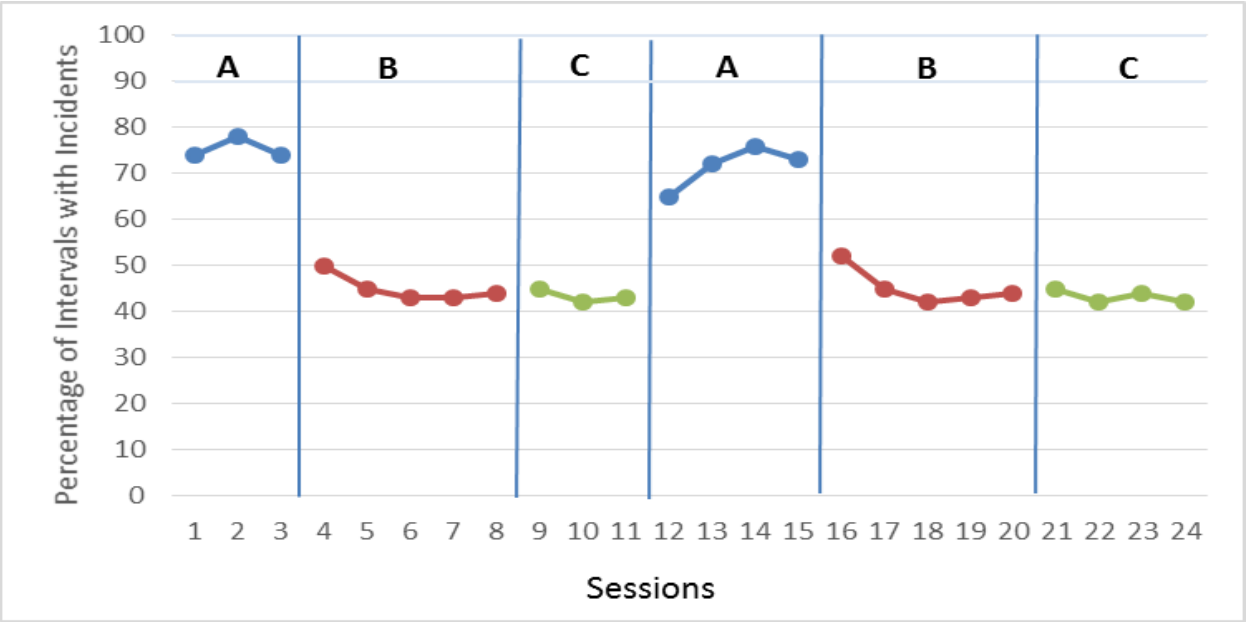
For multiple baseline/probe designs with a third condition, reviewers should review the first two phases. Review team leadership can consider an experiment including the third phase, if: (1) the first two phases do not form an experiment that meet standards with or without reservations, or (2) a comparison 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 A of the *WWC Standards Handbook (version 4.0)* 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 *Standards 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**

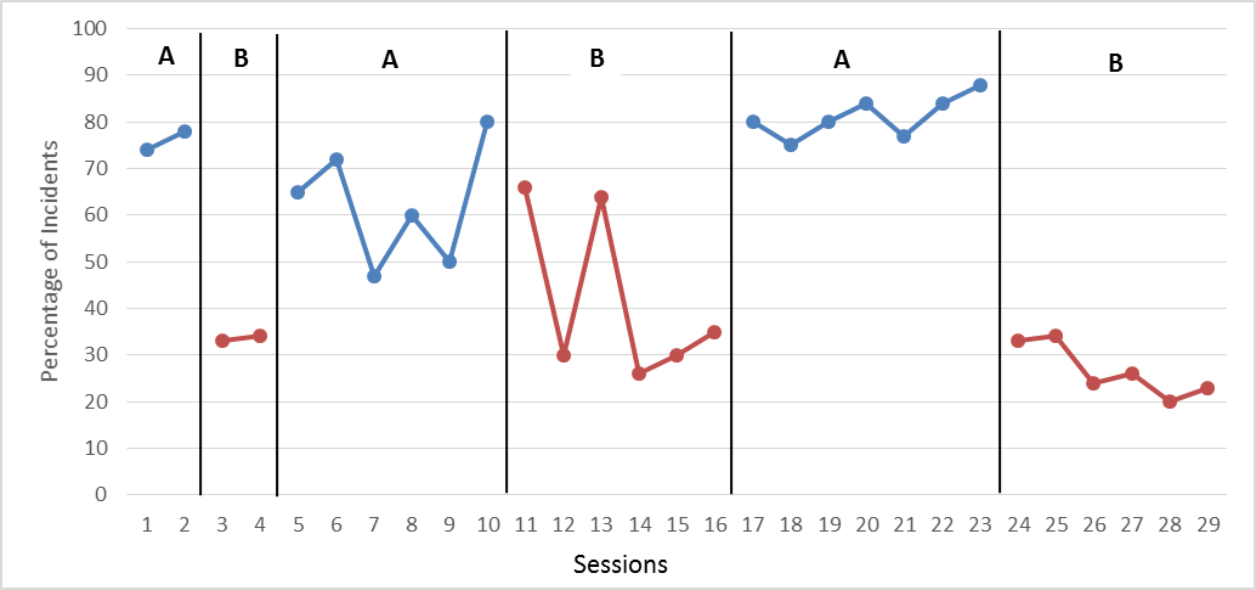
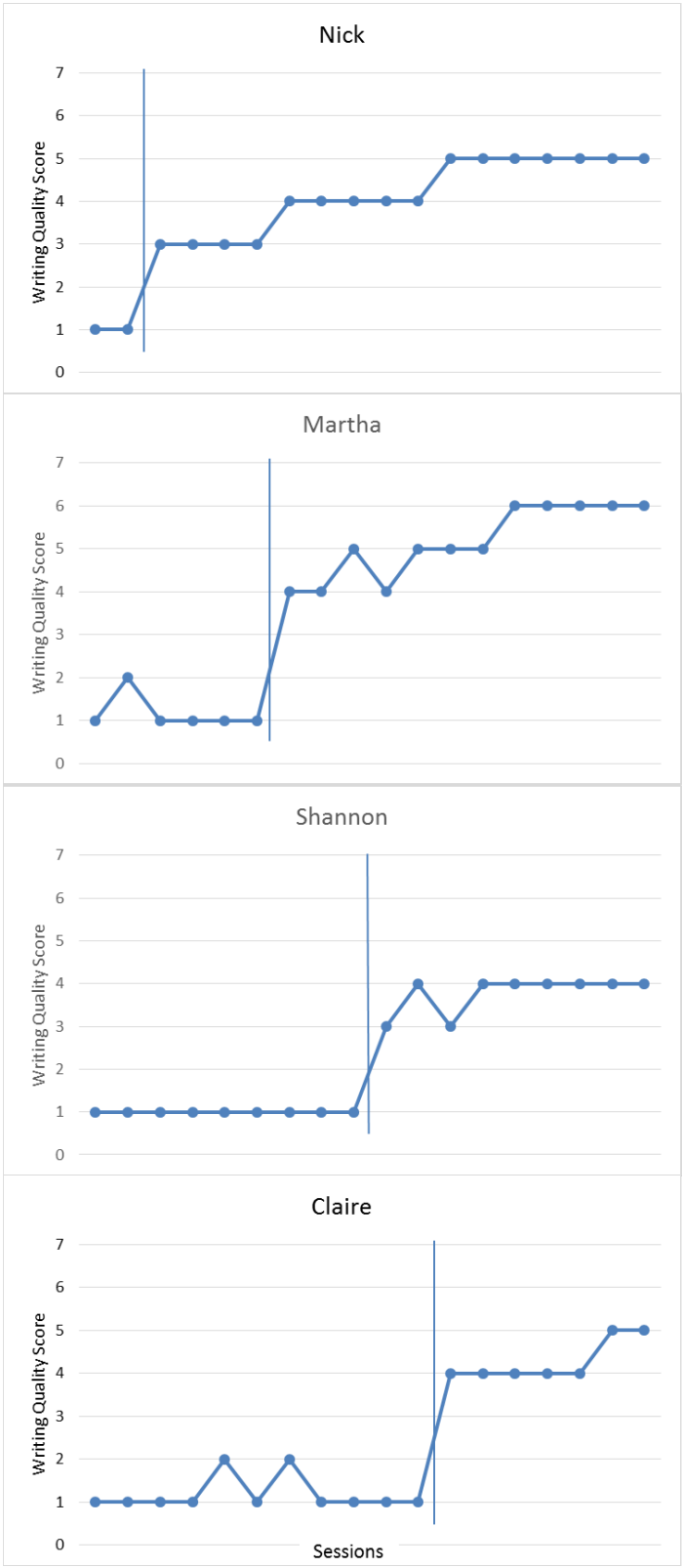


**Figure F.2. Multiple baseline design with four cases**



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

To meet WWC pilot single-case design standards, a single-case design experiment must include at least three attempts to demonstrate an intervention effect *at three different points in time*. In multiple baseline and multiple probe designs, the intervention is introduced to different cases at different points in time (or, in the event of a single case, across different settings at different points in time), resulting in multiple AB data series; these data series are stacked and presented in one figure. Each data series, presented in a single tier of the figure, provides the opportunity to demonstrate one effect of the intervention. When reviewing the full design, comparisons are also made across the data series in each tier, in order determine whether the experiment demonstrates an intervention effect at three different points in time.

The *WWC Standards Handbook (version 4.0)* 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" (p. A-5). The *Standards Handbook* does not have 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. However, the *Standards Handbook* does not: 1) explain how to determine whether the graphical presentations of the tiers present data that was collected at the same time, or 2) specify the exact requirements for "some degree of concurrence." This section provides guidance on these requirements.

**Guidance for determining whether graphical presentations of data show what happened at the same time across tiers**

Time is displayed consistently when the graphical presentations of data show vertically what happened at the same time. Any time period with a data point in one tier should either have a data point at the same time in other tiers or clearly identify the data as missing (if other cases/settings did not have a session). If the graph appears to present time consistently across tiers and nothing in the study text suggests otherwise, reviewers should assume that the graphical display of data shows what happened at the same time for each tier.

- Figures G.1 and G.2 are examples of multiple baseline and multiple probe design experiments, respectively, that illustrate a consistent display of time across all tiers.

- Figure G.3 shows an example of a multiple probe design experiment in which the sessions in the x-axes are labeled differently for each case, but the figure still allows for vertical comparison of what happened at the same time for each case. In this example, the authors are purposely trying to show that specific numbered treatment sessions occurred at different points in time for each student. For example, Kara's 5th session occurred at the same point in time as Wendy's 3rd session and Hannah's 2nd session (rather than their 5th sessions). Unless the study authors provide information to suggest a different interpretation, reviewers should assume that the graphical display of data shows what happened at the same time, even though the numbered sessions do not vertically line up.

If it is clear that the graphical display of data does not show what happened at the same time for each tier, or the reviewer has concerns based on the study text, he or she should raise those concerns with the review team leadership. Studies that do not present time consistently across the tiers 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.

- Figure G.4 shows an example of a multiple baseline design experiment in which the x-axes for each tier are different, and it is clear that the graphical display of data *does not* show what happened at the same time. For example, Jen's data were measured between January 8 and January 15, while Grace's data were measured between January 14 and January 21. As a result, this experiment *Does Not Meet WWC Pilot Single-Case Design Standards* because there are insufficient data to evaluate the attempts to demonstrate an intervention effect.

If the x-axes mostly line up across the tiers, but it is not clear if they entirely line up, due to issues with printing or distorted graphical displays, reviewers should raise any concerns with the review team leadership. An author query may be needed to obtain a consistent display. If a clear graphical display cannot be obtained and concurrence cannot be assessed, the study 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.

**Guidance for determining whether concurrence exists**

Reviewers should assess concurrence once they have determined that the session timing is displayed consistently. For studies relying on multiple baseline and multiple probe designs, reviewers should examine the tiers for cases (or settings) 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 requirement. Cases must also continue to have baseline data (for at least one session) after the intervention is administered to preceding cases.

- For example, consider a multiple baseline design experiment with three cases (see Figure G.1). For this multiple baseline design to have adequate concurrence, baseline data collection for all three cases must begin before Session 6, when Katie first receives the intervention. In addition, Tommy and Steve must have at least one baseline data point after Katie first receives the intervention (after Session 5), and Steve must have at least one baseline data point after Tommy first receives the intervention (after Session 9). In the figure below, both of these requirements are met. However, if such concurrent data were not collected, the design could not exclude threats to internal validity (such as history or maturation) 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. Reviewers should indicate this in the study review guide.

For multiple probe design experiments, the Handbook describes three additional requirements, which may relate to timing and concurrence:

- Initial preintervention sessions must overlap vertically. Within the first three sessions, the design must include three consecutive probe points for each case to *Meet Pilot Single-Case Design Standards Without Reservations* and at least one probe point for each case to *Meet Pilot Single-Case Design Standards With Reservations*.

- 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 *Meet Pilot Single-Case Design Standards Without Reservations* and at least one probe point for each case to *Meet Pilot Single-Case Design Standards With Reservations.*

- Each case not receiving the intervention must have a probe point in a session where another case either (a) first receives the intervention or (b) reaches the prespecified intervention criterion. [Note: some review team protocols do not require this third requirement.]

Figure G.2 shows an example of a multiple probe design experiment across three cases that meets the multiple probe design requirements necessary to receive a rating of *Meets Pilot Single-Case Design Standards With Reservations*. The initial baseline sessions for Teri, Kate, and Dan overlap vertically and include three consecutive probe points for each case. In addition, all three cases have at least one probe point just prior to introducing the intervention, and each case not receiving the intervention has a probe point in a session where another case first receives the intervention (e.g., both Kate and Dan have a probe point at Session 7, and Dan has a probe point at Session 9). If all three cases had at least three probe points just prior to introducing the intervention, this design could *Meets Pilot Single-Case Design Standards Without Reservations* (but Teri has only two probe points just prior to introducing the intervention).

Figure G.3 shows another example of a multiple probe design experiment across three cases that can be rated as *Meet Pilot Single-Case Design Standards With Reservations,* for three reasons:

- Each phase has at least three data points (rather than five).

- The initial baseline session for all three students overlaps vertically. Within the first three sessions of the design, Kara has three consecutive baseline probe points, but Wendy and Hannah only have one baseline probe point; the rest of Wendy and Hannah's baseline probe points come after the first three sessions of this design.

- All three cases have at least one probe point just prior to introducing the intervention, and each case not receiving the intervention has a probe point in a session where another case first receives the intervention. However, Wendy and Hannah do not have at least *three consecutive* probe points just prior to introducing the intervention,

because time passed between Session 1 and Session 2 for each student; thus, this design can *Meet Pilot Single-Case Design Standards With Reservations.*

Note that some experiments intentionally do not collect data during a **training phase** for either the teacher or student. A separate guidance document describes how to assess concurrence in these types of designs.

**Figure G.1. A multiple baseline design experiment that uses a consistent display of time across tiers and demonstrates concurrence**

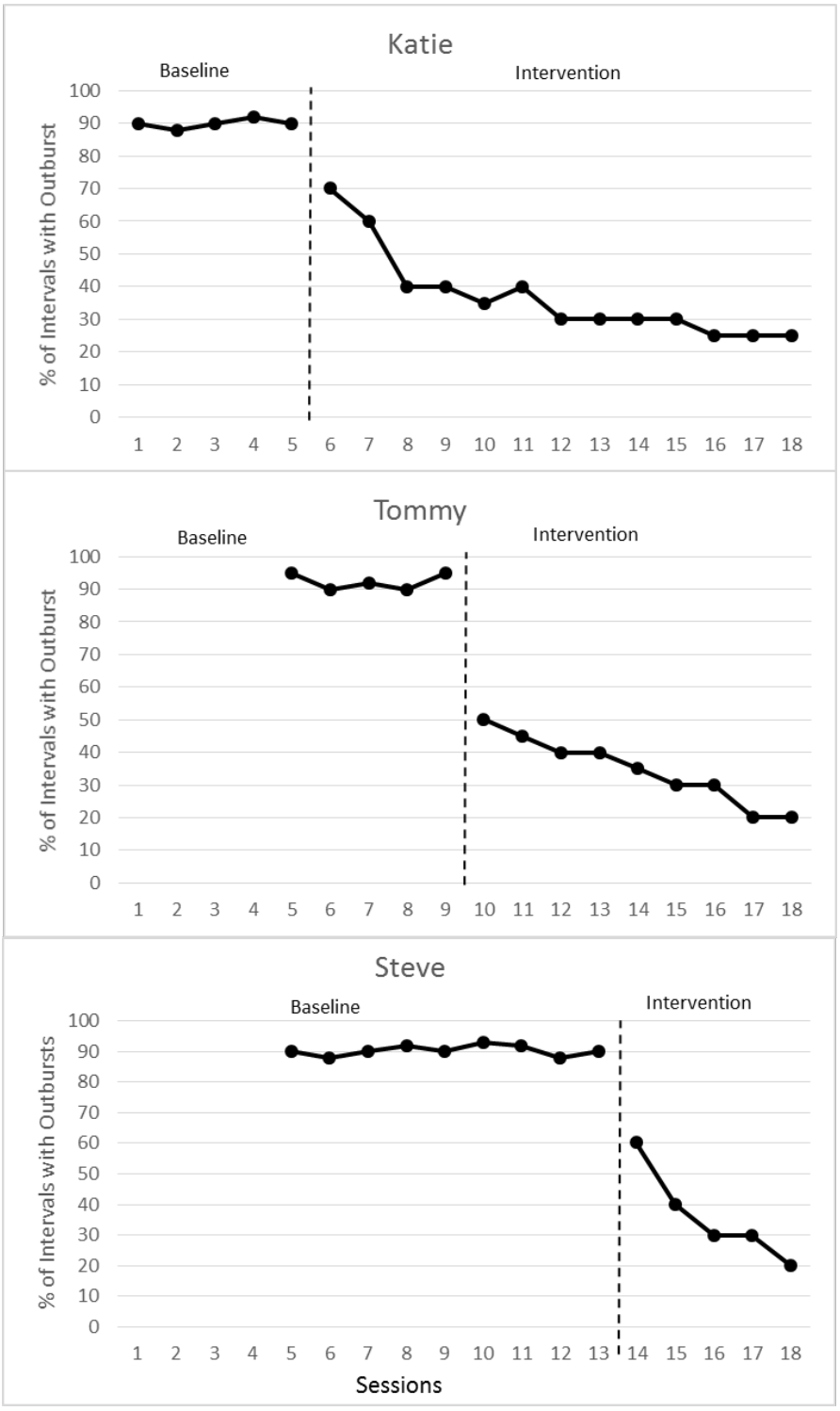


**Figure G.2. A multiple probe design experiment that uses a consistent display of time across tiers and demonstrates concurrence**

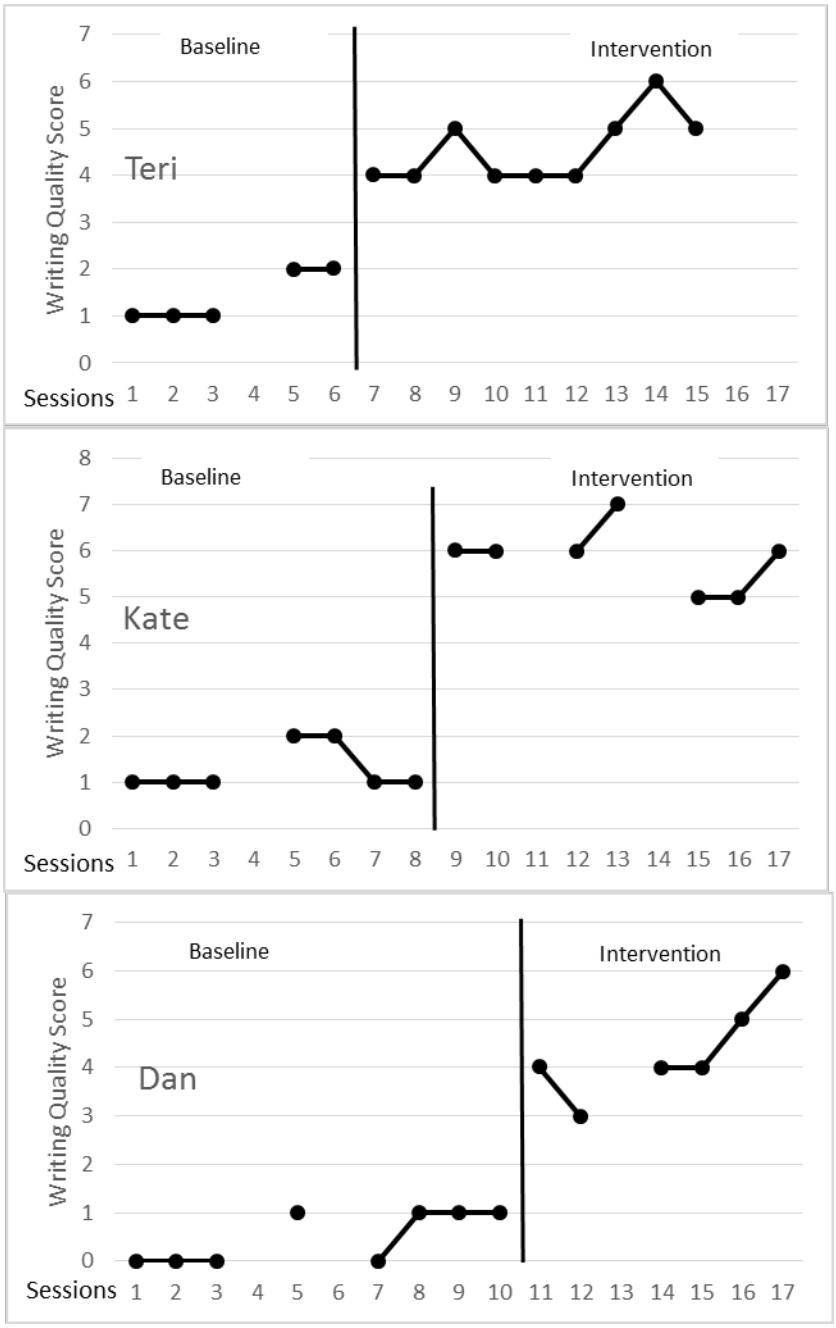


**Figure G.3. A multiple probe design experiment that uses different x-axes but still allows for vertical comparison of what happened at the same time for each case**

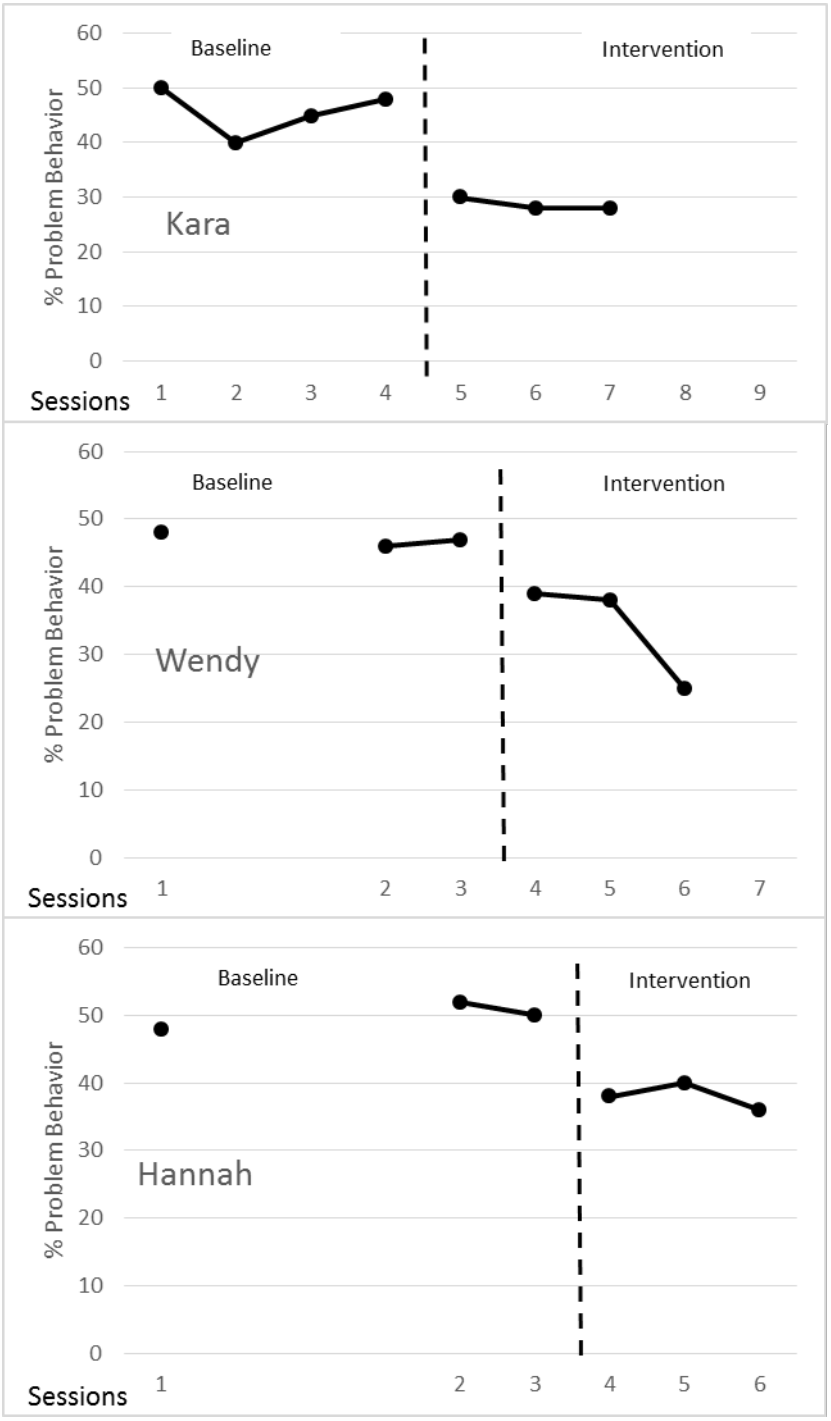
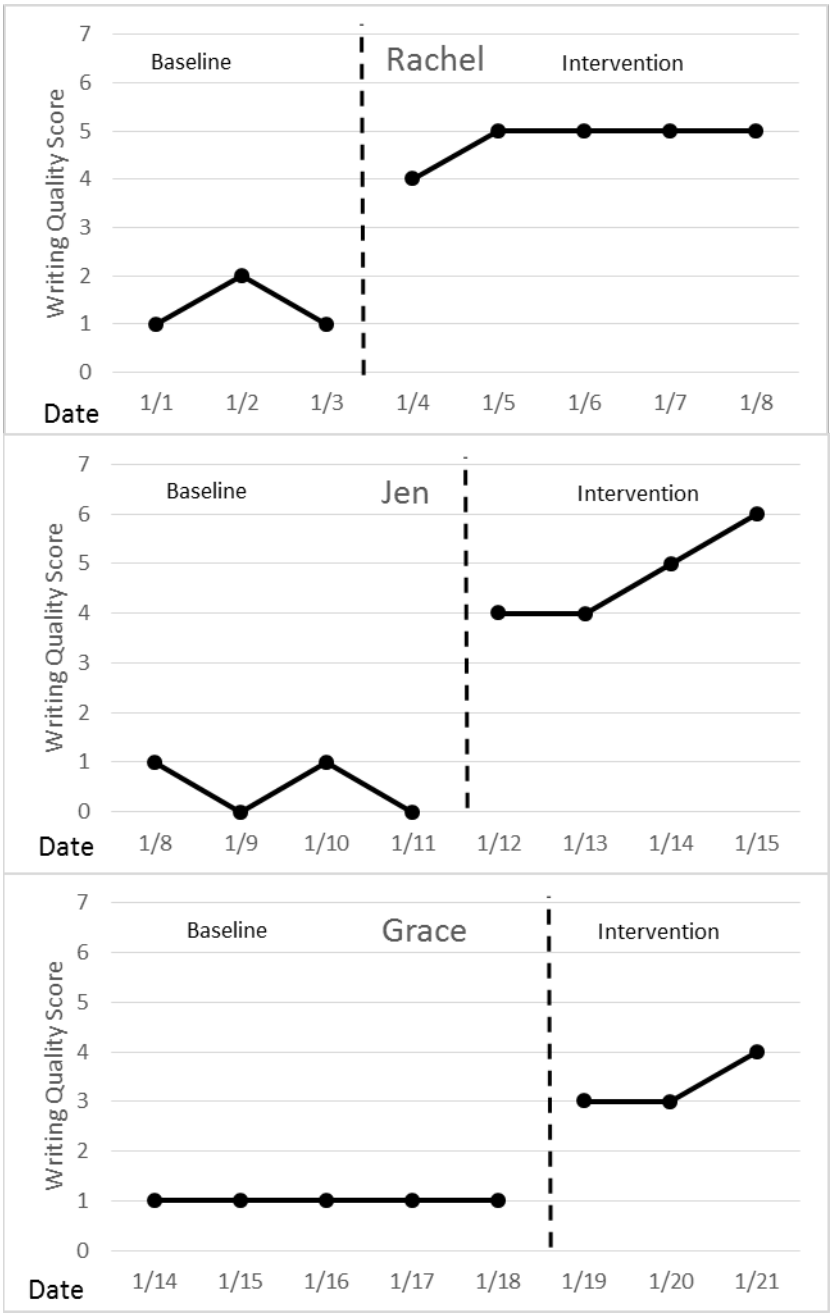


**Figure G.4. Multiple baseline design experiment: Graphical display of data does not show what happened vertically at the same time**



## H. Training Phases in Multiple Baseline and Multiple Probe Design Experiments

Some multiple baseline and multiple probe design experiments separate the intervention phases into training and post-training phases for each case. This type of design is usually used when researchers do not expect the intervention to have an immediate effect on a student, but instead believe that the student must master the intervention before effects will be seen. The training phase—sometimes called the instruction, treatment, or intervention phase—is used to teach the student the intervention, and then effects after complete delivery of the training are measured during the post-training phase. [Note that teacher or parent training phases might also be documented in single-case design experiments; although the majority of this guidance is focused on student training phases, some guidance on parent/teacher training phases is also included, where relevant.]

Authors may collect and present outcome data during the training phase or they might choose to display the training period without data. The graphical display of the experiment may: (1) present all of the sessions in the training phase as "empty" sessions without data points (see Figure H.1), (2) present the training phase (often using a vertical line) without indicating how many sessions were in the phase or describing the timing of the probes (see Figure H.2), or (3) present the training phase with data (see Figure H.3).

Single-case design experiments with training phases may pose several challenges for reviewers, including the ability to: document an immediate effect, evaluate the timing of sessions, or determine whether concurrence exists. The *WWC Standards Handbook (version 4.0)* gives methodologists and content experts discretion to "make decisions about whether the design is appropriate for evaluating an intervention" (p. E.5), but provides no specific guidance for reviews of studies with training phases.

### Guidance

Reviewers should notify the review team leadership if an experiment has a training phase. In the study, authors should provide information about whether they expected an impact at the onset of training or after specific training criteria have been met – this information may be presented in the Methods section or as part of the research question(s). The review team leadership, in consultation with a content expert, will examine this text and confirm the relevant research question, based on the intervention and outcomes, and then determine if the use of a training phase is appropriate for that question. The research question of interest to the WWC may differ from the research question as stated by the authors, and depends on the focal intervention, the outcome, and other contextual information. Research questions fall into one of the following two categories:

1. **Research questions about immediate impacts at the onset of training.** Some interventions can be fully implemented quickly and any effect of the intervention on outcomes is expected to be immediate—for studies of these interventions, the relevant research question is whether the intervention has an immediate effect. For example, a flash card intervention that teaches students to memorize a small set of math facts such as 2+2=4 should immediately affect knowledge of math facts.

2. **Research questions about impacts after complete delivery of the training (e.g., after all training criteria have been met).** Other interventions require training sessions to take place before an impact is expected—for studies of these interventions, the relevant research question is whether the intervention has an effect following the completion of training session(s) (e.g., after reaching some pre-determined training criterion.) For example, self-regulated strategy development (SRSD) teaches writing strategies and self-regulation using a multi-step process. Because students must master and synthesize skills taught in the earlier steps, SRSD is not expected to affect writing quality immediately when training begins, but only after all of the training steps have been completed.

The guidance for reviewing single-case design experiments that include training phases varies depending on the research question (e.g., whether or not an immediate impact is expected), and whether the training phase is for a student or for a teacher/parent. Please see the sections below for specific guidance on how to approach each of these situations.

*Guidance for interventions that expect immediate impacts at the onset of training*

When an immediate impact is expected, outcome data should be collected as soon as students are exposed to the intervention. Otherwise, the presence or lack of an immediate impact cannot be demonstrated, and threats to internal validity, such as maturation or history, may affect outcomes.

**Training phases without data.** Single-case design experiments that include a student training phase without data cannot be used to answer research questions about immediate impacts. The training phase in these designs is equivalent to an intervention phase, and these designs do not provide the necessary data from the time when the intervention was first administered. Thus, when the relevant research question is about immediate impacts, single-case design figures that include student training phases represented as "empty" sessions or vertical lines without data points (e.g., Figures I.1 and I.2) 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 single-case design experiments might include a *teacher* or *parent* training phase without presenting student data from that time period; these experiments can answer research questions about immediate impacts on students, as long as students do not receive any portion of the intervention until the parent or teacher training period is complete.

**Training phases with data.** Single-case design experiments that have student training phases with data (e.g., Figure H.3) can answer research questions about immediate impacts. If the authors present data from a training phase for students, that phase should be treated as the first intervention phase. When completing visual analysis, the baseline phase should be compared with the training phase to look for evidence of an effect; training and subsequent post-training intervention phases may be combined for purposes of the WWC review, if the conditions are similar. Consult with the review team leadership to determine whether the training and post-training intervention phases should be combined when conducting the review.

Single-case design experiments might include a post-baseline *teacher* or *parent* training phase with student outcome data. Reviewers should consult with the review team leadership to determine if and how to incorporate these data into the review.

*Guidance for interventions that expect impacts only after complete delivery of the training*

When an intervention impact is expected only after complete delivery of the training, the authors do *not* need to present data from the training phase to document an impact. Instead, training phases may be displayed graphically as "empty" sessions without data points or with vertical lines (e.g., Figures I.1, I.2, and I.4).

**Training phases without data.** Single-case design experiments that include a student training phase without data are appropriate for answering research questions about impacts after complete delivery of the training. However, it can be difficult to evaluate timing and concurrence in these designs, both of which are necessary to rule out potential confounding factors.

When training phases are presented without data it may be unclear whether the authors present the timing of sessions consistently across data series (for example, see Figure H.2, where the training phase is represented with a vertical line). Reviewers should look for more information about the timing and duration of the training phase in the text. Unless something in the text or the figure suggests otherwise, reviewers can assume that the timing of sessions is consistent across cases. If something in the text or the figure raises questions, an author query may be necessary to determine whether or not the timing was presented consistently across data series.

- For example, reviewers can assume that timing is consistent across cases in Figure H.2, unless the text says otherwise. However, if the text states that Marcia's training lasted one day and occurred between sessions 6 and 7, but Gary's training lasted five days (and occurred between sessions 8 and 9), an author query might be sent to clarify whether there was also a five-day gap between Karen's sessions 8 and 9. If the timing is not presented consistently across cases, reviewers should ask the authors for a consistent display of data. (Refer to the Timing and Concurrence guidance document for more information.)

Once reviewers have determined that timing of sessions is displayed consistently, they should assess concurrence and effects. In order to have concurrence, the cases still in the baseline phase must continue baseline measurement at or after the time point when a preceding case has the first intervention probe *after completing their training*. In other words, there can be no overlap in the training phases among the cases in the experiment. (See Figure H.1 for an example of an experiment with no overlap of training phases).

- **If this requirement is not met, then there is no concurrence** –the design cannot exclude threats to internal validity and 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.

  - For example, in Figure H.4, the training phases for each student overlap and all three students received training during session 10. It is possible that the class started a new writing module in their regular curriculum around session 10, and that module—not the intervention—was responsible for the improvement in writing quality.

- **If this requirement is met, the experiment can *Meet WWC Pilot Single-Case Design Standards***. In addition, when evaluating concurrence in **multiple probe designs**, the Handbook also requires that "Each case not receiving the intervention must have a probe point in a session where another case either (a) first receives the intervention or (b) reaches the prespecified intervention criterion." When impacts are expected only after complete delivery of the 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. (Note that some review protocols allow studies to *Meet WWC Pilot Single-Case Design Standards With Reservations*, even if they do not meet this multiple probe standard.)

If an experiment meets standards (with or without reservations), visual analysis should be conducted to assess whether an effect is demonstrated. When impacts are only expected after all components of the training are implemented (but not after partial implementation, during the training phase), the evaluation of the "immediacy of effect" should focus on the period directly following the conclusion of training, sometimes referred to as the post-training or post-intervention stage. For example, in Figure H.3, the effect of the intervention would be measured starting at time points 8, 12, and 16 for Luke, Maya, and Elena, respectively. When conducting visual analysis, the reviewer should also evaluate whether the cases still in the baseline phase demonstrate an effect at the same time a change occurs for the preceding case(s). The reviewer should describe the data patterns in the study review guide and factor this into the *evidence rating* (e.g., strong, moderate, or no evidence). However, a change in the baseline data pattern for a case not receiving the intervention will not affect the *study rating*.

**Training phases with data.** Some single-case design experiments include a student training phase with data (e.g., Figure H.3), even though an impact is not expected until after the training has been fully delivered. Concurrence can be evaluated using the same guidance that was provided above for evaluating training phases without data. When conducting visual analysis, the reviewer can look at the data in the training phase but should mainly focus on the baseline and post-training phases. Because impacts are only expected after all components of the training are implemented, the observed effect does not need to be demonstrated during the training phase, but should be apparent in the post-training phase, after the intervention has been fully delivered.

These types of single-case design experiments might include a *teacher or parent* training phase with student outcome data. Reviewers should consult with the review team leadership to determine if and how to incorporate these data into the review.

**Figure H.1. Graphical Display that Denotes Training Phase using "Empty" Sessions without Data Points**

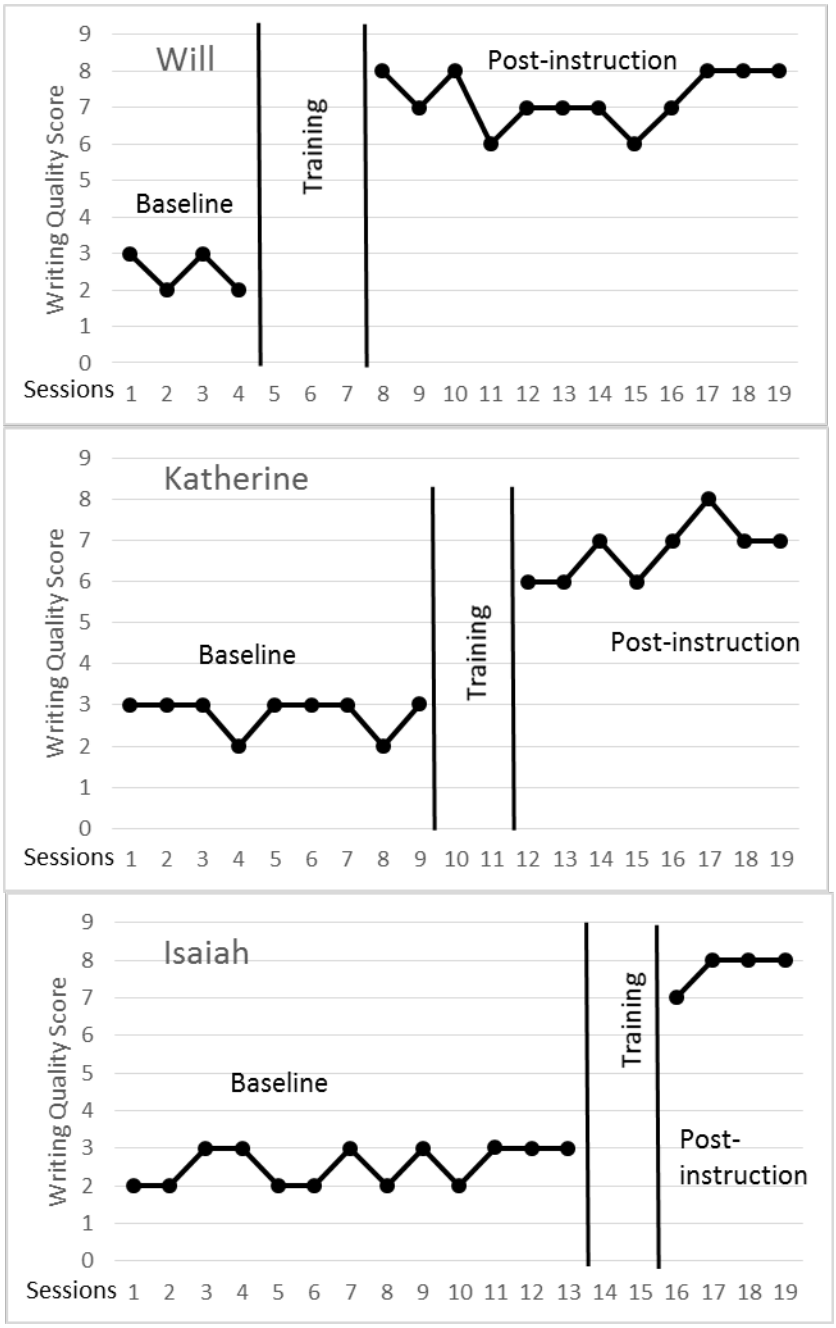


**Figure H.2. Graphical Display using a Vertical Line to Denote the Training Phase**

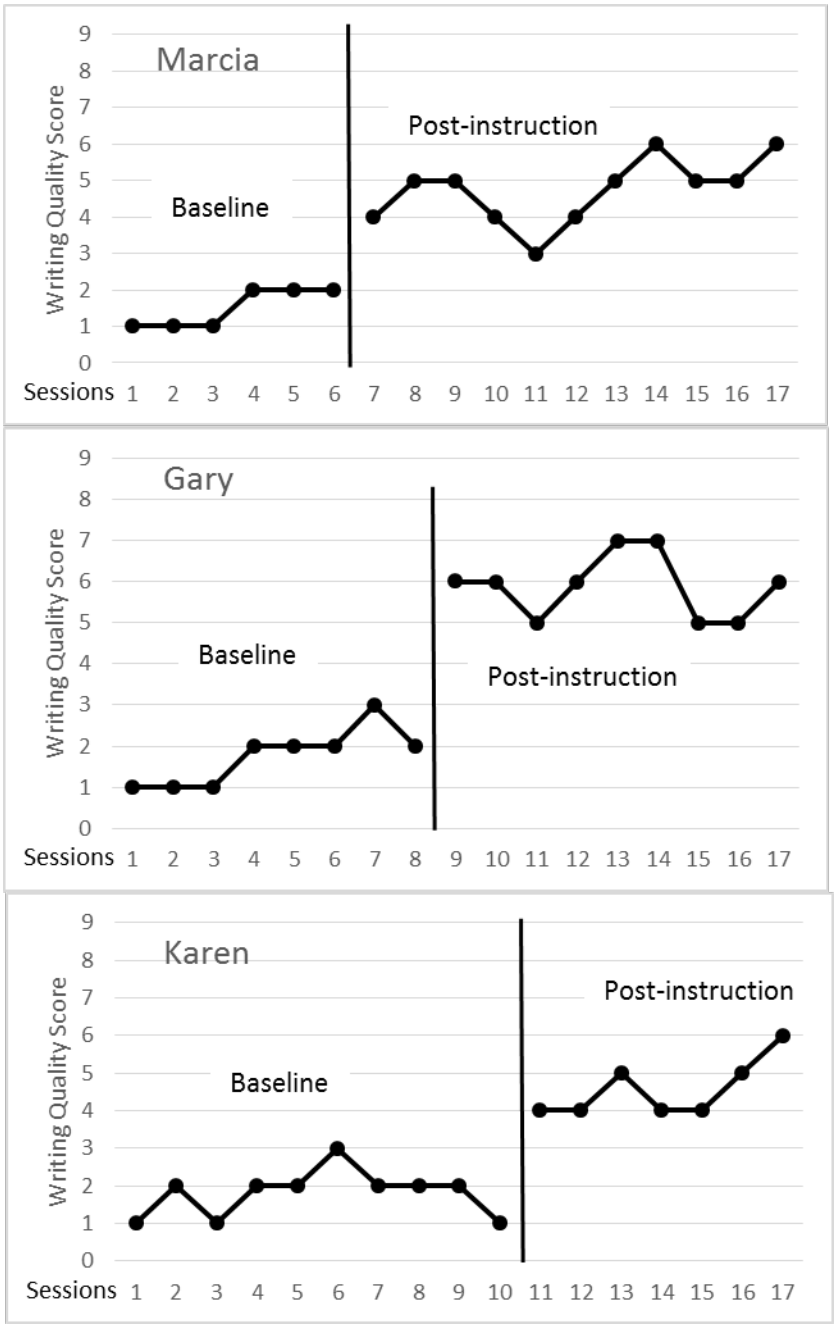


**Figure H.3. Graphical Display that Presents Data from the Training Phase**

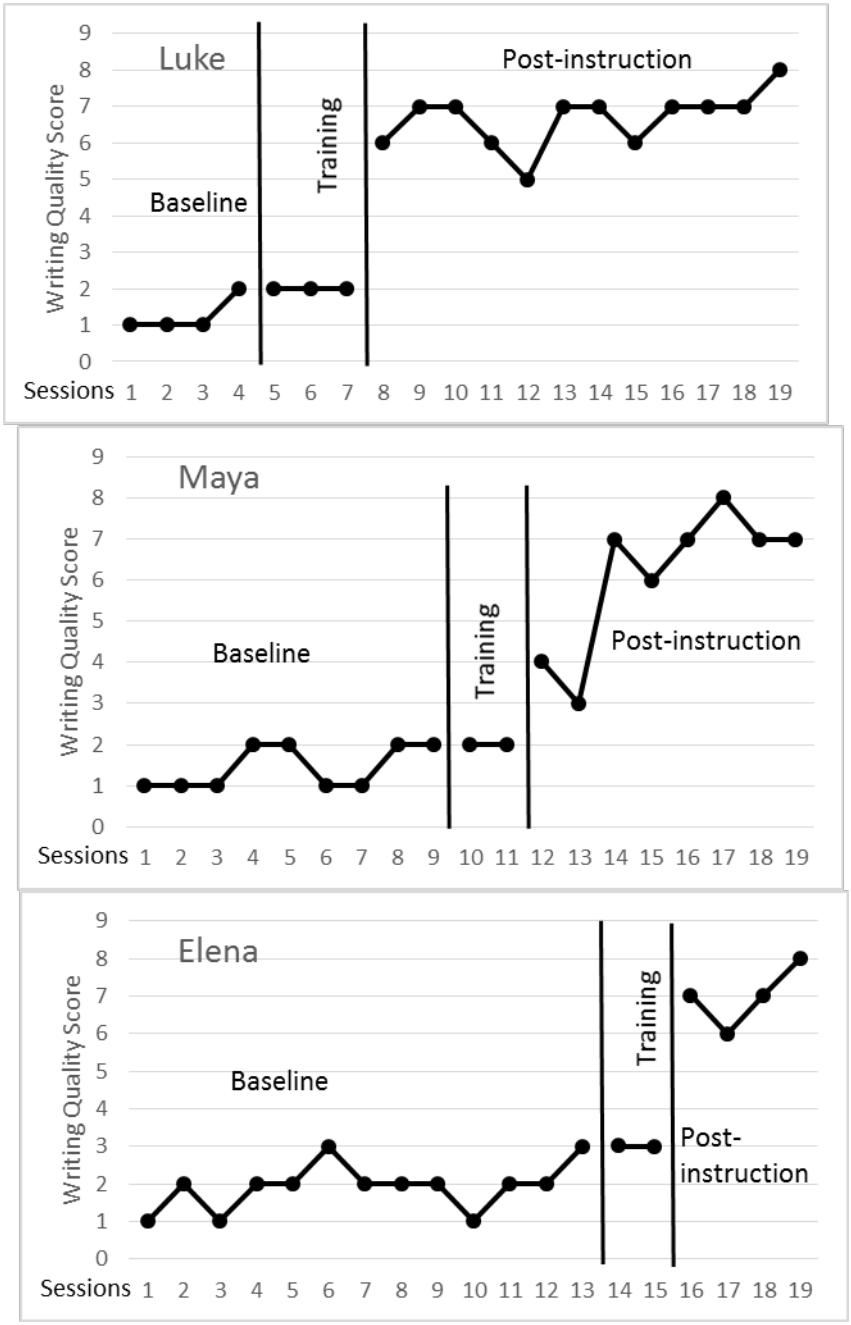
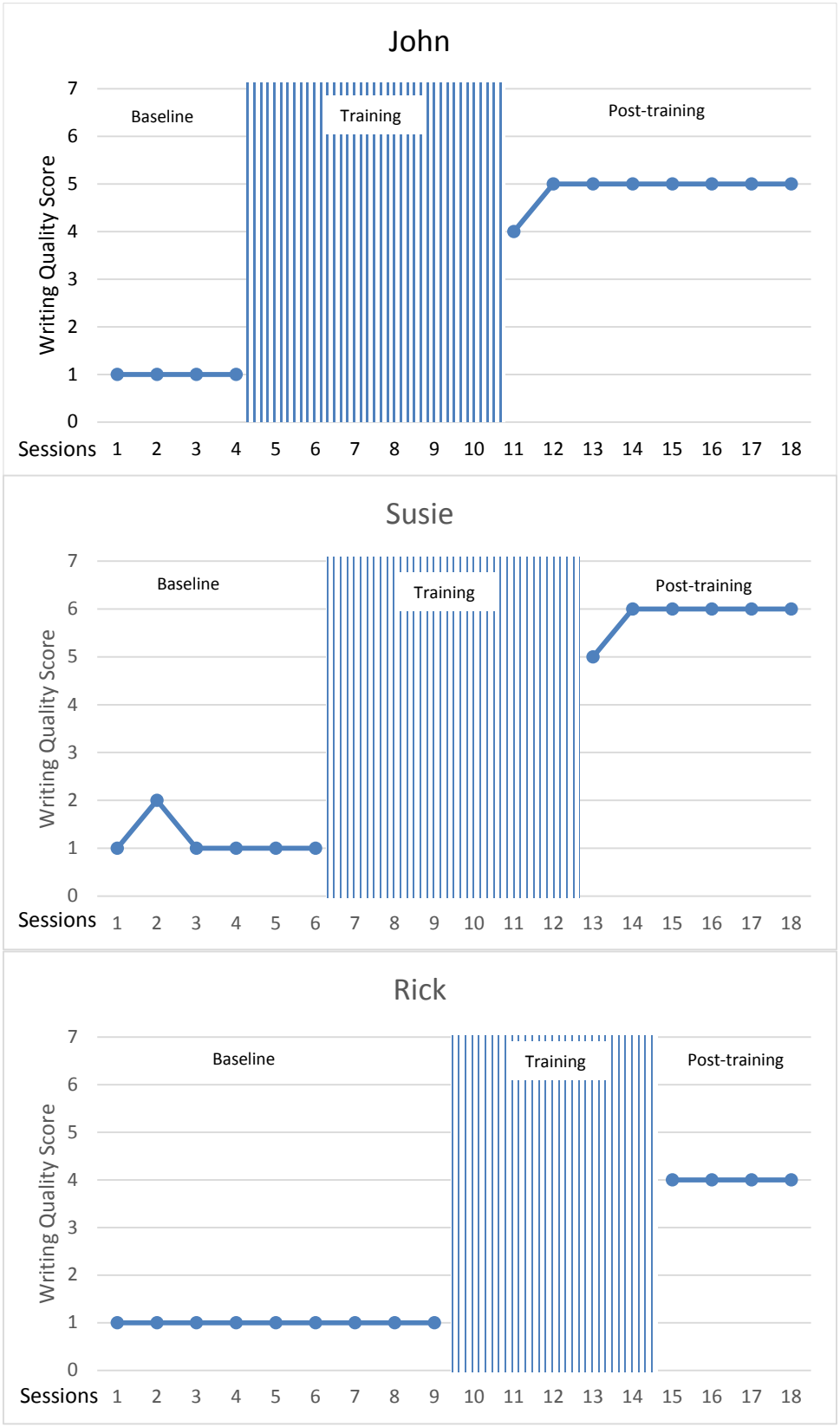


**Figure H.4. Multiple Baseline Design with Overlapping Training Phases**



## REFERENCES

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

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.

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

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