What Works Clearinghouse

SINGLE-CASE DESIGN TECHNICAL DOCUMENTATION
Version 1.0 (Pilot)

Developed for the What Works Clearinghouse by the following panel:

Kratochwill, T. R.
Hitchcock, J.
Horner, R. H.
Levin, J. R.
Odom, S. L.
Rindskopf, D. M.
Shadish, W. R.

June 2010

Recommended citation:
Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M. & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

SINGLE-CASE DESIGNS TECHNICAL DOCUMENTATION

In an effort to expand the pool of scientific evidence available for review, the What Works Clearinghouse (WWC) assembled a panel of national experts in single-case design (SCD) and analysis to draft SCD Standards. In this paper, the panel provides an overview of SCDs, specifies the types of questions that SCDs are designed to answer, and discusses the internal validity of SCDs. The panel then proposes SCD Standards to be implemented by the WWC. The Standards are bifurcated into Design and Evidence Standards (see Figure 1). The Design Standards evaluate the internal validity of the design. Reviewers assign the categories of *Meets Standards, Meets Standards with Reservations* and *Does not Meet Standards* to each study based on the Design Standards. Reviewers trained in visual analysis will then apply the Evidence Standards to studies that meet standards (with or without reservations), resulting in the categorization of each outcome variable as demonstrating *Strong Evidence, Moderate Evidence, or No Evidence.*

A. OVERVIEW OF SINGLE-CASE DESIGNS

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

SCDs are identified by the following features:

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

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

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

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

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

B. CAUSAL QUESTIONS THAT SCDS ARE DESIGNED TO ANSWER

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

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

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

Because the underlying goal of SCDs is most often to determine "Which intervention is effective for this case (or these cases)?" the designs are intentionally flexible and adaptive. For example, if a participant is not responding to an intervention, then the independent variables can be manipulated while continuing to assess the dependent variable (Horner et al., 2005). Because of the adaptive nature of SCD designs, nonresponders might ultimately be considered "responders" under particular conditions.[1] In this regard, SCDs provide a window into the process of participant change. SCDs can also be flexible in terms of lengthening the number of data points collected during a phase to promote a stable set of observations, and this feature may provide additional insight into participant change.

C. THREATS TO INTERNAL VALIDITY IN SINGLE-CASE DESIGN[2]

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

In SCD research, effect replication is an important mechanism for controlling threats to internal validity and its role is central for each of the various threats discussed below. In fact, the

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

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

replication criterion discussed by Horner et al. (2005, p. 168) represents a fundamental characteristic of SCDs: "In most [instances] experimental control is demonstrated when the design documents *three* demonstrations of the experimental effect at *three* different points in time with a single case (within-case replication), or across different cases (inter-case replication) (emphasis added)." As these authors note, an experimental effect is demonstrated when the predicted changes in the dependent measures covary with manipulation of the independent variable. This criterion of three replications has been included in the *Standards* for designs to "meet evidence" standards. Currently, there is no formal basis for the "three demonstrations" recommendation; rather, it represents a conceptual norm in published articles, research, and textbooks that recommend methodological standards for single-case experimental designs (Kratochwill & Levin, in press).

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

Table 1, adapted from Hayes (1981) but without including the original "design type" designations, presents the three major types of SCDs and their variations. In AB designs, a case's performance is measured within each condition of the investigation and compared between or among conditions. In the most basic two-phase AB design, the A condition is a baseline or preintervention series/phase and the B condition is an intervention series/phase. It is difficult to draw valid causal inferences from traditional two-phase AB designs because the lack of replication in such designs makes it more difficult to rule out alternative explanations for the observed effect (Kratochwill & Levin, in press). Furthermore, repeating an AB design across several cases in separate or independent studies would typically not allow for drawing valid inferences from the data (Note: this differs from multiple baseline designs, described below, which introduce the intervention at different points in time). The *Standards* require a minimum of four A and B phases, such as the ABAB design.

There are three major classes of SCD that incorporate phase repetition, each of which can accommodate some form of randomization to strengthen the researcher's ability to draw valid causal inferences (see Kratochwill & Levin, in press, for discussion of such randomization applications). These design types include the ABAB design (as well as the changing criterion design, which is considered a variant of the ABAB design), the multiple baseline design, and the alternating treatments design. Valid inferences associated with the ABAB design are tied to the design's structured repetition. The phase repetition occurs initially during the first B phase, again in the second A phase, and finally in the return to the second B phase (Horner et al., 2005). This design and its effect replication standard can be extended to multiple repetitions of the treatment (e.g., ABABABAB) and might include multiple treatments in combination that are introduced in a repetition sequence as, for example, A/(B+C)/A/(B+C)/A (see Table 1). In the case of the changing criterion design, the researcher begins with a baseline phase and then schedules a series of criterion changes or shifts that set a standard for participant performance over time. The criteria are typically pre-selected and change is documented by outcome measures changing with the criterion shifts over the course of the experiment.

TABLE 1

EXAMPLE SINGLE-CASE DESIGNS AND ASSOCIATED CHARACTERISTICS

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

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

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

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

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

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

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

Embedded in the SCD *Standards* is a criterion that the independent variable is actively manipulated by the researcher, with measurement of the dependent variable occurring after that manipulation. This sequencing ensures the presumed cause precedes the presumed effect. A SCD cannot meet *Standards* unless there is active manipulation of the independent variable.[3]

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

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

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

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

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

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

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

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

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

In single-case research, cases are often selected because their pre-experimental or baseline measures suggest high need or priority for intervention (e.g., immediate treatment for some problem is necessary). If only pretest and posttest scores were used to evaluate outcomes, statistical regression would be a major concern. However, the repeated assessment identified as a distinguishing feature of SCDs in the *Standards* (wherein performance is monitored to evaluate level, trend, and variability, coupled with phase repetition in the design) makes regression easy to diagnose as an internal validity threat. As noted in the *Standards*, data are repeatedly collected during baseline and intervention phases and this repeated measurement enables the researcher to examine characteristics of the data for the possibility of regression effects under various conditions.

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

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

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

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

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

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

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

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

D. THE SINGLE-CASE DESIGN STANDARDS

The PI within each topic area will: (1) define the independent and outcome variables under investigation,[4] (2) establish parameters for considering fidelity of intervention implementation,[5] and (3) consider the reasonable application of the *Standards* to the topic area and specify any deviations from the *Standards* in that area protocol. For example, when measuring self-injurious behavior, a baseline phase of fewer than five data points may be appropriate. PIs might need to make decisions about whether the design is appropriate for evaluating an intervention. For example, an intervention associated with a permanent change in participant behavior should be evaluated with a multiple baseline design rather than an ABAB design. PIs will also consider the various threats to validity and how the researcher was able to address these concerns, especially in cases in which the *Standards* do not necessarily mitigate the validity threat in question (e.g., testing, instrumentation). Note that the SCD *Standards* apply to both observational measures and standard academic assessments. Similar to the approach with group designs, PIs are encouraged to define the parameters associated with "acceptable" assessments in their protocols. For example, repeated measures with alternate forms of an assessment may be acceptable and WWC psychometric criteria would apply. PIs might also need to make decisions about particular studies. Several questions will need to be considered, such as: (a) Will generalization variables be reported? (b) Will follow-up phases be assessed? (c) If more than one consecutive baseline phase is present, are these treated as one phase or two distinct phases? and (d) Are multiple treatments conceptually distinct or multiple components of the same intervention?

SINGLE-CASE DESIGN STANDARDS

These Standards are intended to guide WWC reviewers in identifying and evaluating SCDs. The first section of the *Standards* assists with identifying whether a study is a SCD. As depicted in Figure 1, a SCD should be reviewed using the 'Criteria for Designs that Meet Evidence Standards', to determine those that *Meet Evidence Standards,* those that *Meet Evidence Standards with Reservations,* and those that *Do Not Meet Evidence Standards.*

Studies that meet evidence standards (with or without reservations) should then be reviewed using the 'Criteria for Demonstrating Evidence of a Relation between an Independent Variable and a Dependent Variable' (see Figure 1).[6] This review will result in a sorting of SCD studies into three groups: those that have *Strong Evidence of a Causal Relation,* those that have *Moderate Evidence of a Causal Relation,* and those that have *No Evidence of a Causal Relation.*

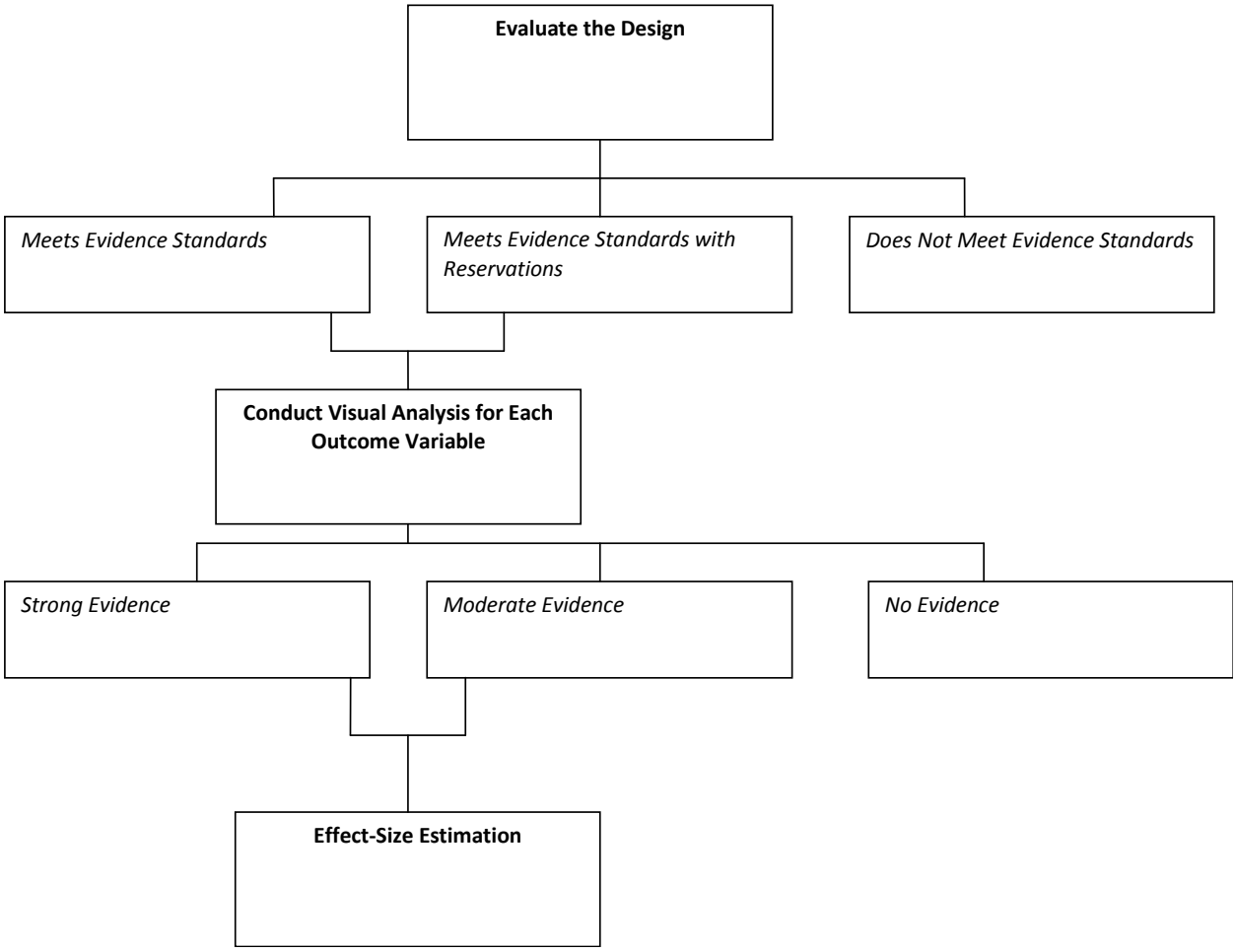
[4] Because SCDs are reliant on phase repetition and effect replication across participants, settings, and researchers to establish external validity, specification of the intervention materials, procedures, and context of the research is particularly important within these studies (Horner et al., 2005).

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

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

FIGURE 1

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



A. SINGLE-CASE DESIGN CHARACTERISTICS

SCDs are identified by the following features:

- An individual "case" is the unit of intervention and the unit of data analysis. A case may be a single participant or a cluster of participants (e.g., a classroom or community).
- Within the design, the case provides its own control for purposes of comparison. For example, the case's series of outcome variables prior to the intervention is compared with the series of outcome variables during (and after) the intervention.
- The outcome variable is measured *repeatedly* within and across *different* conditions or levels of the independent variable. These different conditions are referred to as "phases" (e.g., baseline phase, intervention phase).[7]

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

B. CRITERIA FOR DESIGNS THAT MEET EVIDENCE STANDARDS

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

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

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

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

- **Each outcome variable must be measured systematically over time by more than one assessor, and the study needs to collect inter-assessor agreement in each phase and on at least twenty percent of the data points in each condition (e.g., baseline, intervention) and the inter-assessor agreement must meet minimal thresholds**. Inter-assessor agreement (commonly called interobserver agreement) must be documented on the basis of a statistical measure of assessor consistency. Although there are more than 20 statistical measures to represent inter-assessor agreement (see Berk, 1979; Suen & Ary, 1989), commonly used measures include percentage agreement (or proportional agreement) and Cohen's kappa coefficient (Hartmann, Barrios, & Wood, 2004). According to Hartmann et al. (2004), minimum acceptable values of inter-assessor agreement range from 0.80 to 0.90 (on average) if measured by percentage agreement and at least 0.60 if measured by Cohen's kappa. Regardless of the statistic, inter-assessor agreement must be assessed for each case on each outcome variable. A study needs to collect inter-assessor agreement in all phases. It must also collect inter-assessor agreement on at least 20% of all sessions (total across phases) for a condition (e.g., Baseline, Intervention.).[8] If this standard is not met, the study *Does Not Meet Evidence Standards.*

- **The study must include at least three attempts to demonstrate an intervention effect at three different points in time or with three different phase repetitions**. If this standard is not met, the study *Does Not Meet Evidence Standards*.[9] Examples of designs meeting this standard include ABAB designs, multiple baseline designs with at least three baseline conditions, alternating/simultaneous treatment designs with either at least three alternating treatments compared with a baseline condition or two alternating treatments compared with each other, changing criterion designs with at least three different criteria, and more complex variants of these designs. Examples of designs not meeting this standard include AB, ABA, and BAB designs.[10]

- **For a phase to qualify as an attempt to demonstrate an effect, the phase must have a minimum of three data points**.[11]

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

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

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

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

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

To *Meet Standards with Reservations* a reversal /withdrawal (e.g., ABAB) design must have a minimum of four phases per case with at least 3 data points per phase. Any phases based on fewer than three data *points cannot* b*e used to demonstrat*e existence or lack of an effect.

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

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

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

For studies that meet standards (with and without reservations), the following rules are used to determine whether the study provides *Strong Evidence*, *Moderate Evidence*, or *No Evidence* of a causal relation. In order to provide *Strong Evidence*, at least two WWC reviewers certified in visual (or graphical) analysis must verify that a causal relation was documented. Specifically this is operationalized as at least three demonstrations of the intervention effect along with no non-effects by[12]

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

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

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

- Data within the baseline phase do not provide sufficient demonstration of a clearly defined pattern of responding that can be used to extrapolate the expected performance forward in time assuming no changes to the independent variable
- Failure to establish a consistent pattern within any phase (e.g., high variability within a phase)
- Either long latency between introduction of the independent variable and change in the outcome variable or overlap between observed and projected patterns of the outcome variable between baseline and intervention phases makes it difficult to determine whether the intervention is responsible for a claimed effect
- Inconsistent patterns across similar phases (e.g., an ABAB design in which the first time an intervention is introduced the outcome variable data points are high, the second time an intervention is introduced the outcome variable data points are low, and so on)
- Comparing the observed and projected patterns of the outcome variable between phases does not demonstrate evidence of a causal relation

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

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

D. VISUAL ANALYSIS OF SINGLE-CASE RESEARCH RESULTS[13]

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

E. RECOMMENDATIONS FOR COMBINING STUDIES

When implemented with multiple design features (e.g., within- and between-case comparisons), SCDs can provide a strong basis for causal inference (Horner et al., 2005). Confidence in the validity of intervention effects demonstrated within cases is enhanced by replication of effects across different cases, studies, and research groups (Horner & Spaulding, in press). The results from single-case design studies will not be combined into a single summary rating unless they meet the following threshold:[14]

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

F. EFFECT-SIZE ESTIMATES FOR SINGLE-CASE DESIGNS[15]

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

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

[15] Prepared by David Rindskopf and William Shadish.

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

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

A number of nonparametric methods have been used to analyze SCDs (e.g., Percentage of Nonoverlapping Data [PND], Percentage of All Nonoverlapping Data [PAND], or Percent Exceeding the Median [PEM]). Some of these have been accompanied by efforts to convert them to parametric estimators such as the phi coefficient, which might in turn be comparable to typical between-groups measures. If that could be done validly, then one could use distribution theory from standard estimators to create standard errors and significance tests. However, most such efforts make the erroneous assumption that nonparametric methods do not need to be concerned with the assumption of independence of errors, and so the conversions might not be valid. In such cases, the distributional properties of these measures are unknown, and so standard errors and statistical tests are not formally justified. Nonetheless, if all one wanted was a rough measure of the approximate size of the effect without formal statistical justification or distribution theory, selecting one of these methods would make sense. However, none of these indices deal with trend, so the data would need to be detrended[16] with, say, first-order differencing before computing the index. One could combine the results with ordinary unweighted averages, or one could weight by the number of cases in a study.

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

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

likely to be present in SCD data. For multi-level models, many researchers are less familiar with both the analytic methods and the interpretation of results, so that their widespread use is probably less likely than with regression. Also, practical implementation of multilevel models for SCDs is technically challenging, probably requiring the most intense supervision and problem-solving of any method. Even if these technical developments were to be solved, the resulting estimates would still be in a different metric than effect-size estimates based on between-group studies, so one could not compare effect sizes from SCDs to those from group studies.

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

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

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

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

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

REFERENCES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

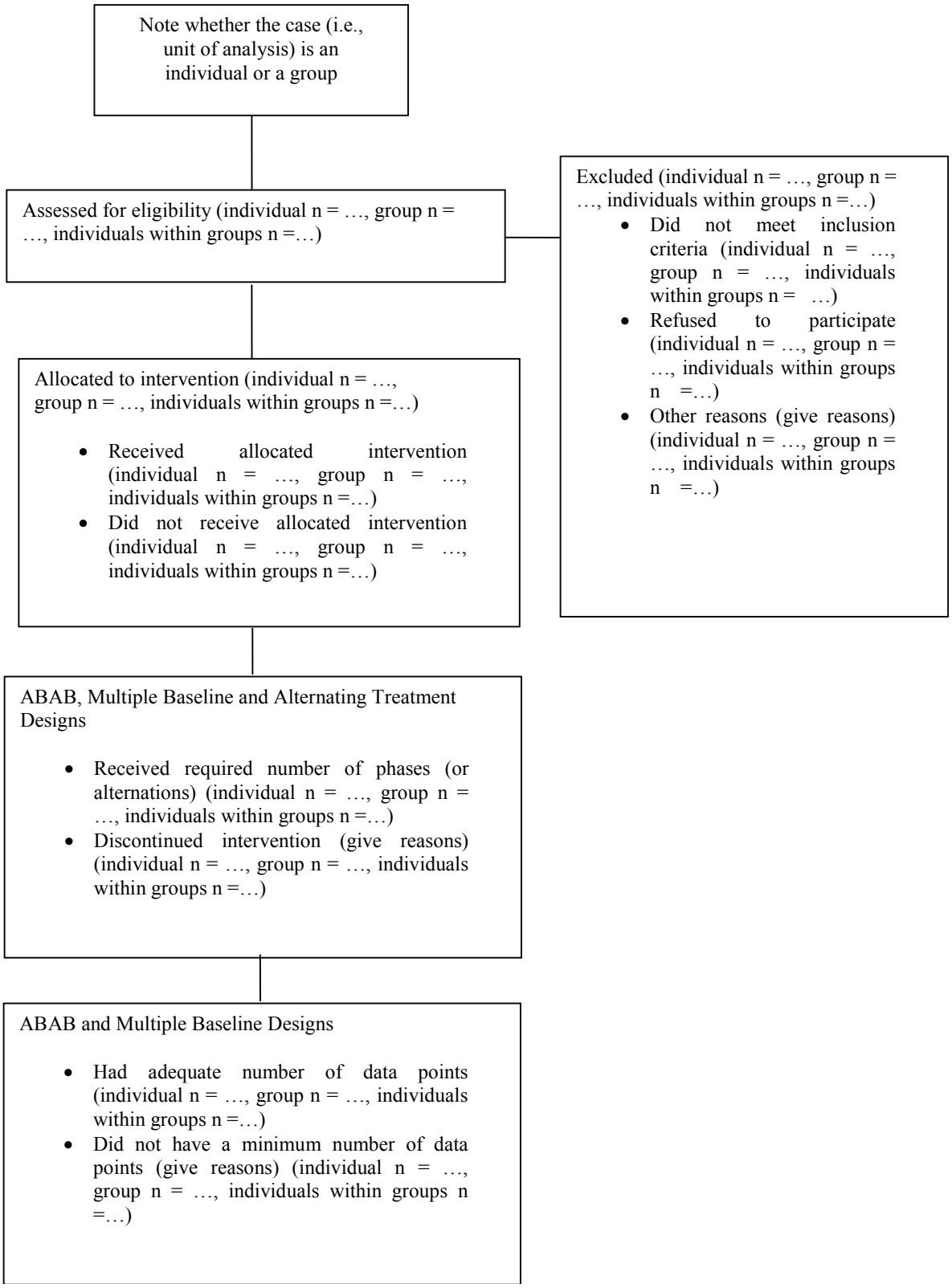
What Works Clearinghouse. (2008). *Procedures and standards handbook* (version 2.0). Retrieved July 10, 2009, from http://ies.ed.gov/ncee/wwc/references/idocviewer/doc.aspx?docid=19&tocid=1

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

APPENDIX A

ATTRITION DIAGRAM

ATTRITION DIAGRAM



APPENDIX B

VISUAL ANALYSIS

Figure 1. Depiction of an ABAB Design

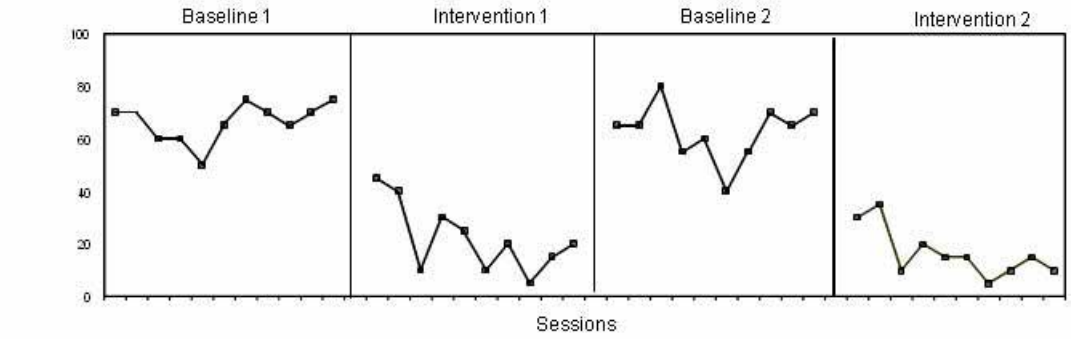


Figure 2. An Example of Assessing Level with Four Phases of an ABAB Design

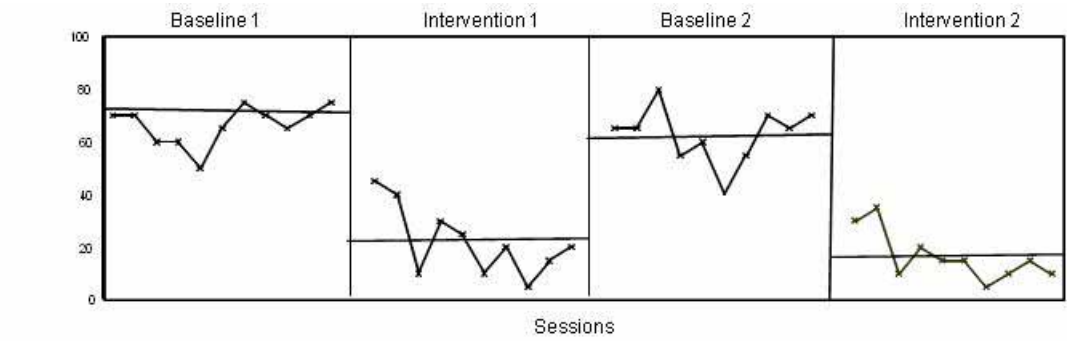


Figure 3. An Example of Assessing in Each Phase of an ABAB Design

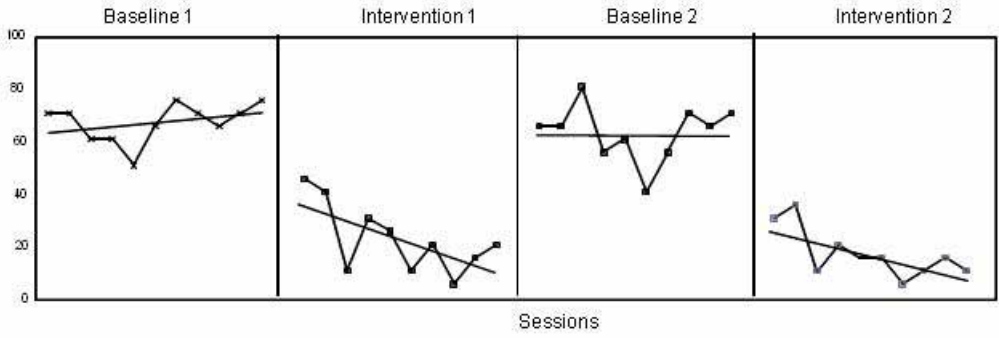


Figure 4. Assess Variability Within Each Phase

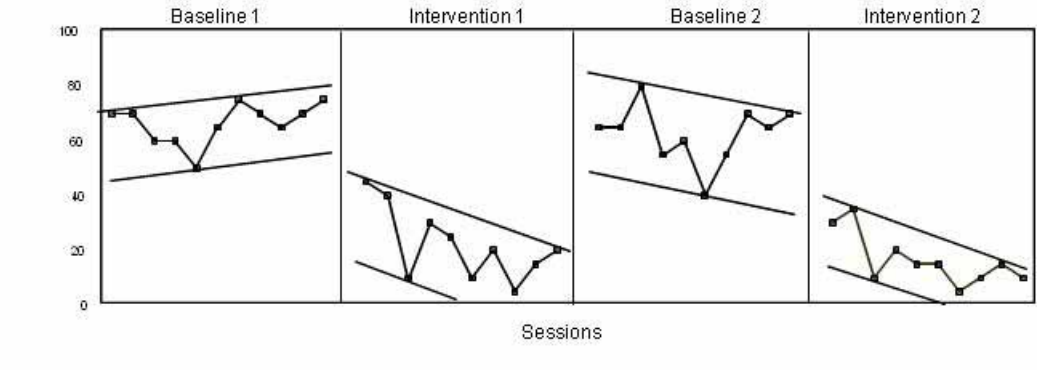


Figure 5. Consider Overlap Between Phases

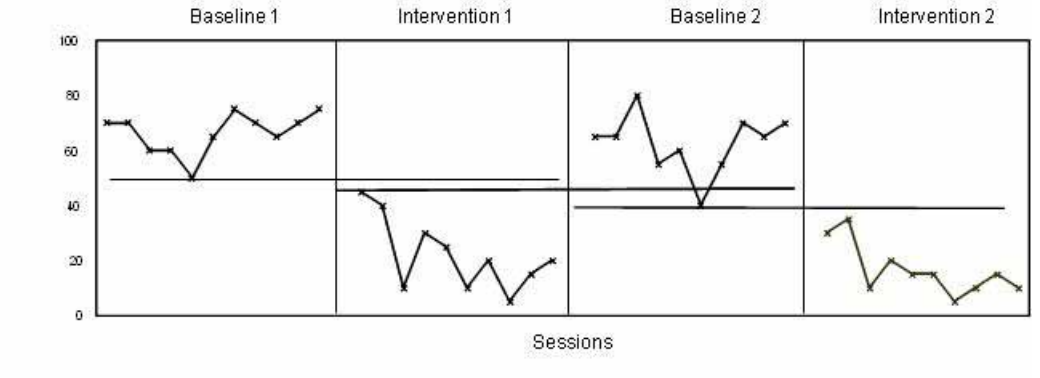


Figure 6. Examine the Immediacy of Effect with Each Phase Transition

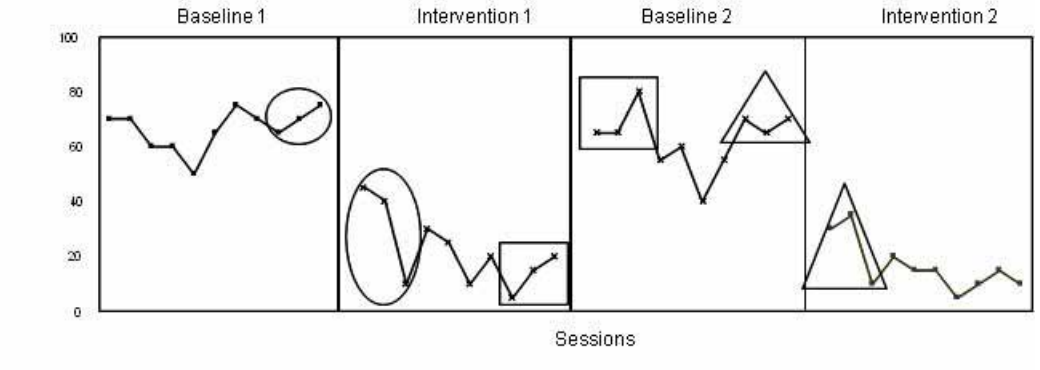


Figure 7. Examine Consistency Across Similar Phases

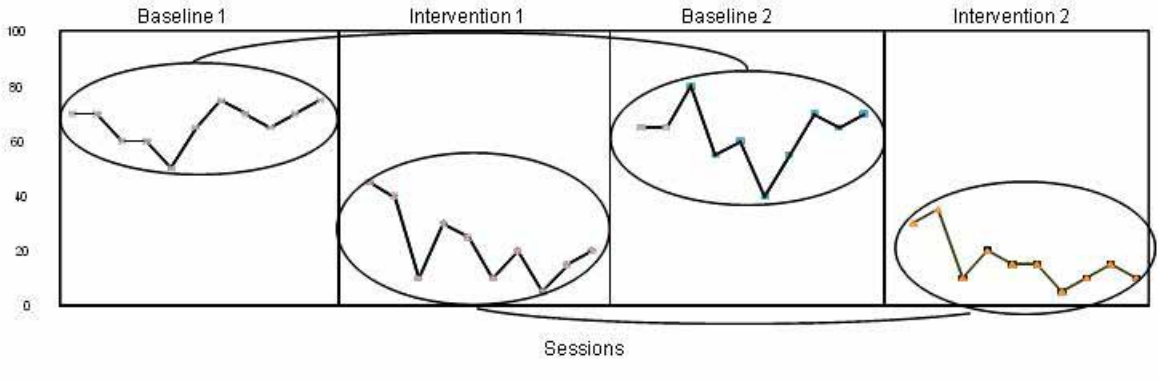


Figure 8A. Examine Observed and Projected Comparison Baseline 1 to Intervention 1

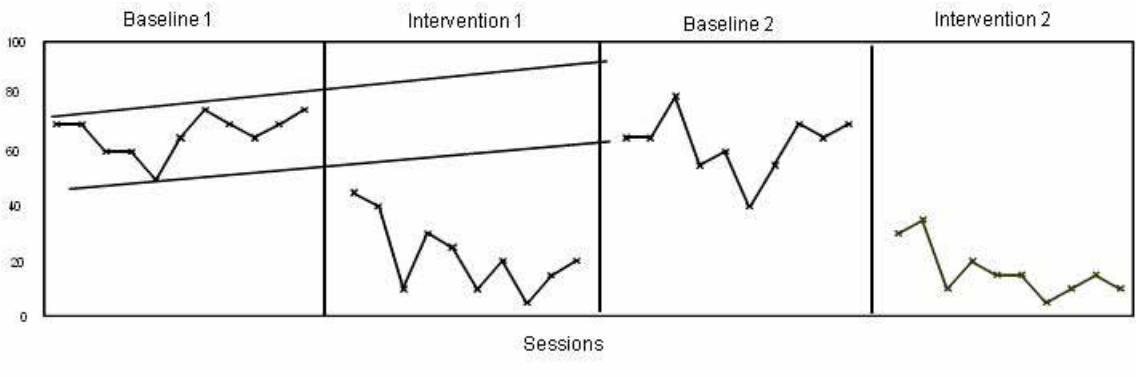


Figure 8B. Examine Observed and Projected Comparison Intervention to Baseline 2

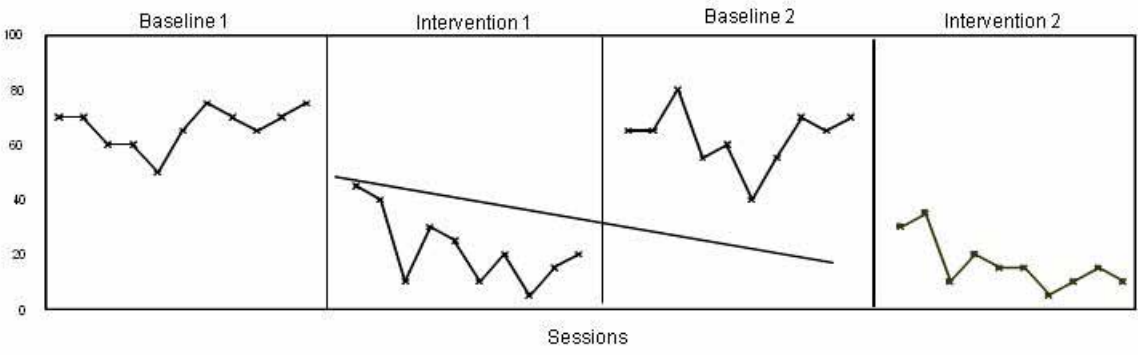


Figure 8C. Examine Observed and Projected Comparison Baseline 2 to Intervention 2