Remedial and Special Education
http://rse.sagepub.com/

Single-Case Intervention Research Design Standards

Thomas R. Kratochwill, John H. Hitchcock, Robert H. Horner, Joel R. Levin, Samuel L. Odom, David M. Rindskopf and William R. Shadish

Remedial and Special Education 2013 34: 26 originally published online 15 August 2012
DOI: 10.1177/0741932512452794

The online version of this article can be found at:
http://rse.sagepub.com/content/34/1/26

Published by:
Hammill Institute on Disabilities

and
SAGE
http://www.sagepublications.com

Hammill Institute on Disabilities

Remedial and Special Education
34(1) 26–38

DOI: 10.1177/0741932512452794
http://rase.sagepub.com
SAGE

Single-Case Intervention Research Design Standards

Thomas R. Kratochwill, PhD[1], John H. Hitchcock, PhD[2], Robert H. Horner, PhD[3], Joel R. Levin, PhD[4], Samuel L. Odom, PhD[5], David M. Rindskopf, PhD[6], and William R. Shadish, PhD[7]

Abstract

In an effort to responsibly incorporate evidence based on single-case designs (SCDs) into the What Works Clearinghouse (WWC) evidence base, the WWC assembled a panel of individuals with expertise in quantitative methods and SCD methodology to draft SCD standards. In this article, the panel provides an overview of the SCD standards recommended by the panel (henceforth referred to as the *Standards*) and adopted in Version 1.0 of the WWC's official pilot standards. The *Standards* are sequentially applied to research studies that incorporate SCDs. The design standards focus on the methodological soundness of SCDs, whereby reviewers assign the categories of *Meets Standards, Meets Standards With Reservations*, and *Does Not Meet Standards* to each study. Evidence criteria focus on the credibility of the reported evidence, whereby the outcome measures that meet the design standards (with or without reservations) are examined by reviewers trained in visual analysis and categorized as demonstrating *Strong Evidence, Moderate Evidence, or No Evidence.* An illustration of an actual research application of the *Standards* is provided. Issues that the panel did not address are presented as priorities for future consideration. Implications for research and the evidence-based practice movement in psychology and education are discussed. The WWC's Version 1.0 SCD standards are currently being piloted in systematic reviews conducted by the WWC. This document reflects the initial standards recommended by the authors as well as the underlying rationale for those standards. It should be noted that the WWC may revise the Version 1.0 standards based on the results of the pilot; future versions of the WWC standards can be found at http://www.whatworks.ed.gov.

Keywords

single-case research design, WWC single-case design standards, design standards, evidence criteria

The demands for accountability in education require the identification of effective, evidence-based interventions. As such, the What Works Clearinghouse (WWC) was established in 2002 by the Institute for Education Sciences (IES) to provide independent evaluations of the internal validity of studies that test the affect of educational and psychological interventions. Although the initial efforts of the WWC focused on studies using group-based methodologies, there has been recent recognition of the importance of single-case research for estimating the effectiveness of certain types of interventions, particularly those aimed at low-incidence problems. The purpose of this article, therefore, is to provide an overview of the recently released Pilot WWC standards for evaluating single-case intervention research to inform policy and practice (Kratochwill et al., 2010). These single-case design (SCD) standards have been designed to complement other standards developed by the WWC Panel for group-based methods such as randomized controlled trials (WWC, 2008) and regression discontinuity designs (Schochet et al., 2010).

The authors of this article are the members of the panel convened by the WWC to develop standards for reviewing evidence based on SCD studies. The standards discussed in this article (henceforth referred to as the *Standards*) were

[1]University of Wisconsin–Madison, USA
[2]Ohio University, Athens, USA
[3]University of Oregon, Eugene, USA
[4]University of Arizona, Tucson, USA
[5]University of North Carolina at Chapel Hill, USA
[6]City University of New York, New York City, USA
[7]University of California, Merced, USA

Corresponding Author:
Thomas R. Kratochwill, PhD, School Psychology Program, University of Wisconsin–Madison, 1025 West Johnson Street, Madison, WI 53706, USA
Email: tomkat@education.wisc.edu

proposed by the panel and adopted as Version 1.0 Pilot Standards by the WWC. These standards are currently being piloted in systematic reviews being conducted by the WWC. As they are tested in various domains of research, we hope to learn about their strengths and limitations. The knowledge acquired during their application could then be used to modify the various design and evidence criteria. Later in the manuscript, we provide some examples of future revisions that might be made to the *Standards*. The official documentation for Version 1.0 Pilot version of the WWC standards and future updates to these standards can be found at http://www.whatworks.ed.gov.

There are several reasons for developing standards for evaluating the causal validity of SCDs. First, as noted above, the WWC has developed standards for review of other research methodologies and was interested in extending scientific knowledge by reviewing the SCD database that is relevant to educational and psychological interventions (Shadish & Sullivan, 2011). In this regard, we were guided by a keen awareness that a large research literature exists on interventions that have been examined in SCD studies, especially in applied and clinical areas of psychology and education (e.g., clinical and school psychology, special education); this SCD literature was not being considered by the WWC. Traditionally, the large-sample randomized controlled trials experiment was held as the gold standard, and SCDs were sometimes relegated to a lower status among traditional methodologists. In the development of *Standards* for SCD research, we wanted to advance a high standard so as to establish the existence of a rigorous database that would compare favorably to traditional large-group designs. In doing so, however, we also were fully aware that not all single-case researchers would applaud our efforts, especially when our proposed criteria are applied to literatures in which the *Standards* are not met. At the same time, there may be critics of the SCD *Standards*, with some traditional methodologists noting that they are not rigorous enough and others suggesting that they are too rigorous.

Second, although other groups and organizations have developed coding criteria for SCDs (such as Divisions 12 [clinical] and 16 [school psychology] of the American Psychological Association; National Reading Panel, 2000), these criteria have not been extended broadly into areas of research reviewed by the WWC. Moreover, to our knowledge, these coding criteria are not being routinely used in most areas of psychology and education. Thus, we were interested in establishing criteria that could be used to standardize reviews of the current literature on various topics and interventions. A recent review by Smith (in press) indicates that the SCD *Standards* compare favorably with other standards that have been developed in psychology and education on such dimensions as research design, measurement, and data analysis.

Third, SCDs have contributed greatly to the "evidence-based practice" movement (Flay et al., 2005), and we deemed that inclusion of these methodologies was important for advancing the scientific knowledge base on educational practices in schools. However, no consensus criteria currently exist for reviewing the SCD literature, including the credibility of various SCDs and a format for showing readers how to become more informed consumers of this literature. Many textbooks on SCD exist (e.g., Barlow, Nock, & Hersen, 2009; Gast, 2010; Johnston & Pennypacker, 2009; Kazdin, 2011), but none of these texts provide coding criteria for research reviews. In the absence of consensus on coding criteria, psychological and educational researchers do not have a standard or baseline from which to design and/or summarize research. The current *Standards* were developed not only to assist in the review of current research but also to increase the quality of future research on important topics in psychology and education.

The SCD Standards

The SCD *Standards* apply to all SCDs, except those that are augmented by including one or more independent comparison cases (i.e., comparison groups, such as classrooms and schools). As depicted in Figure 1, eligible studies are reviewed using the *Design Standards* and are classified as either (a) *Meets Design Standards*, (b) *Meets Design Standards With Reservations*, or (c) *Does Not Meet Design Standards.* Those studies that meet Design Standards with or without reservations are then reviewed using the *Evidence Criteria* to determine whether there is evidence of a functional relation between the independent and outcome variable (see Figure 1). Following the application of the *Evidence Criteria*, the cases within each study will be classified as either demonstrating (a) *Strong Evidence of a Causal Relation*, (b) *Moderate Evidence of a Causal Relation*, or (c) *No Evidence of a Causal Relation.* The specific criteria and procedures used to apply the standards are provided in the following sections.

Criteria for Designs That Meet Standards

The following four criteria are used to assess whether the study's design *Meets Design Standards, Meets Design Standards With Reservations*, or *Does Not Meet Design Standards.* First, a basic protocol for minimizing threats to internal validity requires that an independent variable (i.e., intervention) must be systematically manipulated (i.e., the researcher, rather than some naturally occurring event, must determine when and how changes in independent-variable conditions will occur). If this standard is not met, the study *Does Not Meet Design Standards.*

Second, each outcome variable must be measured systematically over time by more than one assessor. Interobserver

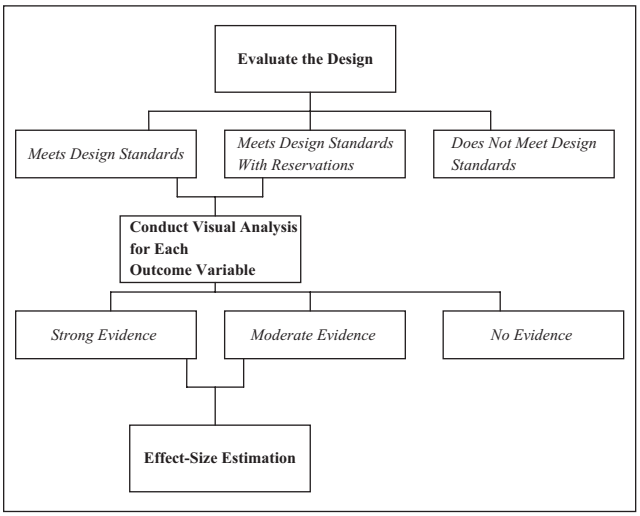

Figure 1. A procedure for applying SCD *Standards*: First evaluate the design and then, if applicable, evaluate the evidence.
Source: What Works Clearinghouse single-case design technical documentation Version 1.0. Available at http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

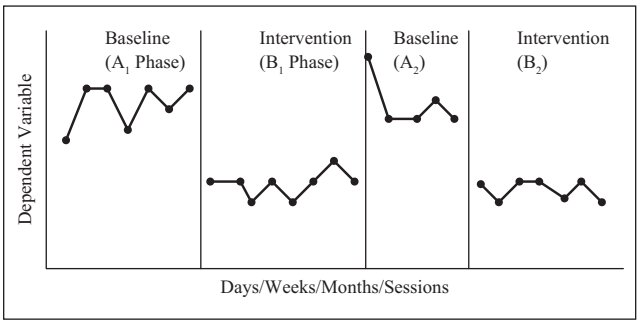

Figure 2. An example of an ABAB single-case intervention research design.
Source: Kratochwill and Levin (2010), reproduced with permission.

agreement (for the dependent variable) must be documented on the basis of an accepted psychometric measure of agreement. Although more than 20 psychometric measures have been proposed to represent interassessor agreement (see Berk, 1979; Suen & Ary, 1989), commonly used techniques 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 interassessor 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, interassessor agreement must be obtained for each case on each outcome variable. A summary of interassessor agreement for a variable must be based on at least 20% of the data points within each condition (e.g., baseline, intervention). If this standard is not met, the study *Does Not Meet Design Standards.*

Third, the study must include at least three attempts to demonstrate an intervention effect each at a different point in time. The heart of interpreting single-case results lies in replication of an unlikely change in the pattern of the data correlated with the researcher either systematically or randomly manipulating the independent variable. A study has the opportunity to demonstrate experimental control (e.g., a functional relation) only if the design provides at least three different opportunities for the researcher to manipulate the independent variable and observe if there is a researcher-predicted change in the pattern of the data. To control effects of confounding variables, these three opportunities must occur at three different points in time. If this procedure is not followed, the study *Does Not Meet Design Standards*. Examples of designs meeting this standard include ABAB designs and their extensions, multiple-baseline designs with at least three baseline conditions, 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. Figures 2–4 illustrate the most common form of SCDs that would meet design standards (i.e., the ABAB, multiple-baseline, and alternating treatment designs, respectively). For alternating and simultaneous treatment designs, the fourth standard that follows supersedes the third standard by effectively requiring five opportunities to demonstrate a treatment effect.

Fourth, for a phase to qualify as an attempt to demonstrate an effect, the phase must include a minimum of three data points (with a preference for at least five). Single-case research examines not only pattern of responding at a moment in time but also the trajectory of responding. The interpretation of single-case data is affected by the trend or slope of the time-ordered data, and as such each phase is assessed to determine whether the phase demonstrates a convincing pattern of responding. In general, the larger the number of data points in a phase, the more confidence there is with respect to the pattern of responding. A phase with fewer than three data points typically offers too little information to allow confident documentation of the pattern of the data. To *Meet Standards*, a reversal/withdrawal design (i.e., ABAB) must include a minimum of four phases per case with at least five data points per phase. To *Meet Standards With Reservations*, a reversal/withdrawal design must include a minimum of four phases per case with three to four data points per phase. Any phases based on fewer than three data points cannot be used to demonstrate existence of or lack of an effect. If the principal investigator conducting the literature review determines that there are exceptions to this standard, the exceptions will be specified

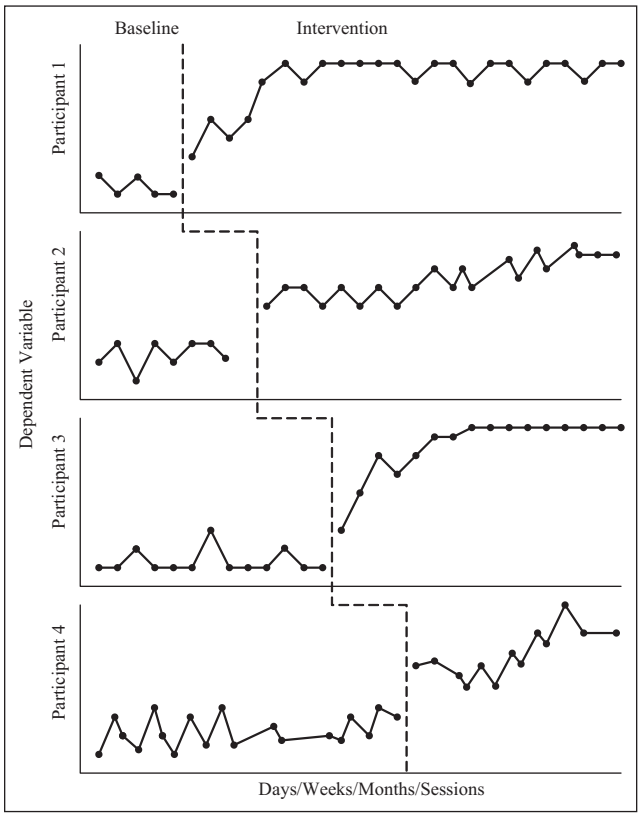

Figure 3. An example of a multiple-baseline single-case intervention research design.
Source: Kratochwill and Levin (2010), reproduced with permission.

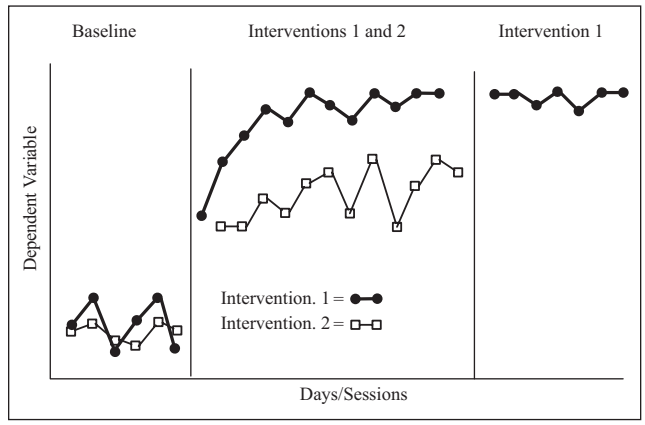

Figure 4. An example of an alternating treatment single-case intervention research design.
Source: Kratochwill and Levin (2010), reproduced with permission.

in the topical area or a practice guide protocol. For example, extreme self-injurious behavior might justify a lower threshold of only one or two data points in a design phase. Thus, there is some flexibility in the application of the design standards.

- To *Meet Standards*, a multiple-baseline design must include a minimum of six phases (i.e., at least three A and three B phases) with at least five data points per phase. To *Meet Standards with Reservations*, a multiple-baseline design must include a minimum of six phases with three to four data points per phase. Any phases based on fewer than three data points cannot be used to demonstrate either existence of or lack of an effect. As of this writing additional criteria have been advanced for variants of the multiple-baseline design, including the nonconcurrent multiple-baseline design and the multiple-probe design.
- Recall that an alternating treatment design requires at least 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 only 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.* In the case of the alternating treatment design, each treatment comparison is rated separately (e.g., A vs. B, A vs. C, and C vs. B in a three-condition design).

Table 1 provides a summary of the design standards for the three most common SCDs.

Comments on Criteria for Designs That Meet Standards

A few additional comments may assist readers in understanding the rationale behind our proposed design standards. First of all, among the various important features of SCD, replication of the intervention is usually considered the most important because it addresses the degree of causal inference that can be established in the experiment. Although there is consensus on the need for replication in the basic types of SCDs, there has been no "gold standard" for the number of replications required to draw conclusions from the experiment. The "three demonstrations" criterion is based on professional convention (R. Horner, Swaminathan, Sugai, & Smolkowski, 2012). More demonstrations further increase confidence that the evidence demonstrates a functional relation between the intervention and the outcome (Kratochwill & Levin, 2010). Although atypical, there might be circumstances in which designs without three replications meet the standards.

As a second consideration, one may be inclined to examine the literature base in major textbooks on SCD for

Table 1. Summary of the What Works Clearinghouse Design Pilot Standards for SCDs, Version 1.0.

Standards	ABAB design	Multiple-baseline design	Alternating treatment design
Meets standards			
Number of phases	4	6	NA
Number of data points per phase	At least 5	At least 5	At most 2 per phase; at least 5 per condition
Meets standards with reservations			
Number of phases	4	6	NA
Number of data points per phase	At least 3	At least 3	At most 2 phase; at least 4 per condition

Note: SCDs = single-case designs.

guidance regarding the issue of the number of replications required in various SCDs. A review of selected texts reveals differences of opinion about this important criterion. As an illustration, Barlow, Nock, and Herson (2009) noted that the ABA (and BAB) design is "experimental," whereas others (e.g., Kazdin, 2011) suggested that the ABAB design is needed to meet the experimental criterion. Likewise, one will also find different opinions about the number of replications required for multiple-baseline and alternating treatment designs. In our review of textbooks, we note that Johnston and Pennypacker (2009), authors of one of the leading traditional behavior analysis SCD texts, suggested that the ABA variant of the "reversal" design provides only "preliminary evidence" related to the reliability of the treatment effect, and they went on to say that even the ABAB design is problematic in drawing firm conclusions. In considering the ABAB design, they noted,

> This design does not provide reassurance that the independent variable is solely responsible for the associated changes in responding. Someone could still argue that these selected behavioral changes are due to extraneous factors initiated by or selectively associated with the independent variable, and we would have no evidence for a convincing rebuttal. Again, investigators may sometimes find this a difficult constraint to accept. When you have labored to plan and implement a study and have watched everything closely from day to day, it is tempting to believe that the independent variable embedded in the intervention condition is solely responsible for the observed changes in responding. Even colleagues reading a published report of the study might want to make the same assumption, especially if they are also interested in a certain outcome. Nevertheless, the fact of the matter is that neither an ABA nor ABAB design establishes a functional relation because it does not address the role of extraneous factors selectively associated with the independent variable. (Johnston & Pennypacker, 2009, pp. 264–265)

The authors further consider ways to evaluate and control extraneous variables that may compromise experimental causal inference (e.g., variables that occur concurrently with the intervention or B phases). The panel's recommendations for dealing with these causal-inference concerns were however considered within the context of internal validity threats (see Shadish, Cook, & Campbell, 2002; Appendix A of Kratochwill et al., 2010).

Third, some traditional methodologists may contend that even SCDs with replication will never be considered completely experimental (i.e., a methodology that is capable of establishing strong causal inference) because randomization is not incorporated into the basic design structure. As the argument goes, only random assignment of participants to treatment conditions (e.g., experimental and control conditions in the traditional randomized experiment)—and in some contexts, orders of treatment administration—can help reduce (but not eliminate) certain major threats to internal validity. Yet, SCDs can be structured with various forms of randomization, in intervention research that can be regarded as experimental, to improve causal inference (Kratochwill & Levin, 2010). The panel was fully aware that such a standard for SCDs is currently unrealistic given the paucity of investigations that incorporate randomization into experiments. Thus, in many respects, our current focus on replication represents a "Goldilocks" ("just right") position in standards for review of current research. Having said that, we believe the *Standards* err on the side of requiring evidence that makes statements about the presence of a treatment effect (or for that matter, the absence of an effect) the most plausible explanation for the observed data.

Visual Analysis of Single-Case Research Results

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 (Gast, 2010; 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). The emphasis on visual analysis for evaluating evidence of single-case research, rather than quantitative metrics, was decided for several reasons. First, the vast majority of research conducted with SCDs has applied visual analysis to data outcomes and interpretations (Kratochwill, Levin, Horner, & Swoboda, in press). As authors, we wished to develop criteria for visual analysis that would align with the body of evidence in published studies. Second, currently there are no agreed upon criteria for statistical analysis of single-case data. A wide range of methods have been applied, often with violations of statistical assumptions (e.g., independence of error components in the data) or fundamental differences in outcomes when applied to the same data set. Nevertheless, our emphasis on visual analysis does not preclude the use of an appropriate statistical test in SCDs (as one example of many available, the reader is referred to Campbell & Herzinger, 2010, for a review of various statistical procedures that have been and can be used in SCDs). Third, there is also considerable disagreement over the "appropriate" calculation of effect sizes (ESs) as applied to single-case data (see Shadish, Rindskopf, & Hedges, 2008, for a detailed discussion, as well as a later section of this article).

Following the tradition of visual analysis of data displays in SCDs (Parsonson & Baer, 1978), the rules used in the SCD *Standards* for conducting visual analysis involve examining the overall graph, as well as considering four steps and six features of the outcome measure. Step 1 is documentation of a predictable and stable baseline pattern of data (e.g., the student is consistently reading with many errors; the student is consistently engaging in high rates of disruption). If a convincing baseline pattern is documented, then Step 2 consists of examining the data within each phase of the study to assess the within-phase pattern(s). The key issue here is to assess whether there is a sufficient amount of data with sufficient consistency to demonstrate a predictable pattern of responding (i.e., level or trend). Step 3 in the visual-analysis process is to compare the data from each phase with the data in the adjacent (or a similar) phase to assess whether manipulation of the independent variable can be plausibly tied to an "effect." An effect is demonstrated if manipulation of the independent variable is associated with predicted change in the pattern of the dependent variable (with temporal proximity between the two taken into account as well). The fourth step in visual analysis is to integrate 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)—see R. Horner et al. (2012).

To assess the effects within SCDs, six outcome-measure features are used to examine within- and between-phase data patterns: (a) level, (b) trend, (c) variability, (d) immediacy of the effect, (e) overlap, and (f) consistency of data patterns across similar phases (Fisher, Kelley, & Lomas, 2003; Hersen & Barlow, 1976; Kazdin, 1982; Kennedy, 2005; 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 overall average (mean) of the outcome measures within a phase. "Trend" refers to the slope of the best-fitting straight line for the outcome measures within a phase, and in the context of visual analysis, "variability" typically refers to the range, variance, or standard deviation of the outcome measures about the best-fitting line (whether linear or curvilinear). Variability might also refer simply to the degree of overall scatter, in the sense that too much variability of the data points undermines one's ability to make a reasoned prediction in the absence of treatment manipulation. Examination of the data within a phase is used (a) to describe the observed pattern of a case's performance and (b) to extrapolate the expected performance forward in time, assuming that no changes in the independent variable were to occur (Furlong & Wampold, 1981). The six visual-analysis features are used collectively for comparison of projected and expected patterns 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 outcome measures within each phase, the reviewer 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 "three data points" criterion is somewhat arbitrary, with the reviewers having some discretion in this matter. 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 study. However, predicted delayed effects or gradual effects of the intervention may be built into the design of the experiment, which in turn would influence decisions about phase length for a particular study. "Overlap" refers to the proportion of data from one phase that overlaps with data from the previous phase (the data may need to be "detrended" in cases where trends exist within phases). The smaller the proportion of overlapping data points (or conversely, the greater the non-overlap), the more compelling the demonstration of an effect. There are, of course, exceptions to this guideline, such as when across two adjacent phases there is a reversal in trend that not only produces overlap but also demonstrates a therapeutic 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 plausible it is that the outcomes can be attributed to the manipulated independent variable and the more likely the data represent a causal relation. These six features are assessed individually and collectively to determine whether the results from a single-case study plausibly demonstrate a causal relation.

Regardless of the type of SCD used in a study, visual analysis of (a) level, (b) trend, (c) variability, (d) overlap, (e) immediacy of the effect, and (f) consistency of data patterns across similar phases is used to assess whether there are at least indications of an effect at three 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 functionally related to manipulation of the independent variable.

Integrating Design and Evidence Criteria to Assess Experimental Control

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 functional relation between an independent variable (i.e., an intervention) and an outcome variable. To provide *Strong Evidence*, at least two reviewers certified in visual analysis must verify that a functional relation was documented. The certification occurs following a training session and performance on visual-analysis tasks. Specifically, this "evidence of a functional relation" criterion is operationalized as a study exhibiting at least three demonstrations of the intervention effect, each occurring at a different point in time, combined with no failures to observe and effect, by

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

If a SCD does not provide at least three, temporally distinct, demonstrations of an effect, then the study is rated as providing *No Evidence.* If a study provides three demonstrations of an effect and also includes at least one demonstration of a noneffect (this is possible in, for example, a multiple-baseline design with four baselines), the study is rated as providing *Moderate Evidence.* The following characteristics must be considered when identifying a noneffect:

- Data within the baseline phase do not demonstrate a stable enough pattern that can be compared with data patterns obtained in subsequent phases;
- Failure to establish a consistent pattern within any phase (e.g., high variability of outcomes within a phase);
- Difficulty in determining whether the intervention is responsible for a claimed effect as a result of either (a) long latency between introduction of the independent variable and change in the outcome variable or (b) overlap between observed and projected patterns of the outcome variable between baseline and intervention phases;
- Inconsistent patterns across similar phases (e.g., an ABAB design in which the outcome variable data points are high during the first B phase but low during the second B phase);
- Major discrepancies between the projected and observed between-phase patterns of the outcome variable; and
- When examining the outcomes of a multiple-baseline design, reviewers must also consider the extent to which the time at which a basic effect is initially demonstrated with one series (e.g., first 5 days following introduction of the intervention for Case 1) is associated with change in the data pattern over the same time frame in the other series of the design (e.g. same 5 days for Cases 2, 3, and 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 ES estimation follows.

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, especially given considerations of internal validity threats (R. H. 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 (R. Horner & Spaulding, 2010). The SCD Panel recommended a 5-3-20 threshold for including a body of SCDs in a systematic evidence review that can be designated as an evidence-based practice:

1. A minimum of five SCD studies examining the intervention with a design that either *Meets Standards or Meets Standards With Reservations*.
2. The SCD studies are conducted by at least three research teams with no overlapping authorship at three different institutions.
3. The combined number of cases (i.e., participants, classrooms, etc.) totals at least 20.

The term "studies" (in this context also regarded as "experiments" or "investigations") refers to analyses organized around answering a specific research question. The five-study criterion can be fulfilled in various ways. For example, the same target behavior of five participants may be examined in five separate experiments incorporating an ABAB design. However, five different target behaviors for the same participant, aimed at answering a particular research question in a single ABAB design, would not qualify. A research article may include one or more than one study, experiment, or investigation in it. Reviewers can rely on authors of the original research to indicate when this situation is the case. The author(s) would clearly describe the separate research questions, describe the method for addressing each research question, and report analyses specific to each research question.

Although the 5-3-20 criterion might be considered arbitrary by some researchers, there is some precedent for this criterion (R. H. Horner et al., 2005; Odom, Collet-Klingenberg, Rogers, & Hatton, 2010). Nevertheless, there is no denying that we could have adopted a criterion more or less demanding than the one selected. In the absence of a widely adopted standard in the field of psychology and education for an evidence-based practice, we believe that there must be a starting point from which to draw conclusions from a SCD literature review. In the future, the 5-3-20 criterion will likely be reconsidered in light of research reviews conducted to provide insights into its reasonableness.

Note that the 5-3-20 criterion takes into account the design standards. In addition, reviewers can investigate details about authorship and institution, documenting any unusual circumstances in a report on a particular intervention or practice. Once the 5-3-20 criterion is met, reviewers consider the evidence for a particular intervention as the next step in the review process. In Version 1.0 of the Pilot WWC SCD standards, the 5-3-20 rule must be met before the findings of the review of SCDs can be summarized.

Because appropriate SCD ES estimators are still being developed (see the next section), details for combining group and SCD ESs have not yet been provided by IES.

A Sample Application of the SCD Standards

With release of the WWC Version 1.0 Pilot SCD standards, there have been some reviews on various topics in the SCD literature using these criteria. As a recent example, Maggin, Chafouleas, Goddard, and Johnson (2011) reviewed the literature on the use of token economies as an intervention for students demonstrating behavior difficulties in school settings. Out of 834 studies examined, the authors found that 24 met the review criteria, based on a total of 90 cases (with the cases consisting of 67 individual students and 23 classrooms). Overall, the authors reported that research on token economies does not support evidence for a "best practice" in classroom management—contradicting early reviews in which it had been concluded that in many other areas of treatment, token economies have produced positive outcomes (e.g., Kazdin, 1982).

Specific findings from the Maggin et al. (2011) review help to illustrate the application of the *Standards* and certain considerations of the criteria as they are applied to a topical area of research. First, in terms of design standards, the most common limitation of the research was related to the reporting of interobserver agreement (i.e., researchers failed to report agreement data across phases, sessions, or both). This criterion was followed by limitations in designs not meeting the replication requirement in the design standards (such as AB and ABA designs). In addition, the authors reported that for two cases, there were insufficient data points within a phase. Four studies demonstrated strong evidence (one student case and three classroom cases) and eight (all student cases) demonstrated moderate evidence. Finally, in application of the 5-3-20 criterion, the authors reported that there is still insufficient evidence to support token economies for either individual student or classroom applications based on the number of studies, replication across investigators, and number of participants.

Certain considerations in these findings deserve further comment. To begin with, the SCD *Standards* were applied to a relatively old research literature on token economies (i.e., 11 of the studies reviewed were published prior to 1980 and only 3 studies reviewed were published in 2005 or later). On further analysis, in the token economy research review, we expected that SCD research has generally improved in meeting design standards over the past decade. Additional information obtained from the authors bore out this prediction: Findings revealed that the cases that met design standards with or without reservations all came from studies conducted in 1987 or later. The specific breakdown of case type by year was as follows:

- Classes as cases—1987 and 1990
- Individual students as cases—2004, 2001, 1990, and 2007. The 2007 study was a dissertation and not all cases met standards due to attrition and logistics; however, those students who completed the study, all met standards (Maggin, personal communication, January 13, 2012).

Thus, when reporting the results of literature reviews in the future, some consideration might be given to placing greater weight on more recent research that addresses a broad spectrum of improvements that may occur in SCD methodology. The methodological improvements in later research reviewed by Maggin et al. (2011) generally mirror findings reported by Sullivan and Shadish (2011) in their review of 616 SCD studies published in 2008. The authors found that approximately 45% of the SCDs met the strictest criteria for the Version 1.0 Pilot WWC SCD standards, with an additional 30% meeting the Pilot WWC standards with reservation. Nevertheless, they also found that although interobserver agreement was generally acceptable in published SCD research (i.e., 94% of the research included interobserver agreement over 20% of the sessions), the "number of data points" criterion in baseline and intervention phases was less likely to be met.

It is also noteworthy that interobserver agreement and replication were not the only methodological shortcomings of research on token economies in the Maggin et al. (2011) review. The authors pointed to the failure of researchers to report information on treatment integrity (see Sanetti & Kratochwill, 2009). In the absence of data on the integrity/fidelity of the intervention, little can be concluded from the extant research about whether it was implemented properly and at sufficient dosage levels. Although the Version 1.0 Pilot WWC SCD standards do not have a separate category for coding treatment integrity, it is taken into account during the review process by reviewers.

ES Estimates for SCDs

ES estimates are available for most designs involving traditional group intervention research, and in associated meta-analyses, there is widespread agreement about how these ESs 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's *g*). This situation is not true for SCDs in that researchers have not reached consensus on appropriate methods or standards for ES estimation. What follows is a brief summary of the main issues, with a more extensive discussion in an article by Shadish et al. (2008).

Several issues are involved in creating ES estimates. First is the general issue of how to quantify the size of an effect. One can quantify the effect for a single case, 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 ESs. Next is the issue of comparability of different ESs for SCDs. Finally, there is the issue of the comparability of ES estimates for SCDs and for group-based designs.

Most researchers implementing SCDs still base their inferences on visual analysis, even though several quantitative ES methods have been proposed. Each has flaws, but some methods are likely to be more useful than others. The panel recommended using some of these current methods until better ones have been developed. Note that Maggin et al. (2011) reported four different ES statistics in their review, including percentage of nonoverlapping data (PND), improvement rate difference, standardized mean difference, and raw-data multilevel ESs (see pp. 538–539 for an overview of their rationale).

A number of ad hoc methods have been used to analyze SCDs (e.g., PND, percentage of all nonoverlapping data [PAND], percent exceeding the median [PEM]). Some of these methods have been accompanied by efforts to convert them to parametric estimators such as the phi coefficient (φ), which might in turn be comparable to traditional group-study 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 of such efforts make the erroneous assumption that these 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, and so the data would need to be detrended with, say, first-order differencing before computing the index. Lacking a formal inverse–variance weighting system, one could combine the results with ordinary unweighted averages or one could weight by the number of cases in a study.

Several formally justified parametric methods have been proposed, including regression estimates and multilevel models. Regression estimates have three advantages. First, many primary researchers are familiar with regression, and so both the analyses and the results are likely to be easily understood. Second, these methods can model trends in the data, and so they do not require prior detrending of the data. Third, regression methods can be applied to obtain an ES from a single case, whereas multilevel models require several cases within a study. However, 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 likely to be present in SCD data. For multilevel models, many researchers are less familiar with the analytic methods and the interpretation of results, so that their widespread use is probably less likely than with regression. In addition, practical implementation of multilevel models for SCDs is more technically challenging than ordinary regression, likely

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 ES estimates based on traditional group studies, and so one could not compare ESs from SCDs with those from group studies.

An exception to the latter statement 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 ES estimator because the data are already in the same metric. However, other technical problems remain: Estimators are still not comparable with those from traditional group studies and such instances tend to be rare across studies.

As we have pointed out several times, the quantitative methods that have been proposed are not in the same metric as 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 standard deviation or a pooled within-group standard deviation. These variances reflect variation across people. In contrast, SCDs, by definition, involve comparison of behavior within an individual (or other entity), across different conditions. Attempts to standardize these effects have usually involved dividing by some version of a within-phase standard deviation, 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 ES index such as those mentioned 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 ES metric is used. This latter hypothesis could be partially tested by computing more than one of these indices and comparing their rank orderings.

An ES estimator for SCDs that is comparable to those used in traditional group studies is badly needed. Shadish et al. (2008; Hedges, Shadish, & Pustejovsky, 2011) have developed an estimator for continuous outcomes that is promising in this regard, though the distribution theory is still being derived and tested. A preliminary application of this estimator to a program for treating autistic children suggests that it does yield ESs that are comparable with those from randomized between-groups designs (Shadish, Hedges, & Pustejovsky, 2011). However, the method is so new that a number of problems need work (although these problems also apply to most of the estimators described previously). The small number of cases in most SCDs would make such an estimate imprecise (i.e., it would have a large standard error and an associated wide confidence interval). Furthermore, 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 limitation to using the Shadish et al. (2008) procedure—although this problem applies to nearly every other estimator that has been proposed. Fortunately, some reason exists to think that normal approximations to these data might work reasonably well. Finally, this method requires detrending the data prior to ES estimators. Hence, it is premature to advise use of this method except to investigate further their statistical properties.

Until all these methods receive more thorough investigation, we suggest the following guidelines for estimating ESs in SCDs. First, in those cases in which all the outcome measures are already in a common metric (such as proportions or rates), these may be preferred to the existing standardized ES estimators. Second, if only one standardized ES estimate is to be chosen, the regression-based estimators are best justified from technical and practical points of view given the fit with visual-analysis logic. Third, we strongly recommend conducting sensitivity analyses. For example, one could report one or more nonparametric estimates (but not the PND estimator, because it has undesirable statistical properties; Wolery, Busick, Reichow, & Barton, 2010) 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 ESs) 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.

Summary and Conclusions

In this article, we reviewed the pilot standards developed by the WWC for SCDs. As authors, we perceive the Pilot WWC standards as a work in progress in that their application to the research review process will likely yield need for modification and additional criteria for coding studies. In addition, there were several issues that the WWC Version 1.0 Pilot SCD standards do not address; these

issues and any new criteria may be included in future versions of the standards. First, as noted above, there is no consensus on the "correct" form of an ES estimate for SCD (Shadish et al., 2008). It bears repeating that in the future, a high priority is that researchers provide a stronger statistical justification for one or more ES estimates, as well as advance strategies for combining SCD summaries with those based on traditional group studies. Clearly, this is an area of great importance if evidence reviews that use the WWC Pilot standards are to incorporate evidence from several study design types.

Second, and as noted at the outset, the *Standards* apply primarily to SCDs in which the "case" is an individual participant. Although the *Standards* apply to SCDs in which a group (e.g., classroom, school) comprises the case, special issues can be raised when a group or other aggregate is used in the experiment. Consider a multiple-baseline design across participants in which the researcher implements an intervention across four classrooms. In addition to the proposed SCD *Standards*, the investigator adopting this design should consider other validity issues that are specifically related to group investigations including, for example, differential selection, differential attrition, control of extraneous variables, and the possibility of within-school contagion effects. Reporting of results when a group is the case should also prompt the researcher to consider additional factors such as group composition, within-group variance, nonresponders to treatment, and a rationale for the metric used to report the data in the graph, among other issues.

Third, the *Standards* recommend replication as a basis for drawing valid inferences for SCDs. Recently, a class of designs in which some form of randomization can be incorporated has been proposed for single-case researchers (Kratochwill & Levin, 2010). As was noted previously, such designs are now used infrequently in applied and clinical research in psychology and education despite their advantages for improving the internal validity of SCDs. In the future, these designs could be included in the *Standards*, along with a set of criteria for their review. Aside from increasing their validity, randomized SCDs are associated with a class of randomization statistical tests that can improve the statistical conclusion validity of the research study (see, for example, Edgington & Onghena, 2007; Todman & Dugard, 2001).

Fourth, with respect to data-analysis methods, the *Standards* focus exclusively on the use of visual analysis as these procedures apply to the majority of published research using SCDs without having to deal with the difficult statistical issues we outlined previously. Strategies to improve visual analysis, including the training of reviewers in these methods, were provided as part of the process of development of the *Standards* (Horner & Spaulding, 2010). Nevertheless, there are a number of considerations in visual analysis of single-case data, among which are reliability and validity of the process (Kratochwill et al., 2011). Research on visual analysis of data demonstrates some inconsistent findings. For example, Lieberman, Yoder, Reichow, and Wolery (2010) tested various characteristics of multiple-baseline designs to determine whether the data features affected the judgments of visual-analysis experts (N = 36 editorial board members of journals who publish SCDs) regarding the presence of a functional relation and agreement on the outcomes. It was found that graphs with steep slopes (versus shallow slopes) when the intervention was introduced were judged as more often having a functional relation. Nevertheless, there was still some disagreement on whether the functional relation had been established. Lieberman et al. noted that training visual judges to address conditions in which there is change long after the intervention, and where there is inconsistent latency of change across units, may be helpful in reviewers' concurrence about a functional relation. Kahng et al. (2010) replicated and extended earlier research on visual analysis by including editorial board members of the *Journal of Applied Behavior Analysis* as participants in the study. Board members were asked to judge 36 ABAB design graphs on a 100-point scale while rating the degree of experimental control. These authors reported high levels of agreement among judges, noting that the reliability of visual analysis has improved over the years, due in part to better training in visual-analysis methods.

Fifth, aside from the psychometric and statistical concerns with visual analysis, visual analysis of data is typically conducted in a *response-guided* format, wherein decisions are made about the point of intervention in the data series based on the preceding data pattern, taking into account changes in trend, variability, and level. Response-guided procedures (in contrast, to either an a priori decision or determining the intervention point at random) have been criticized for introducing bias and increasing Type I errors in the research (Ferron & Jones, 2006; Todman & Dugard, 2001). Options for dealing with these issues have been proposed by Ferron and Jones (2006), but these options have not yet been widely adopted in actual SCD investigations.

Sixth, the Version 1.0 WWC SCD Pilot standards did not elaborate upon the range of possible statistical methods that might be applied to SCDs. Over the years, a large number of procedures have been recommended including, for example, analysis of variance, time-series analysis, hierarchical linear modeling, and nonparametric randomization tests (see Campbell & Herzinger, 2010; Kazdin, 2011; Kratochwill, 1978; Kratochwill & Levin, 1992). Nevertheless, there is considerable disagreement among statisticians on many of these procedures, generally because of violating the distributional assumptions of traditional methods when they are applied to single-case data. The autocorrelated nature of

time-series data poses special challenges for SCD researchers who attempt to adopt one or more of these statistical applications.

We close with two general conclusions. One is that continued work is needed to improve primary SCD work as well as efforts to summarize findings across SCD studies via research syntheses. A review of the research design literature shows that this is equally true for other types of investigations. A second point is that SCDs represent an important part of our empirical understanding of interventions, particularly when dealing with small samples sizes. SCDs can yield unambiguous causal evidence for particular cases, but their greatest limitation is that most studies are based on small samples. This perspective justifies our efforts (as well as prior efforts) to summarize systematically SCD evidence and continue to explore ways for better combining SCD findings with findings from other types of studies. With such efforts, researchers will be able to offer empirically sound commentary about interventions that do and do not work.

Acknowledgments

The authors express appreciation to Scott Cody, Steve Lipscomb, and Shannon Monahan at Mathematica Policy Research for their assistance with our work on the single-case design panel.

Authors' Note

The information contained herein is based on the What Works Clearinghouse's (WWC's) *Single-case design (SCD) technical documentation Version 1.0* (pilot) produced by the current authors and available at http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf. The standards that are described in the technical documentation were developed by the authors for the Institute of Education Sciences (IES) under Contract ED-07-CO-0062 with Mathematica Policy Research, Inc., to operate the WWC. The content of this article does not necessarily represent the views of the IES or of the WWC. The Version 1.0 SCD standards are currently being piloted in systematic reviews conducted by the WWC. This document reflects the initial standards recommended by the authors as well as the underlying rationale for those standards. It should be noted that the WWC may revise the Version 1.0 standards based on the results of the pilot; future versions of the WWC standards can be found at http://www.whatworks.ed.gov. With the exception of the first author, all authors are listed alphabetically.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single-case experimental designs: Strategies for studying behavior change*. Allyn and Bacon.

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

Campbell, J. M., & Herzinger, C. V. (2010). Statistics and single-subject research methodology. In D. L. Gast (Ed.), *Single-subject research methodology in behavioral sciences* (pp. 417–453). New York, NY: Routledge.

Edgington, E. S., & Onghena, P. (2007). *Randomization tests* (4th ed.). Baca Raton, FL: Chapman & Hall/CRC.

Ferron, J., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *Journal of Experimental Education*, *75*, 66–81.

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.

Flay, B. R., Biglan, A., Boruch, R. F., Castro, F. G., Gottfredson, D., Kellam, S., … Ji, P. (2005). Standards of evidence: Criteria for efficacy, effectiveness and dissemination. *Prevention Science*, *6*, 151–175.

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

Gast, D. L. (2010). *Single subject research methodology in behavioral sciences*. New York, NY: Routledge.

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.

Hedges, L. V., Shadish, W. R., & Pustejovsky, J. (2011, July). *Effect sizes for single-case designs*. Paper presented at the 5th annual meeting of the Society for research synthesis methodology, Ottawa, Ontario, Canada.

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

Horner, R., & Spaulding, S. (2010). Single-case research designs. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1386–1394). Thousand Oaks, CA: SAGE.

Horner, R., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis of single-case research. *Education and Treatment of Children, 35*, 269-290.

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.

Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). New York, NY: Routledge.

Kahng, S. W., Chung, K.-M., Gutshall, K., Pitts, S. C., Kao, J., & Girolami, K. (2010). Consistent visual analysis of intrasubject data. *Journal of Applied Behavior Analysis*, *43*, 35–45.

Kazdin, A. E. (1982). The token economy: A decade later. *Journal of Applied Behavior Analysis*, *15*, 431–445.

Kazdin, A. E. (2011). *Single-case research designs: Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.

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

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

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

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

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*, 122–144.

Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swoboda, C. (in press).Visual analysis of single-case intervention research: Conceptual and methodological considerations. In T. R. Kratochwill and J. R. Levin (Eds.) *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association.

Lieberman, R. G., Yoder, P. J., Reichow, B., & Wolery, M. (2010). Visual analysis of multiple baseline across participants graphs when change is delayed. *School Psychology Quarterly*, *25*, 28–44.

Maggin, D. M., Chafouleas, S. M., Goddard, K. M., & Johnson, A. H. (2011). A systematic evaluation of token economies as a classroom management tool for students with challenging behavior. *Journal of School Psychology*, *49*, 529–554.

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

National Reading Panel. (2000). *Report of the national reading panel: Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Jessup, MD: National Institute for Literacy.

Odom, S. L., Collet-Klingenberg, L., Rogers, S. J., & Hatton, D. D. (2010). Evidence-based practices in interventions for children and youth with autism spectrum disorders. *Preventing School Failure*, *54*, 275–282.

Parsonson, B., & Baer, D. (1978). The analysis and presentation of graphic data. In T. Kratochwill (Ed.), *Single subject research* (pp. 101–166). New York, NY: 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.

Sanetti, L. M. H., & Kratochwill, T. R. (2009). Toward developing a science of treatment integrity: Introduction to the special series. *School Psychology Review*, *38*, 445–459.

Schochet, P., Cook, T., Deke, J., Imbens, G., Lockwood, J. R., Porter, J., & Smith, J. (2010). *Standards for regression discontinuity designs*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/wwc_rd.pdf

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

Shadish, W. R., Hedges, L. V., & Pustejovsky, J. (2011, July). *An application of a new d-estimator for single-case designs*. Presented at the Society for Research Synthesis Methodology. Ottawa, Ontario, Canada.

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.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*, 971–980. doi:10.3758/s13428-011-0111-y

Smith, J. D. (in press). *Single-case experimental designs: A systematic review of published research and recommendations for researchers and reviewers*.

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

Sullivan, K. J., & Shadish, W. R. (2011). *An assessment of single-case designs by the What Works Clearinghouse standards*. Manuscript submitted for publication.

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

Todman, J. B., & Dugard, P. (2001). *Single-case and small-n experimental designs: A practical guide to randomization tests*. Mahwah, NJ: Lawrence Erlbaum.

What Works Clearinghouse. (2008). *Procedures and standards handbook* (Version 2.0). Retrieved from http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19

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

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *Journal of Special Education*, *44*, 18–28.