Published in final edited form as:
Am J Speech Lang Pathol. 2012 November ; 21(4): 397–414. doi:10.1044/1058-0360(2012/11-0036).

Single-Subject Experimental Design for Evidence-Based Practice

Breanne J. Byiers[a], **Joe Reichle**[a], and **Frank J. Symons**[a]

[a]University of Minnesota, Minneapolis

Abstract

Purpose—Single-subject experimental designs (SSEDs) represent an important tool in the development and implementation of evidence-based practice in communication sciences and disorders. The purpose of this article is to review the strategies and tactics of SSEDs and their application in speech-language pathology research.

Method—The authors discuss the requirements of each design, followed by advantages and disadvantages. The logic and methods for evaluating effects in SSED are reviewed as well as contemporary issues regarding data analysis with SSED data sets. Examples of challenges in executing SSEDs are included. Specific exemplars of how SSEDs have been used in speech-language pathology research are provided throughout.

Conclusion—SSED studies provide a flexible alternative to traditional group designs in the development and identification of evidence-based practice in the field of communication sciences and disorders.

Keywords

single-subject experimental designs; tutorial; research methods; evidence-based practice

The use of single-subject experimental designs (SSEDs) has a rich history in communication sciences and disorders (CSD) research. A number of important studies dating back to the 1960s and 1970s investigated fluency treatments using SSED approaches (e.g., Hanson, 1978; Haroldson, Martin, & Starr, 1968; Martin & Siegel, 1966; Reed & Godden, 1977). Several reviews, tutorials, and textbooks describing and promoting the use of SSEDs in CSD were published subsequently in the 1980s and 1990s (e.g., Connell, & Thompson, 1986; Fukkink, 1996; Kearns, 1986; McReynolds & Kearns, 1983; McReynolds & Thompson, 1986; Robey, Schultz, Crawford, & Sinner, 1999). Despite their history of use within CSD, SSEDs are sometimes overlooked in contemporary discussions of evidence-based practice. This article provides a comprehensive overview of SSEDs specific to evidence-based practice issues in CSD that, in turn, could be used to inform disciplinary research as well as clinical practice.

Correspondence to Breanne Byiers: byier001@umn.edu.

In the current climate of evidence-based practice, the tools provided by SSEDs are relevant for researchers and practitioners alike. The American Speech-Language-Hearing Association (ASHA; 2005) promotes the incorporation of evidence-based practice into clinical practice, defining evidence-based practice as "an approach in which current, high-quality research evidence is integrated with practitioner experience and client preferences and values into the process of making clinical decisions." The focus on the individual client afforded by SSEDs makes them ideal for clinical applications. The potential strength of the internal validity of SSEDs allows researchers, clinicians, and educators to ask questions that might not be feasible or possible to answer with traditional group designs. Because of these strengths, both clinicians and researchers should be familiar with the application, interpretation, and relationship between SSEDs and evidence-based practice.

The goal of this tutorial is to familiarize readers with the logic of SSEDs and how they can be used to establish evidence-based practice. The basics of SSED methodology are described, followed by descriptions of several commonly implemented SSEDs, including their benefits and limitations, and a discussion of SSED analysis and evaluation issues. A set of standards for the assessment of evidence quality in SSEDs is then reviewed. Examples of how SSEDs have been used in CSD research are provided throughout. Finally, a number of current issues in SSEDs, including effect size calculations and the use of statistical techniques in the analysis of SSED data, are considered.

The Role of SSEDs in Evidence-Based Practice

Numerous criteria have been developed to identify best educational and clinical practices that are supported by research in psychology, education, speech-language science, and related rehabilitation disciplines. Some of the guidelines include SSEDs as one experimental design that can help identify the effectiveness of specific treatments (e.g., Chambless et al., 1998; Horner et al., 2005; Yorkston et al., 2001). Many research communities, however, hold the position that randomized control trials (RCTs) represent the "gold standard" for research methodology aimed at validating best intervention practices; therefore, RCTs de facto become the only valid research methodology that is necessary for establishing evidence-based practice.

RCTs do have many specific advantages related to understanding causal relations by addressing methodological issues that may compromise the internal validity of research studies. Kazdin (2010), however, compellingly argued that certain characteristics of SSEDs make them an important addition and alternative to large-group designs. He argued that RCTs may not be feasible with many types of interventions, as resources for such large-scale studies may not be available to test the thousands of treatments likely in use in any given field. In addition, the carefully controlled conditions in which RCTs must be conducted to ensure that the results are interpretable may not be comparable and/or possible to implement in real-life (i.e., uncontrolled) conditions. SSEDs are an ideal tool for establishing the viability of treatments in real-life settings before attempts are made to implement them at the large scale needed for RCTs (i.e., scaling up). Ideally, several studies using a variety of methodologies will be conducted to establish an intervention as evidence-based practice. When a treatment is established as evidence based using RCTs, it is often

interpreted as meaning that the intervention is effective with most or all individuals who participated. Unfortunately, this may not be the case (i.e., there are responders and nonresponders). Thus, systematic evaluation of the effects of a treatment at an individual level may be needed, especially within the context of educational or clinical practice. SSEDs can be helpful in identifying the optimal treatment for a specific client and in describing individual-level effects.

Analysis of Effects in SSEDs

Desirable Qualities of Baseline Data

The analysis of experimental control in all SSEDs is based on visual comparison between two or more conditions. The conditions tested typically include a baseline condition, during which no intervention is in place, as well as one or more intervention conditions. The baseline phase establishes a benchmark against which the individual's behavior in subsequent conditions can be compared. The data from this phase must have certain qualities to provide an appropriate basis for comparison. The first quality of ideal baseline data is stability, meaning that they display limited variability. With stable data, the range within which future data points will fall is predictable. The second quality of ideal baseline data is a lack of a clear trend of improvement. The difficulty posed by trends in baseline data is dictated by the direction of behavior change expected during the intervention phase: If the behavior reflected in the dependent measure is expected to increase as a result of the intervention, a decreasing trend during baseline does not pose a significant problem. If, on the other hand, the trend for the dependent measure is increasing during baseline, determining whether or not a continued increase during the intervention phase constitutes a treatment effect is likely to be compromised. By convention, a minimum of three baseline data points are required to establish dependent measure stability (Kazdin, 2010), with more being preferable. If stability is not established in the initial sessions, additional measurements should be obtained until stability is achieved. Alternatively, steps can be taken to introduce additional controls (strengthening internal validity) into the baseline sessions that may contribute to variability.

Visual Data Inspection as a Data Reduction Strategy: Changes in Level, Trend, and Variability

Once the data in all conditions have been obtained, they are examined for changes in one or more of three parameters: level, trend (slope), and variability. *Level* refers to the average rate of performance during a phase. Panel A of Figure 1 shows hypothetical data demonstrating a change in level. In this case, the average rate of performance during the baseline phase is lower than the average rate of performance during the intervention phase. Figure 1 also illustrates that the change in level occurred immediately following the change in phase. The change in level is evident, in part, because there is no overlap between the phases, meaning that the lowest data point from the intervention phase is still higher than the highest data point from the baseline phase.

On the other hand, there is overlap between the baseline and intervention phases in Panel B of Figure 1, and the overall level of the dependent variable does not differ much between the

phases. There is, however, a change in trend, as there is a consistent decreasing trend during the baseline phase, which is reversed in the intervention phase.

Finally, in Panel C, there is no evidence for changes in level or trend. There is, however, a change in variability. During the baseline phase, performance in the dependent measure is highly variable, with a minimum of 0% and a maximum of 100%. In contrast, during the intervention phase, performance is stable, with a range of only 6%. All three of these types of changes may be used as evidence for the effects of an independent variable in an appropriate experimental design.

When such changes are large and immediate, visual inspection is relatively straightforward, as in all three graphs in Figure 1. In many real-life data sets, however, effects are more ambiguous. Take, for example, the graphs in Figure 2. If only the average performance during each phase is considered, each of these graphs includes a between-phase change in level. On closer inspection, however, each presents a problem that threatens the internal validity of the experiment and the ability of the clinical researcher to make a warranted causal inference about the relation between treatment (the independent variable) and effect (the dependent variable).

In Panel A of Figure 2, no change is observed until the third session of the intervention phase. This latency brings into question the assumption that the manipulation of the independent variable is responsible for the observed changes in the dependent variable. It is possible that the observed change may be more appropriately attributed to some factor outside the control of the experimenter. To rule out the plausibility of an extraneous variable, the experimental effect must be replicated, thereby showing that although there may be a delay, changes in the dependent variable reliably occur following changes to the independent variable. This type of replication (within study) is a primary characteristic of SSEDs and is the primary basis for internally valid inferences.

By contrast, Panel B of Figure 2 shows a data set in which an increasing trend is present during the baseline phase. As a result, any increases observed during the intervention phase may simply be a continuation of that trend rather than the result of the manipulation of the independent variable. This underscores the importance of “good” baseline data, and, in particular, of the need to continue collecting baseline data to eliminate the possibility that any trends observed are likely to continue in the absence of an intervention.

Panel C also underscores the importance of “good” baseline data. Although no consistent trend is present in the baseline phase, the data are highly variable. As a result, there is an overlap between many of the sessions in the baseline and intervention phases, even though the average level of performance is higher in the intervention phase (M = 37%) than in the baseline phase (M = 57%). Because the determination of experimental effects in SSEDs is based on visual inspection of the results rather than statistical analyses, such an overlap obscures any potential effects. As a result, when baseline data such as these are collected, the researcher should attempt to eliminate possible sources of variability to help establish a clear pattern of responding.

Threats to the internal validity of SSEDs, such as those demonstrated in Figure 2, are described as "demonstrations of noneffect" in the language of a panel assembled by the What Works Clearinghouse (WWCH), an initiative of the Institute for Education Sciences (IES) that was appointed to develop a set of criteria for determining whether the results of SSEDs provide evidence of sufficient quality to identify an intervention as evidence based (Kratochwill et al., 2010). A description of the criteria developed by the panel as well as their application to evidence-based practice in CSD follows.

Criteria for Evidence Quality in SSEDs

A number of groups from different fields have developed criteria to assess the quality of evidence used to support the effectiveness of interventions and to facilitate the incorporation of research findings into practice. Among the most recent of these criteria focusing specifically on SSEDs are those developed by the WWCH panel. Considering the WWCH criteria, determining whether an intervention qualifies as evidence based involves a three-step sequence. The first step involves assessing the adequacy of the experimental design (see Table 1) to determine whether it meets the standards, with or without reservations. If the design is not found to be adequate, no further steps are needed. If the design meets the standards, the second step is to conduct a visual analysis of the results to determine whether the data suggest an experimental effect. If the visual analysis supports the presence of an effect, the data should be examined for demonstrations of noneffect, such as those depicted in Figure 2. If no evidence of an experimental effect is found, the process is terminated. If the visual analysis suggests that the results support the effectiveness of the intervention, the reviewer can move on to the third step: assessing the overall state of the evidence in favor of an intervention by examining the number of times its effectiveness has been demonstrated, both within and across participants. The importance of replication in SSEDs is discussed in more detail in the next section. If the design meets the standards and the visual analysis indicates that there is an effect, with no demonstrations of noneffect, the study would be considered one that provides strong evidence. If it meets the standards and there is evidence of an effect, but the results include at least one demonstration of noneffect, then the study would be considered one that provides moderate evidence. The results of all studies that reported the effects of a particular intervention can then be examined for overall level of evidence in favor of the treatment.

Replication for Internal and External Validity

Replication is one of the hallmarks of SSEDs. Experimental control is demonstrated when the effects of the intervention are repeatedly and reliably demonstrated within a single participant or across a small number of participants. The way in which the effects are replicated depends on the specific experimental design implemented. For many designs, each time the intervention is implemented (or withdrawn following an initial intervention phase), an opportunity to provide an instance of effect replication is created. This within-study replication is the basis of internal validity for SSEDs.

By replicating an investigation across different participants, or different types of participants, researchers and clinicians can examine the generality of the treatment effects and thus potentially enhance external validity. Kazdin (2010) distinguished between two

types of replication. *Direct replication* refers to the application of an intervention to new participants under exactly, or nearly exactly, the same conditions as those included in the original study. This type of replication allows the researcher or clinician to determine whether the findings of the initial study were specific to the participant(s) who were involved. *Systematic replication* involves the repetition of the investigation while systematically varying one or more aspects of the original study. This might include applying the intervention to participants with more heterogeneous characteristics, conducting the intervention in a different setting with different dependent variables, and so forth. The variation inherent to systematic replication allows the researcher, educator, or clinician to determine the extent to which the findings will generalize across different types of participants, settings, or target behaviors. As noted by Johnston and Pennypacker (2009), conducting direct replications of an effect tells us about the certainty of our knowledge, whereas conducting systematic replications can expand the extent of our knowledge.

An intervention or treatment cannot be considered evidence based following the results of a single study. The WWCH panel recommended that an intervention have a minimum of five supporting SSED studies meeting the evidence standards if the studies are to be combined into a single summary rating of the intervention's effectiveness. Further, these studies must have been conducted by at least three different research teams at three different geographical locations and must have included a combined number of at least 20 participants or cases (see O'Neill, McDonnell, Billingsley, & Jenson, 2011, for a summary of different evidence-based practice guidelines on replication). The panel also suggested the use of some type of effect size to quantify intervention effects within each study, thereby facilitating the computation of a single summary rating of the evidence in favor of the invention (a discussion of the advantages and disadvantages of SSEDs and effects sizes follows later). In the next section, the specific types of SSEDs are described and reviewed.

Types of SSEDs

Six primary design types are discussed: the pre-experimental (or AB) design, the withdrawal (or ABA/ABAB) design, the multiple-baseline/multiple-probe design, the changing-criterion design, the multiple-treatment design, and the alternating treatments and adapted alternating treatments designs (see Table 2).

Pre-Experimental (AB) Design

Although the AB design is often described as a SSED, it is more accurately considered a pre-experimental design because it does not sufficiently control for many threats to internal validity and, therefore, does not demonstrate experimental control. As a result, the AB design is best thought of as one that demonstrates correlation between the independent and dependent variables but not necessarily causation. Nevertheless, the AB design is an important building block for true experimental designs. It is made up of two phases: the A (baseline) phase and the B (intervention) phase. Several baseline sessions establish the pre-intervention level of performance. As previously noted, the purpose of the baseline phase is to establish the existing levels/patterns of the behavior(s) of interest, thus allowing for future performance predictions under the continued absence of intervention. Due to the lack of replication of the experimental effect in an AB design, however, it is impossible to say with

certainty whether any observed changes in the dependent variable are a reliable, replicable result of the manipulation of the independent variable. As a result, it is possible that any number of external factors may be responsible for the observed changes. Nevertheless, these designs can provide preliminary objective data regarding the effects of an intervention when time and resources are limited (see Kazdin, 2010).

Withdrawal (ABA and ABAB) Designs

The withdrawal design is one option for answering research questions regarding the effects of a single intervention or independent variable. Like the AB design, the ABA design begins with a baseline phase (A), followed by an intervention phase (B). However, the ABA design provides an additional opportunity to demonstrate the effects of the manipulation of the independent variable by withdrawing the intervention during a second "A" phase. A further extension of this design is the ABAB design, in which the intervention is re-implemented in a second "B" phase. ABAB designs have the benefit of an additional demonstration of experimental control with the reimplementation of the intervention. Additionally, many clinicians/educators prefer the ABAB design because the investigation ends with a treatment phase rather than the absence of an intervention.

It is worth noting that although they are often used interchangeably in the literature, the terms *withdrawal design* and *reversal design* refer to two related but distinctly different research designs. In the withdrawal design, the third phase represents a change back to pre-intervention conditions or the withdrawal of the intervention. In contrast, the reversal design requires the active reversal of the intervention conditions. For example, reinforcement is provided contingent on the occurrence of a response incompatible with the response reinforced during the intervention (B) phases (see Barlow, Nock, & Hersen, 2009, for a complete discussion of the mechanics and relative advantages of reversal designs).

A recent example of the withdrawal design was executed by Tincani, Crozier, and Alazetta (2006). They implemented an ABAB design to demonstrate the effects of positive reinforcement for vocalizations within a Picture Exchange Communication System (PECS) intervention with school-age children with autism (see Figure 3). A visual analysis of the results reveals large, immediate changes in percentage of vocal approximations emitted by the student each time the independent variable is manipulated, and there are no overlapping data between the baseline and intervention phases. Finally, there are no demonstrations of a noneffect. As a result, this case would be considered strong evidence supporting the effectiveness of the intervention based on the WWCH evidence-based practice criteria. The study meets the standards (with reservations) because (a) the researchers actively manipulated the independent variable (presence/absence of vocal reinforcement), (b) data on the dependent variable were collected systematically over time, (c) a minimum of four data points were collected in each phase (at least five are needed to meet the standards without reservations), and (d) the effect was replicated three times (the intervention was implemented, withdrawn, and implemented again).

Advantages and disadvantages of withdrawal designs—Withdrawal designs (e.g., ABA and ABAB) provide a high degree of experimental control while being relatively

straightforward to plan and implement. However, a major assumption of ABAB designs is that the dependent variable being targeted is *reversible* (e.g., will return to pre-intervention levels when the intervention is withdrawn). If the individual continues to perform the behavior at the same level even though the intervention is withdrawn, a functional relationship between the independent and dependent variables cannot be demonstrated. When this happens, the study becomes susceptible to the same threats to internal validity that are inherent in the AB design.

Although many behaviors would be expected to return to pre-intervention levels when the conditions change, others would not. For example, if one's objective were to teach or establish a new behavior that an individual could not previously perform, returning to baseline conditions would not likely cause the individual to “unlearn” the behavior. Similarly, studies aiming to improve proficiency in a skill through practice may not experience returns to baseline levels when the intervention is withdrawn. In other cases, the behavior of the parents, teachers, or staff implementing the intervention may not revert to baseline levels with adequate fidelity. In other cases still, the behavior may come to be maintained by other contingencies not under the control of the experimenter.

Another potential disadvantage of these designs is the ethical issue associated with withdrawing an apparently effective intervention. Additionally, stakeholders may be unwilling (or unable) to return to baseline conditions, especially given the expectation that the behavior will return to baseline levels (or worse) when the intervention is withdrawn.

Overall, ABAB designs are one of the most straightforward and strongest SSED “treatment effect demonstration” strategies. Ethical considerations regarding the withdrawal of the intervention and the reversibility of the behavior need to be taken into account before the study begins. Further extensions of the ABAB design logic to comparisons between two or more interventions are discussed later in this article.

Multiple-Baseline and Multiple-Probe Designs

Multiple-baseline and multiple-probe designs are appropriate for answering research questions regarding the effects of a single intervention or independent variable across three or more individuals, behaviors, stimuli, or settings. On the surface, multiple-baseline designs appear to be a series of AB designs stacked on top of one another. However, by introducing the intervention phases in a staggered fashion, the effects can be replicated in a way that demonstrates experimental control. In a multiple-baseline study, the researcher selects multiple (typically three to four) conditions in which the intervention can be implemented. These conditions may be different behaviors, people, stimuli, or settings. Each condition is plotted in its own panel, or *leg*, that resembles an AB graph. Baseline data collection begins simultaneously across all the legs. The intervention is introduced systematically in one condition while baseline data collection continues in the others. Once responding is stable in the intervention phase in the first leg, the intervention is introduced in the next leg, and this continues until the AB sequence is complete in all the legs.

Figure 4 shows the results from a study using a multiple-baseline, across-participants design examining the collateral language effects of a question-asking training procedure for

children with autism (Koegel, Koegel, Green-Hopkins, & Barnes, 2010). The design meets the WWCH standards. The independent variable (the question-asking procedure) was actively manipulated, and the dependent variable (percentage of unprompted questions asked by each child) was measured systematically across time, with appropriate levels of interobserver agreement reported. Except for the generalization phase, at least five data points were collected in each phase. Because the generalization phase is not integral to the demonstration of the experimental control, this does not affect the sufficiency of the design: The effects were replicated across three activities.

Visual analysis of the results supports the effectiveness of the intervention in that there was an immediate change in unprompted question-asking with the implementation of the intervention for all three children, with no overlap between the baseline and intervention phases. No indications of noneffect are present in the data. As a result, this study provides strong evidence that the question-asking intervention results in increases in collateral question-asking.

The data from the final phase of the study depicted in Figure 4 are worth noting because they show the continued performance of the dependent variable in the absence of the treatment. In some ways, this is akin to a return to baseline conditions, as in the second “A” condition of a withdrawal design. In this case, however, the behavior does not return to pre-intervention levels, suggesting that the behavior is nonreversible and that using a reversal design to demonstrate the effects of the intervention would have been inappropriate. For this study, the maintenance of the behavior after the intervention was withdrawn supports its long-term effectiveness without undermining the experimental control.

In some cases, the simultaneous and continuous data collection in all legs of multiple-baseline designs is not feasible or necessary. Multiple-probe designs are a common variation on multiple baselines in which continuous baseline assessment is replaced by intermittent probes to document performance in each of the conditions during baseline. Probes reduce the burden of data collection because they remove the need for continuous collection in all phases simultaneously (see Horner & Baer, 1978, for a full description of multiple-probe designs). Pre-intervention probes in Condition 1 are obtained continuously until a stable pattern of performance is established. Meanwhile, single data collection sessions would be conducted in each of the other conditions to assess pre-intervention levels. Once responding has reached the criterion threshold in the intervention phase of the first leg, continuous measurement of pre-intervention levels is introduced in the second. When stable responding during the intervention phase is observed, intermittent probes can be implemented to demonstrate continued performance, and intervention is introduced in the second leg. This pattern is repeated until the effects of the intervention have been demonstrated across all the conditions.

Multiple-probe designs may not be appropriate for behaviors with significant variability because the intermittent probes may not provide sufficient data to demonstrate a functional relationship. If a stable pattern of responding is not clear during the baseline phase with probes, the continuous assessment of a multiple-baseline format may be necessary.

When selecting conditions for a multiple-baseline (or multiple-probe) design, it is important to consider both the independence and equivalence of the conditions. Independence means that changing behavior in one condition will not affect performance in the others. If the conditions are not independent, implementing the intervention in one condition may lead to changes in behavior in another condition while it remains in the baseline phase (McReynolds & Kearns, 1983). This makes it challenging (if not impossible) to demonstrate convincingly that the intervention is responsible for changes in the behavior across all the conditions. When implementing the intervention across individuals, it may be necessary—to avoid diffusion of the treatment—to ensure that the participants do not interact with one another. When the intervention is implemented across behaviors, the behaviors must be carefully selected to ensure that any learning that takes place in one will not transfer to the next. Similarly, contexts or stimuli must be sufficiently dissimilar so as to minimize the likelihood of effect generalization.

Although an assumption of independence suggests that researchers should select conditions that are clearly dissimilar from one another, the conditions must be similar enough that the effects of the independent variable can be replicated across each of them. If the multiple baselines are conducted across participants, this means that all the participants must be comparable in their behaviors and other characteristics. If the multiple baselines are being conducted across behaviors, those behaviors must be similar in function, topography, and the effort required to produce them while remaining independent of one another.

Advantages and disadvantages of multiple-baseline/multiple-probe designs— Because replication of the experimental effect is across conditions in multiple-baseline/multiple-probe designs, they do not require the withdrawal of the intervention. This can make them more practical with behaviors for which a return to baseline levels cannot occur. Depending on the speed of the changes in the previous conditions, however, one or more conditions may remain in the baseline phase for a relatively long time. Thus, when multiple baselines are conducted across participants, one or more individuals may wait some time before receiving a potentially beneficial intervention.

The need for multiple conditions can make multiple-baseline/multiple-probe designs inappropriate when the intervention can be applied to only one individual, behavior, and setting. Also, potential generalization effects such as these must be considered and carefully controlled to minimize threats to internal validity when these designs are used. Nevertheless, multiple-baseline designs often are appealing to researchers and interventionists because they do not require the behavior to be reversible and do not require the withdrawal of an effective intervention.

Changing-Criterion Designs

Similar to withdrawal and multiple-baseline/multiple-probe designs, changing-criterion designs are appropriate for answering questions regarding the effects of a single intervention or independent variable on one or more dependent variables. In the previous designs, however, the assumption is that manipulating the independent variable will result in large, immediate changes to the dependent variable(s). In contrast, a major assumption of the changing-criterion is that the dependent variable can be increased or decreased

incrementally with stepwise changes to the dependent variable. Typically, this is achieved by arranging a consequence (e.g., reinforcement) contingent on the participant meeting the predefined criterion. The changing-criterion design can be considered a special variation of multiple-baseline designs in that each phase serves as a baseline for the subsequent one (Hartmann & Hall, 1976). However, rather than having multiple baselines across participants, settings, or behaviors, the changing-criterion design uses multiple levels of the independent variable. Experimental control is demonstrated when the behavior changes repeatedly to meet the new criterion (i.e., level of the independent variable).

Figure 5 shows the results of a study by Facon, Sahiri, and Riviere (2008). In this study, a token reinforcement procedure was used to increase the speech volume of a child with selective mutism and mental retardation. During the baseline phase, the child's speech was barely audible, averaging 43 dB. For each new phase in the treatment condition, a criterion level for speech volume was set, which dictated what level of performance the child had to demonstrate to earn the reinforcement tokens. The horizontal lines on the graph represent the criterion set for each phase. To ensure the student's success during the intervention, the initial criterion was set at 43 dB. Researchers established a priori decision rules for changes to the criterion: The criterion would be increased when 80% of the child's utterances during three consecutive sessions were equal to or above the current criterion. Each new criterion value was equal to the mean loudness of the five best verbal responses during the last session of the previous phase.

The design of this study meets the WWCH standards, but with reservations. The independent variable (in this case, the token reinforcement system with the increasing dB criterion) was actively manipulated by the researchers, and the dependent variable was measured systematically over time. Each phase included a minimum of three data points (but not the five points required to meet the standards fully), and the number of phases with different criteria far exceeded the minimum three required.

Upon visual inspection, the results support the effectiveness of the intervention. There were few overlapping data points between the different criterion phases, and changes to the criterion usually resulted in immediate increases in the target behavior. These results would have been further strengthened by the inclusion of bidirectional changes, or mini-reversals, to the criterion (Kazdin, 2010). Such temporary changes in the level of the dependent measure(s) in the direction opposite from that of the treatment effect enhance experimental control because they demonstrate that the dependent variable covaries with the independent variable. As such, bidirectional changes are much less likely to be the result of extraneous factors. Nevertheless, the results did not show any evidence of noneffect, and the results would be considered strong evidence in favor of the intervention.

Advantages and disadvantages of changing-criterion designs—Changing-criterion designs are ideal for behaviors for which it is unrealistic to expect large, immediate changes to coincide with manipulation of the independent variable. They do not require the withdrawal of treatment and, therefore, do not present any ethical concerns associated with removing potentially beneficial treatments. Unlike multiple-baseline/multiple-probe designs, changing-criterion studies require only one participant, behavior, and setting. Not all

interventions, however, can be studied using a changing-criterion design; only interventions in which consequences for meeting or not meeting the established criterion levels of the behavior can be used. In addition, because the participant must be able to meet a certain criterion to contact the contingency, the participant must have some level of the target behavior in his or her repertoire before the study begins. Changing-criterion designs are not appropriate for behaviors that are severe or life threatening because they do not result in immediate, substantial changes. For teaching many complex tasks, however, shaping a behavior through a series of graduated steps is an appropriate strategy, and the changing-criterion design is a good option for a demonstrating the intervention's effectiveness.

Multiple-Treatment Designs

Thus far, the designs that we have described are only appropriate to answer questions regarding the effects of a single intervention or variable. In many cases, however, investigators—whether they are researchers, educators, or clinicians—are interested in not only whether an intervention works but also whether it works better than an alternative intervention. One strategy for comparing the effects of two interventions is to simply extend the logic of withdrawal designs to include more phases and more conditions. The most straightforward design of this type is the ABACAC design, which begins with an ABA design and is followed by a CAC design. The second “A” phase acts as both the withdrawal condition for the ABA portion of the experiment and the baseline phase for the ACAC portion. This design is ideal in situations where an ABA or ABAB study was planned but the effects of the intervention were not as sizable as had been hoped. Under these conditions, the intervention can be modified, or another intervention selected, and the effects of the new intervention can be demonstrated. The design has the same advantages and disadvantages of basic withdrawal designs but allows for a comparison of effects for two different treatments. A major drawback, however, is that the logic of SSEDs allows only for the comparison of adjacent conditions. This restriction helps to minimize threats to internal validity, such as maturation, that can lead to gradual changes in behavior over time, independent of study conditions. As a result, it is not appropriate to comment on the relative effects of the interventions (i.e., the “B” and “C” phases) in an ABACAC study because they never occur next to one another. Rather, one can only conclude that one, both, or neither intervention is effective relative to baseline. On the other hand, beginning with a full reversal or withdrawal design (ABAB), with it followed by a demonstration of the effects of the second intervention (CAC, resulting in ABABCAC), allows for the direct comparison of the two interventions. The BC comparison, however, is never repeated in this sequence, limiting the internal validity of the comparison.

Besides comparing the relative effects of two or more distinct interventions, multiple-treatment-phase designs can be used to assess the additive effects of treatment components. For example, if a treatment package consists of two separate components (components “B” and “C”), one can determine whether the intervention effects are due to one component alone or whether both are needed. Ward-Horner and Sturmey (2010) identified two methods for conducting component analyses: *dropout*, in which components were systematically removed from the treatment package to determine whether the treatment retained its effectiveness, and *add-in*, in which components were assessed individually before the

implementation of the full treatment package. Each of these methods has its own advantages and disadvantages (see Ward-Horner & Sturmey, 2010, for a full discussion), but taken together, component analyses can provide a great deal of information about the necessity and sufficiency of treatment components. In addition, they can inform strategies for fading out treatments while maintaining their effects.

Wacker and colleagues (1990) conducted dropout-type component analyses of functional communication training (FCT) procedures for three individuals with challenging behavior. The data presented in Figure 6 show the percentage of intervals with hand biting, prompts, and mands (signing) across functional analysis, treatment package, and component analysis phases. The functional analysis results indicated that the target behavior (hand biting) was maintained by access to tangibles as well as by escape from demands. In the second phase, a treatment package that included FCT and time-out was implemented. By the end of the phase, the target behavior was eliminated, prompting had decreased, and signing had increased. To identify the active components of the treatment package, a dropout component analysis was conducted. First, the time-out component of the intervention was removed, leaving the FCT component alone. A decreasing trend in signing and an increasing trend in hand biting were observed. This was reversed when the full treatment packaged was reimplemented. In the third phase of the component analysis, the FCTcomponent was removed, leaving time-out and differential reinforcement of other behavior (DRO). Again, a decreasing trend in signing and an increasing trend in hand biting were observed, which were again reversed when the full treatment package was applied.

Overall, visual inspection of these data provides a strong argument for the necessity of both the FCT and time-out components in the effectiveness of the treatment package, and no indications of noneffect are present in the data. The design, however, does not meet the standards set forth by the WWCH panel. This is because (a) the final two final treatment phases do not include the minimum of three data points and (b) the individual treatment component phases (FCT only and time-out/DRO) were implemented only once each. As a result, the data from this study could not be used to support the treatment package as an evidence-based practice by the IES standards. Additional data points within each phase, as well as replications of the phases, would strengthen the study results.

One disadvantage of all designs that involve two or more interventions or independent variables is the potential for multiple-treatment interference. This occurs when the same participant receives two or more treatments whose effects may not be independent. As a result, it is possible that the order in which the interventions are given will affect the results. For example, the effects of two interventions may be additive, so that the effects of Intervention 2 are enhanced beyond what they should be because Intervention 2 followed Intervention 1. In essence, this creates the potential for an order effect (or a carryover effect). Alternatively, Intervention 1 may have measurable but delayed effects on the dependent variable, making it appear that Intervention 2 is effective when the results should be attributed to Intervention 1. Such possibilities should be considered when multi-treatment studies are being planned (see Hains & Baer, 1989, for a comprehensive discussion of multiple-treatment interference). A final, longer phase in which the final “winning”

treatment is implemented for an extended time can help alleviate some of the concerns regarding multiple-treatment interference.

Advantages and disadvantages of multiple-treatment designs—Designs such as ABCABC and ABCBCA can be very useful when a researcher wants to examine the effects of two interventions. These designs provide strong internal validity evidence regarding the effectiveness of the interventions. External validity, however, may be compromised by the threat of multiple-treatment interference. Additionally, the same advantages and disadvantages of ABAB designs apply, including issues related to the reversibility of the target behavior. Despite their limitations, these designs can provide strong empirical data upon which to base decisions regarding the selection of treatments for an individual client. Although, in theory, these types of designs can be extended to compare any number of interventions or conditions, doing so beyond two becomes excessively cumbersome; therefore, the alternating treatments design should be considered.

Alternating Treatments and Adapted Alternating Treatments Designs

Alternating treatments design (ATD)—The logic of the ATD is similar to that of multiple-treatment designs, and the types of research questions that it can address are also comparable. The major distinction is that the ATD involves the rapid alternation of two or more interventions or conditions (Barlow & Hayes, 1979). Data collection typically begins with a baseline (A) phase, similar to that of a multiple-treatment study, but during the next phase, each session is randomly assigned to one of two or more intervention conditions. Because there are no longer distinct phases of each intervention, the interpretation of the results of ATD studies differs from that of the studies reviewed so far. Rather than comparing between phases, all the data points within a condition (e.g., all sessions of Intervention 1) are connected (even if they do not occur adjacently). Demonstration of experimental control is achieved by having differentiation between conditions, meaning that the data paths of the conditions do not overlap.

In ATDs, it is important that all potential “nuisance” variables be controlled or counterbalanced. For example, having different experimenters conduct sessions in different conditions, or running different session conditions at different times of day, may influence the results beyond the effect of the independent variables specified. Therefore, all experimental procedures must be analyzed to ensure that all conditions are identical except for the variable(s) of interest. Presenting conditions in random order can help eliminate issues regarding temporal cycles of behavior as well as ensure that there are equal numbers of sessions for each condition.

Lang and colleagues (2011) used an ATD to examine the effects of language of instruction on correct responding and inappropriate behavior (tongue clicks) with a student with autism from a Spanish-speaking family. To ensure that the conditions were equivalent, all aspects of the teaching sessions except for the independent variable (language of instruction) were held constant. Specifically, the same teacher, materials, task demands, reinforcers, and reinforcer schedules were used in both the English and Spanish sessions.

The results of this study (see Figure 7) demonstrated that the student produced a higher number of correct responses and engaged in fewer challenging behaviors when instruction was delivered in Spanish than in English. The superiority of the Spanish instruction was evident in this case because there was no overlap in correct responding or inappropriate behaviors between the English and Spanish conditions.

Although visual analysis supported the inference that treatment effects were functionally related to the independent variable, the results of this study did not meet the design standards set out by the WWCH panel because the design consisted of only two treatments in comparison with each other. To meet the criterion of having at least three attempts to demonstrate an effect, studies using an ATD must include a direct comparison of three interventions, or two interventions compared with a baseline. To be considered as support for an evidence-based practice, this design would need to have incorporated a third intervention condition or to have begun with a baseline condition.

Adapted alternating treatments design (AATD)—One commonly used alternative to the ATD is called the *adapted alternating treatments design* (AATD; Sindelar, Rosenburg, & Wilson, 1985). Whereas the traditional ATD assesses the effects of different interventions or independent variables on a single outcome variable, in the AATD, a different set of responses is assigned to each intervention or independent variable. The resulting design is similar to a multiple-baseline, across-behaviors design with concurrent training for all behaviors. For example, Conaghan, Singh, Moe, Landrum, and Ellis (1992) assigned a different set of 10 phrases to each of three conditions (directed rehearsal, directed rehearsal plus positive reinforcement, and control). This strategy allowed the researchers to determine whether the acquisition of new signed phrases differed across the three conditions. Figure 8 shows one participant's correct responses during sessions across baseline phases, alternating treatments phases, and extended treatment phases.

Unlike the Lang et al. (2011) study, the design used in this study met the WWCH standards. This was because, in addition to meeting the minimum number of sessions per phase, it included a direct comparison between three conditions as well as a direct comparison with a baseline phase. The data from the baseline phase established that the participant did not respond correctly in the absence of the intervention. The data from the alternating treatments phase supported the effectiveness of the directed rehearsal and directed rehearsal plus positive reinforcement conditions compared with the control condition. They also supported the relative effectiveness of the directed rehearsal with reinforcement compared with directed rehearsal alone.

During the initial four sessions of the alternating treatments phase, responding remained at zero for all three word sets. Steadily increasing trends were observed in both of the directed rehearsal conditions beginning in the fifth session, whereas responding remained at zero in the control condition. The rate of acquisition in the directed rehearsal plus positive reinforcement condition was higher than in directed rehearsal alone throughout the alternating treatments phase. The latency in correct responding observed during the initial sessions of the alternating treatments was a demonstration of noneffect. The fact that no change in responding was observed in the control condition, however, is evidence that the

changes were due to the intervention rather than a result of some factor outside of the study. As further demonstration of the experimental effect of directed rehearsal plus reinforcement, a final condition was implemented in which the treatment package was used to teach the phrases from the other two conditions. This condition further strengthened the evidence for the effectiveness of the intervention, as performance on all three words sets reached 100% by the end of the phase. In sum, the latency to change observed during the alternating treatments phase meant that this study merits a rating of moderate evidence in favor of the intervention.

Advantages and disadvantages of ATDs and AATDs—ATDs and AATDs can be useful in comparing the effects of two or more interventions or independent variables. Unlike multiple-treatment designs, these designs can allow multiple comparisons in relatively few sessions. The issues related to multiple-treatment interference are also relevant with the ATD because the dependent variable is exposed to each of the independent variables, thus making it impossible to disentangle their independent effects. To ensure that the selected treatment remains effective when implemented alone, a final phase demonstrating the effects of the best treatment is recommended (Holcombe & Wolery, 1994), as was done in the study by Conaghan et al., 1992. Many researchers pair an independent but salient stimulus with each treatment (i.e., room, color of clothing, etc.) to ensure that the participants are able to discriminate which intervention is in effect during each session (McGonigle, Rojahn, Dixon, & Strain, 1987). Nevertheless, outcome behaviors must be readily reversible if differentiation between conditions is to be demonstrated.

The AATD eliminates some of the concerns regarding multiple-treatment interference because different behaviors are exposed to different conditions. As in the multiple-baseline/multiple-probe designs, the possibility of generalization across behaviors must be considered, and steps should be taken to ensure the independence of the behaviors selected. In addition, care must be taken to ensure equal difficulty of the responses assigned to different conditions.

Having reviewed the logic underlying SSED, the basic approach to analysis (visual inspection relying on observed changes in level, trend, and variability), and the core strategies for arranging conditions (i.e., design types), in the following section we briefly discuss a number of quantitative evaluation issues concerning SSED. The issues are germane because of the WWCH and related efforts to establish standard approaches for evaluating SSED data sets as well as the problem of whether and how to derive standardized effect sizes from SSED data sets for inclusion in quantitative syntheses (i.e., meta-analysis).

Evaluating Results in SSED Research

Statistical Analysis and SSED

The issue of when, if ever, the data generated from SSEDs should be statistically analyzed has a long and, at times, contentious history (Iwata, Neef, Wacker, Mace, & Vollmer, 2000). We approach this issue by breaking it into four related but distinct parts that include detecting effects, determining their magnitude and the quality of the causal inference, and data-based decision making. Subsequently, relevant considerations for research and practice

are delineated. Space considerations preclude treating any one aspect of this issue exhaustively (suggestions for further reading are provided).

Effect detection—Conventional approaches to single-subject data analysis rely on visual inspection (as reviewed earlier in this article). From the perspective of clinical significance, supporting a "visual inspection–only" approach is warranted because the practitioner (and, ultimately, the field of practice) is interested in identifying only those variables that lead to large, unambiguous changes in behavior. One argument against the exclusive reliance on visual inspection is that it is prone to Type 1 errors (inferring an effect when there is none), particularly if the effects are small to medium (Franklin, Gorman, Beasley, & Allison, 1996; Todman & Dugard, 2001). Evidence for experimental control is not always as compelling from a visual analysis perspective. This was showcased in the Tincani et al. (2006) study discussed previously. In many cases, the clinical significance of behavior change between conditions is less clear and, therefore, is open to interpretation.

From the perspective of scientific significance, one can argue that statistical analysis may be warranted as a judgment aid for determining whether there were any effects, regardless of size, because knowing this would help determine whether to continue investigating the variable (i.e., intervention). If it is decided that, under some circumstances, it is scientifically sensible to use statistical analyses (e.g., *t* tests, analyses of variance [ANOVAs], etc.) as judgment aids for effect detection within single case data sets, the next question is a very practical one—can we? In other words, can parametric inferential statistical techniques be applied safely? In this context, the term *safely* refers to whether the outcome variables are sufficiently robust that they withstand violating the assumptions underlying the statistical test. The short answer seems to be "no," with the qualifier "under almost all circumstances." The key limitation and common criticism of generating statistics based on single-subject data is auto-correlation (any given data point is dependent or interacts with the data point preceding it). Because each data point is generated by the same person, the data points are not independent of one another (violating a core assumption of statistical analysis—technically, that the error terms are not independent of one another). Thus, performance represented in each data point may likely be influencing the next (Todman & Dugard, 2001). Autocorrelated data will, in turn, artificially inflate *p* values and affect Type 1 error rates.

One argument for statistically analyzing single-subject data sets, mentioned above, is that visual inspection is prone to Type 1 error in the presence of medium to small effects (Franklin et al., 1996). Unfortunately, the proposed solution of implementing conventional inferential statistical tests with single-subject data based on repeated measurement of the same subject is equally prone to Type 1 error because of autocorrelation. Traditional nonparametric approaches have been advocated, but they do not necessarily avoid the autocorrelation problem and, depending on the size of the data array, there are power issues. Alternatively, if single-subject data are regarded as time-series data, there have been some novel applications of bootstrapping methodologies relying on using the data set itself along with resampling approaches to determine exact probabilities rather than probability estimates (Wilcox, 2001). For example, Borckardt et al. (2008) described a "simulation modeling analysis" for time-series data, which allows a statistical comparison between phases of a single-subject experiment while controlling for serial dependency in the data

(i.e., quantifying the autocorrelation and modeling it in the analysis). In the end, effect detection is determined by data patterns in relation to the phases of the experimental design. It seems that the clearer one is about the logic of the design and the criteria that will be used to determine an effect in advance, the less one needs to rely on searching for a "just-in-case" test after the fact.

Magnitude of effect—An emphasis on accountability is embodied in the term *evidence-based practice*. One of the tools used to help answer the question of "what works" that forms the basis for the evidence in evidence-based practice is *meta-analysis*—the quantitative synthesis of studies from which standardized and weighted effect sizes can be derived. Meta-analysis methodology provides an objective estimate of the magnitude of an intervention's effect. One of the main problems of SSEDs is that the evidence generated is not always included in meta-analyses. Alternatively, if studies based on SSEDs are used in meta-analysis, there is no agreement on the correct metric to estimate and quantify the effect size.

An obvious corollary to the issue of effect magnitude is that visual inspection, per se—although sensitive to a range of holistic information embodied in a data display (trend, recurrent pattern, delayed/lagged response, variability, etc.; Parker & Hagan-Burke, 2007)—does not generate a quantitative index of intervention strength (i.e., effect magnitude) that is recognizable to the broader scientific community. The determination of which practices and interventions are evidence based (and which will, therefore, be promoted and funded) increasingly involves quantitative synthesis of data and exposes the need for a single, agreed-upon effect size metric to reflect magnitude in SSEDs (Parker, Hagan-Burke, & Vannest, 2007). Accordingly, the changing scientific standards across practice communities (e.g., ASHA, American Psychological Association, American Educational Research Association) are reflected in the organization's editorial policies and publication practices, which increasingly require effect sizes to be reported.

There has been a small but steady body of work addressing effect size calculation and interpretation for SSEDs. Space precludes an exhaustive review of all the metrics (for comprehensive reviews, see Parker & Hagan-Burke, 2007, and related papers from this group). There are, however, a number of points that can be made regarding the use (derivation, interpretation) of effect size indices that are common to all. The simplest and most common effect size metric is the percentage of nonoverlapping data (PND; Scruggs, Mastropieri, & Casto, 1987). It is easy to calculate by hand and, therefore, is easily accessible to practitioners. The most extreme positive (the term *positive* is used in relation to the clinical problem being addressed; therefore, it could be the highest or lowest score) baseline data point is selected, from which a straight line is drawn across the intervention phase of the graph (for simplicity's sake, assume an AB-only design). Then, the number of data points that fall above (or below) the line is tallied and divided by the total number of intervention data points. If, for example, in a study of a treatment designed to improve (i.e., increase) communication fluency, eight of 10 data points in the intervention phase are greater in value than the largest baseline data point value, the resulting PND would equal 80%.

Although the clinical/educational appeal of such a metric seems obvious (easy to calculate, it is consistent with visual inspection of graphic data), there are potential problems with the approach. For example, there are ceiling effects for PND, making comparisons across or between interventions difficult (Parker & Hagan-Burke, 2007; Parker et al., 2007), and PND is based on a single data point, making it sensitive to outliers (Riley-Tillman & Burns, 2009). In addition, there is no known sampling distribution, making it impossible to derive a confidence interval (CI). CIs are important because they help create an interpretive context for the dependability of the effect by providing upper and lower bounds for the estimate. As a result, PND is a statistic of unknown reliability.

Most work on effect sizes for SSEDs has been conducted implicitly or explicitly to address the limits of PND. Some work has conserved the approach by continuing to calculate some form of descriptive *degrees of overlap*, including percentage of data points exceeding the median (PEM; Ma, 2006), percentage of zero data points (PZD; Johnson, Reichle, & Monn, 2009), and the percentage of all nonover-lapping data (PAND; Parker et al., 2007). Olive and Smith (2005) compared a set of descriptive effect size statistics (including a regression-based effect size, PND, standard mean difference, and mean baseline reduction) to visual analysis of several data sets and found that each consistently estimated relative effect size. Other investigators have attempted to integrate degree of overlap with general linear model approaches such as linear regression. The regression-based techniques (e.g., Gorman & Allison, 1996, pp. 159–214) make use of predicted values derived from baseline data to remove the effects of trend (i.e., predicted values are subtracted from observed data). Subsequently, adjusted mean treatment scores can be used in calculating familiar effect size statistics (e.g., Cohen's *d*, Hedge's *g*). This application may be more commonly accepted among those familiar with statistical procedures associated with group design.

As with each of the issues discussed in this section, there are advantages and disadvantages to the regression and non-regression methods for determining effect size for SSEDs. Nonregression methods involve simpler hand calculations, map on to visual inspection of the data, and are less biased in the presence of small numbers of observations (Scruggs & Mastropieri, 1998). But, as recently argued by Wolery, Busick, Reichow, and Barton (2010), the overlap approaches for calculating effect sizes do not produce metrics that adequately reflect magnitude (i.e., in cases where the intervention was effective and there is no overlap between baseline and treatment, the degree of the nonoverlap of the data—the magnitude—is not indexed by current overlap-based effect sizes). Regression methods are less sensitive to outliers, control for trend in the data, and may be more sensitive to detecting treatment effects in slope and intercept (Gorman & Allison, 1996). As work in this area continues, novel effect size indices will likely emerge. Parker and Hagan-Burke (2007), for example, demonstrated that the *improvement rate difference* metric (IRD—an index frequently used in evidenced-based medicine) was superior to both PND and PEM (it produces a CI and discriminates among cases [i.e., reduced floor/ceiling effects]) but conserved many of their clinically appealing features (hand calculation, based on nonoverlapping data) without requiring any major assumptions of the data.

Although effect sizes may not be a requirement for databased decision making for a given specific case—because the decision about effect is determined primarily by the degree of

differentiation within the data set as ascertained through visual inspection and by the logical ordering of conditions (see also the *Practice and data-based decisions* section below)—their calculation and reporting may be worth considering with respect to changing publication standards and facilitating future meta-analyses. Note also that lost in the above discussion concerning effect size metrics is the issue of statistical versus clinical significance. Although one of the scientific goals of research is to discover functional relations between independent and dependent variables, the purpose of applied research is discovering the relations that lead to clinically meaningful outcomes (i.e., clinical significance; see Barlow & Hersen, 1973) or socially relevant behavior changes (i.e., social validity; see Wolf, 1978). From a practice perspective, one of the problems of statistical significance is that it can over- or underestimate clinical significance (Chassan, 1979). In principle, the notion of quantifying how large (i.e., magnitude) of an effect was obtained is in keeping with the spirit of clinical significance and social validity, but the effect size interpretation should not blindly lead to assertions of clinically significant results divorced from judgments about whether the changes were clinically or socially meaningful.

Quality of inference—One of the great scientific strengths of SSEDs is the premium placed on internal validity and the reliance on effect replication within and across participants. One of the great clinical strengths of SSEDs is the ability to use a response-guided intervention approach such that phase or condition changes (i.e., changes in the independent variable) are made based on the behavior of the participant. This notion has a long legacy and reflects Skinner's (1948) early observation that the subject (“organism”) is always right. In contrast with these two strengths, there is a line of thinking that argues for incorporating randomization into SSEDs (Kratochwill & Levin, 2009). This notion has a relatively long history (Edgington, 1975) and continues to be mentioned in contemporary texts (Todman & Dugard, 2001). The advantages and disadvantages of the practice are worth addressing (albeit briefly).

The argument for incorporating randomization into SSEDs is to further improve the quality of the causal inference (i.e., strengthening internal validity) by randomizing phase order or condition start times (there are numerous approaches to randomizing within SSEDs; see Kratochwill & Levin, 2009, or almost any of Edgington's work). However, doing so comes at the cost of practitioner flexibility in making phase/condition changes based on patterns in the data (i.e., how the participant is responding). This cost, it is argued, is worth the expense because randomization is superior to replication for reducing plausible threats to internal validity. The within-series intervention conditions are compared in an unbiased (i.e., randomized) manner rather than in a manner that is researcher determined and, hence, prone to bias. The net effect is to further enhance the scientific credibility of the findings from SSEDs. At this point, it seems fair to conclude that it remains an open question about whether randomization is superior to replication with regard to producing clinically meaningful effects for any given participant in an SSED.

One potential additional advantage to incorporating randomization into an SSED is that the data series can be analyzed using randomization tests (Bulte & Onghena, 2008; Edgington, 1996; Todman & Dugard, 2001) that leverage the ease and availability of computer-based resampling for likelihood estimation. Exact *p* values are generated, and the tests appear to be

straightforward ways to supplement the visual analysis of single-subject data. It should be noted, however, that randomization tests in and of themselves do not necessarily address the problem of autocorrelation.

Practice and data-based decisions—Finally, related to several different comments in the preceding sections regarding practical significance, there is the issue of interpreting effects directly in relation to practice in terms of eventual empirically based decision making for a given client or participant. At issue here is not determining whether there was an effect and its standardized size but whether there is change in behavior or performance over time —and the rate of that change. Riley-Tillman and Burns (2009) argued that effect size estimates may make valuable contributions for future quantitative syntheses; however, for a given practitioner, data interpretation and subsequent practice decisions are driven more by slope changes, not by average effect sizes. Nontrivial practice issues, such as special education eligibility, entitlement decisions, and instructional modification, depend on repeated measurement of student growth (i.e., time series data) that is readily translatable into single-subject design logic with judgment aids in the form of numerical slope values and aim lines.

Key advantages of relying on visual inspection and quantifying slope are not only that student growth rates can be interpreted for an individual student in relation to an intervention but also that the growth rate values can be compared to a given student's respective grade or class (or other local norms). For a clear example, interested readers are referred to Silberglitt and Gibbons' (2005) documentation of a slope-standard approach to identifying, intervening, and monitoring reading fluency and at-risk students. Of course, the approach (relying on slope values from serially collected single-subject data) is not without its problems. Depending on the frequency and duration of data collection, the standard error of the estimate for slope values can vary widely (Christ, 2006), leading to interpretive problems for practice. Thus, consistent with all of the points made above, sound methodology (design, measurement) is the biggest determinant of valid decision making. Overall, the four issues discussed above—effect detection, magnitude of effect, quality of the inference, and practice decisions—reflect the critical dimensions involved in the analysis of SSED. The importance of any one dimension over the other will likely depend on the purpose of the study and the state of the scientific knowledge about the problem being addressed.

Conclusions

Unlike the research questions often addressed by studies using traditional group designs, studies employing SSEDs can address the effects that intervention strategies and environmental variables have on performance at the individual level. SSED methodology permits flexibility within a study to modify the independent variable when it does not lead to the desired or expected effect, and it does not compromise the integrity of the experimental design. As a result, SSED methodology provides a useful alternative to RCTs (and quasi-experimental group designs) for the goal of empirically demonstrating that an intervention is effective, or alternatively, determining the better of two or more potential interventions. SSEDs are ideal for both researchers and clinicians working with small or very

heterogeneous populations in the development and implementation of evidence-based practice. The strong internal validity of well-implemented SSED studies allows for visual and, under some circumstances, statistical data analyses to support confident conclusions about—in the words of U.S. Department of Education—"what works."

Kazdin (2010), Horner et al. (2005), and others have highlighted the issue of RCTs within traditional probabilistic group design research being favored among policymakers, granting agencies, and practitioners in the position of selecting interventions from the evidence base. They also highlight the important role that SSEDs can and should play in this process. The specific criteria developed by the WWCH panel emphasize the importance of strong experimental designs—and replication, if SSEDs are to be taken seriously as a tool within the establishment of evidence-based practice. Speech, language, and hearing interventions, by their nature, strive to improve outcomes for individual clients or research participants. Evaluating those interventions within SSEDs and associated visual and statistical data analyses lends rigor to clinical work, is logically and methodologically consistent with intervention research in the field, and can serve as a common framework for decision making with colleagues within and outside the CSD field.

References

American Speech-Language-Hearing Association. Evidence-based practice in communication disorders [Position statement]. 2005. Available from www.asha.org/policy

Barlow DH, Hayes SC. The alternating treatments design: One strategy for comparing the effects of two treatments in a single subject. Journal of Applied Behavior Analysis. 1979; 12:199–210. [PubMed: 489478]

Barlow DH, Hersen M. Single case experimental designs: Uses in applied clinical research. Archives of General Psychiatry. 1973; 29:319–325. [PubMed: 4724141]

Barlow, DH.; Nock, MK.; Hersen, M. Single case experimental designs: Strategies for studying behavior change. 3rd ed.. Allyn & Bacon; Boston, MA: 2009.

Borckardt JJ, Nash MR, Murphy MD, Moore M, Shaw D, O'Neil P. Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time-series analysis. American Psychologist. 2008; 63:77–95. [PubMed: 18284277]

Bulte I, Onghena P. An R package for single-case randomization tests. Behavior Research Methods. 2008; 40:467–478. [PubMed: 18522057]

Chambless DL, Baker MJ, Baucom DH, Beutler LE, Calhoun KS, Crits-Cristoph P, Woody SR. Update on empirically validated therapies, II. The Clinical Psychologist. 1998; 51:3–16.

Chassan, JB. Research design in clinical psychology and psychiatry. 2nd ed.. Irvingston; New York, NY: 1979.

Christ T. Short-term estimates of growth using curriculum-based measurement of oral reading fluency: Estimating standard error of the slope to construct confidence intervals. School Psychology Review. 2006; 35:128–133.

Conaghan BP, Singh NN, Moe TL, Landrum TJ, Ellis CR. Acquisition and generalization of manual signs by hearing-impaired adults with mental retardation. Journal of Behavioral Education. 1992; 2:175–203.

Connell PJ, Thompson CK. Flexibility of single subject experimental designs. Part III: Using flexibility to design or modify experiments. Journal of Speech and Hearing Disorders. 1986; 51:214–225. [PubMed: 3525988]

Edgington ES. Randomization tests for one-subject operant experiments. Journal of Psychology. 1975; 90:57–68.

Edgington ES. Randomized single subject experimental designs. Behavior Research and Therapy. 1996; 34:567–574.

Facon B, Sahiri S, Riviere V. A controlled single-case treatment of severe long-term selective mutism in a child with mental retardation. Behavior Therapy. 2008; 39:313–321. [PubMed: 19027428]

Franklin, RD.; Gorman, BS.; Beasley, TM.; Allison, DB. Graphical display and visual analysis.. In: Franklin, RD.; Allison, DB.; Gorman, BS., editors. Design and analysis of single-case research. Erlbaum; Mahwah, NJ: 1996. p. 119-158.

Fukkink R. The internal validity of aphasiological single-subject studies. Aphasiology. 1996; 10:741–754.

Gorman, BS.; Allison, DB. Statistical alternatives for single-case designs.. In: Franklin, RD.; Allison, DB.; Gorman, BS., editors. Design and analysis of single-case research. Erlbaum; Mahwah, NJ: 1996. p. 159-214.

Hains AH, Baer DM. Interaction effects in multi-element designs: Inevitable, desirable, and ignorable. Journal of Applied Behavior Analysis. 1989; 22:57–69. [PubMed: 2651376]

Hanson BR. The effects of a contingent light-flash on stuttering and attention to stuttering. Journal of Communication Disorders. 1978; 11:451–458. [PubMed: 730837]

Haroldson SK, Martin RR, Starr CD. Timeout as a punishment for stuttering. Journal of Speech and Hearing Research. 1968; 14:356–364.

Hartman DP, Hall RV. The changing criterion design. Journal of Applied Behavior Analysis. 1976; 9:527–532. [PubMed: 1002635]

Holcombe A, Wolery M. Comparative single-subject research: Description of designs and discussion of problems. Topics in Early Childhood Special Education. 1994; 14:119–145.

Horner RH, Baer DM. Multiple-probe technique: A variation of the multiple baseline. Journal of Applied Behavior Analysis. 1978; 11:189–196. [PubMed: 16795582]

Horner RH, Carr EG, Halle J, McGee G, Odom S, Wolery M. The use of single-subject research to identify evidence-based practice in special education. Exceptional Children. 2005; 71:165–179.

Iwata, BA.; Neef, NA.; Wacker, DP.; Mace, FC.; Vollmer, TR., editors. Methodological and conceptual issues in applied behavior analysis. 2nd ed. Society for the Experimental Analysis of Behavior; Lawrence, KS: 2000.

Johnston, JM.; Pennypacker, HS. Strategies and tactics of behavioral research. Routledge; New York, NY: 2009.

Johnson L, Reichle J, Monn E. Longitudinal mentoring with school-based positive behavioral support teams: Influences on staff and learner behavior. Evidence-Based Communication Assessment and Intervention. 2009; 3:113–130.

Kazdin, AE. Single-case research designs: Methods for clinical and applied settings. 2nd ed.. Oxford University Press; New York, NY: 2010.

Kearns KP. Flexibility of single-subject experimental designs. Part II: Design selection and arrangement of experimental phases. Journal of Speech and Hearing Disorders. 1986; 51:204–214. [PubMed: 3525987]

Koegel LK, Koegel RL, Green-Hopkins I, Barnes CC. Question-asking and collateral language acquisition in children with autism. Journal of Autism and Developmental Disorders. 2010; 40:509–515.

Kratochwill, TR.; Hitchcock, J.; Horner, RH.; Levin, JR.; Odom, SL.; Rindskopf, DM.; Shadish, WR. Single-case design technical documentation. Version 1.0 (Pilot). 2010. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_scd.pdf

Kratochwill TR, Levin JR. Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. Psychological Methods. 2009; 15:124–144. [PubMed: 20515235]

Lang R, Rispoli M, Sigafoos J, Lancioni G, Andrews A, Ortega L. Effects of language instruction on response accuracy and challenging behavior in a child with autism. Journal of Behavioral Education. 2011; 20:252–259.

Ma HH. An alternative method for quantitative synthesis of single subject research: Percentage of data points exceeding the median. Behavior Modification. 2006; 30:598–617. [PubMed: 16894232]

Martin RR, Siegel GM. The effects of contingent shock on stuttering. Journal of Speech and Hearing Research. 1966; 9:340–352.

McGonigle JJ, Rojahn J, Dixon J, Strain PS. Multiple treatment intererence in the alternating treatments design as a function of the intercomponent interval length. Journal of Applied Behavior Analysis. 1987; 20:171–178. [PubMed: 3610896]

McReynolds, LV.; Kearns, KP. Single-subject experimental designs in communicative disorders. University Park Press; Baltimore, MD: 1983.

McReynolds LV, Thompson CK. Flexibility of single-subject experimental designs. Part I: Review of the basics of single-subject designs. Journal of Speech and Hearing Disorders. 1986; 51:194–203. [PubMed: 3525986]

Olive ML, Smith BW. Effect size calculations and single-subject designs. Educational Psychology. 2005; 25:313–324.

O'Neill, RE.; McDonnell, JJ.; Billingsley, FF.; Jenson, WR. Single case research designs in educational and community settings. Pearson; Upper Saddle River, NJ: 2011.

Parker RI, Hagan-Burke S. Median-based overlap analysis for single case data: A second study. Behavior Modification. 2007; 31:919–936. [PubMed: 17932244]

Parker RI, Hagan-Burke S, Vannest K. Percentage of all non-overlapping data (PAND): An alternative to PAND. Journal of Special Education. 2007; 40:194–204.

Reed CG, Godden AL. An experimental treatment using verbal punishment with two preschool stutterers. Journal of Fluency Disorders. 1977; 2:225–233.

Riley-Tillman, TC.; Burns, MK. Single case design for measuring response to education intervention. Guilford; New York, NY: 2009.

Robey RR, Schultz MC, Crawford AB, Sinner CA. Single-subject clinical-outcome research: Designs, data, effect sizes, and analyses. Aphasiology. 1999; 13:445–473.

Scruggs TE, Mastropieri MA. Synthesizing single subject studies: Issues and applications. Behavior Modification. 1998; 22:221–242. [PubMed: 9722473]

Scruggs TE, Mastropieri MA, Casto G. The quantitative synthesis of single-subject research design: Methodology and validation. Remedial and Special Education. 1987; 8:24–33.

Silberglitt, B.; Gibbons, KA. Establishing slope targets for use in a response to intervention model [Technical manual]. St. Croix River Education District; Rush City, MN: 2005.

Sindelar PT, Rosenburg MS, Wilson RJ. An adapted alternating treatments design for instructional research. Education and Treatment of Children. 1985; 8:67–76.

Skinner, BF. Walden two. Hackett Publishing; New York, NY: 1948.

Tincani M, Crozier S, Alazetta L. The Picture Exchange Communication System: Effects on manding and speech development for school-aged children with autism. Education and Training in Developmental Disabilities. 2006; 41:177–184.

Todman, JB.; Dugard, P. Single-case and small-n experimental designs: A practical guide to randomization tests. Erlbaum; Mahwah, NJ: 2001.

Wacker DP, Steege MW, Northup J, Sasso G, Berg W, Reimer T, Donn L. Component analysis of functional communication training across three topographies of severe behavior problems. Journal of Applied Behavior Analysis. 1990; 23:417–429. [PubMed: 2150069]

Ward-Horner J, Sturmey P. Component analyses using single-subject experimental designs: A review. Journal of Applied Behavior Analysis. 2010; 43:685–704. [PubMed: 21541152]

Wilcox, RR. Fundamentals of modern statistical methods: Substantially improving power and accuracy. Springer-Verlag; New York, NY: 2001.

Wolery M, Busick M, Reichow B, Barton EE. Comparison of overlap methods for quantitatively synthesizing single-subject data. Journal of Special Education. 2010; 44:18–28.

Wolf M. Social validity: The case for subjective measurement or how applied behavior analysis is findings its heart. Journal of Applied Behavior Analysis. 1978; 11:203–214. [PubMed: 16795590]

Yorkston KM, Spencer KA, Duffy JR, Beukelman DR, Golper LA, Miller RM, Sullivan M. Evidence-based medicine and practice guidelines: Application to the field of speech-language pathology. Journal of Medical Speech-Language Pathology. 2001; 9:243–256.

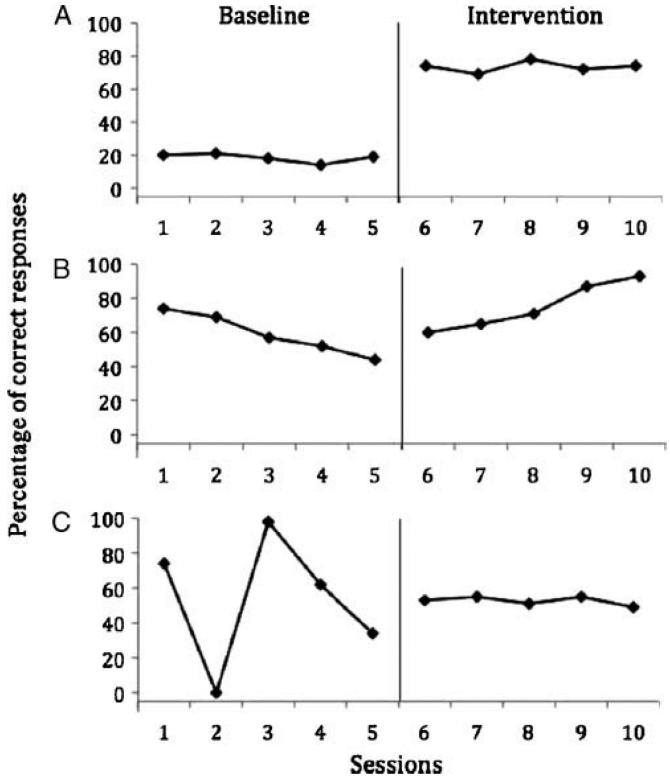

FIGURE 1.

Hypothetical data demonstrating unambiguous changes in level (Panel A), trend (Panel B), and variability (Panel C).

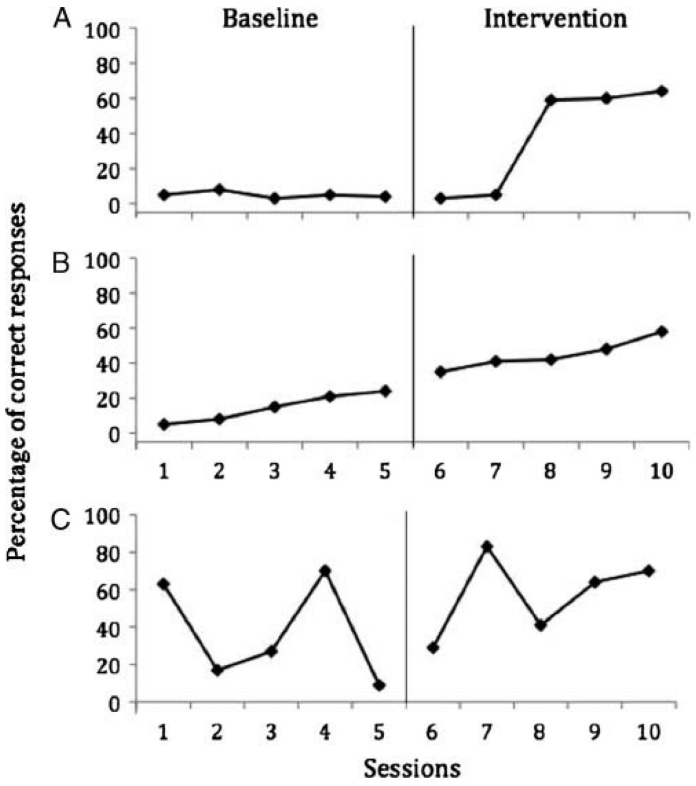

FIGURE 2.

Hypothetical data demonstrating demonstrations of non-effect: delayed latency to change (Panel A), trend in desired direction during baseline phase (Panel B), highly variable data with overlap between baseline and intervention phases (Panel C).

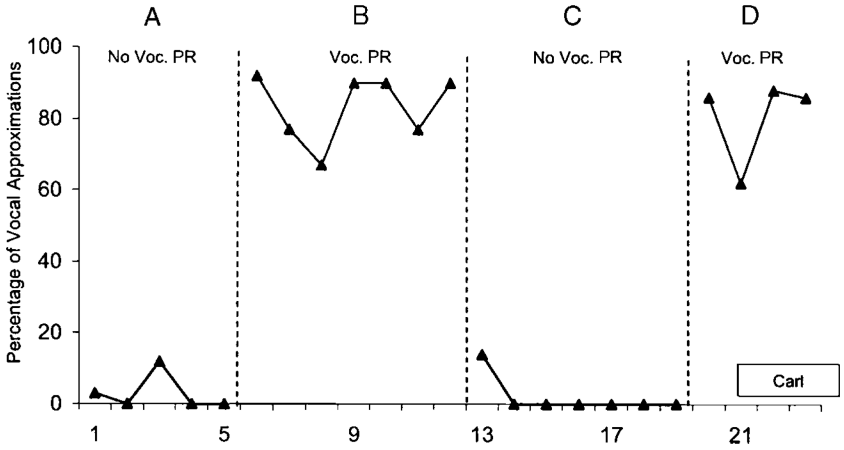

FIGURE 3.

Percentage of trials containing vocal approximations during no positive reinforcement of vocalization (baseline; see Panel A) and positive reinforcement of vocalization (see Panel B), using an ABAB design. Voc. = vocal; PR = positive reinforcement. From “The Picture Exchange Communication System: Effects on manding and speech development for school-aged children with autism,” by Tincani, Crozier, and Alazetta, 2006, Education and Training in Developmental Disabilities, 41, p. 183. Copyright 2006 by Council for Exceptional Children, Division on Developmental Disabilities. Reprinted with permission.

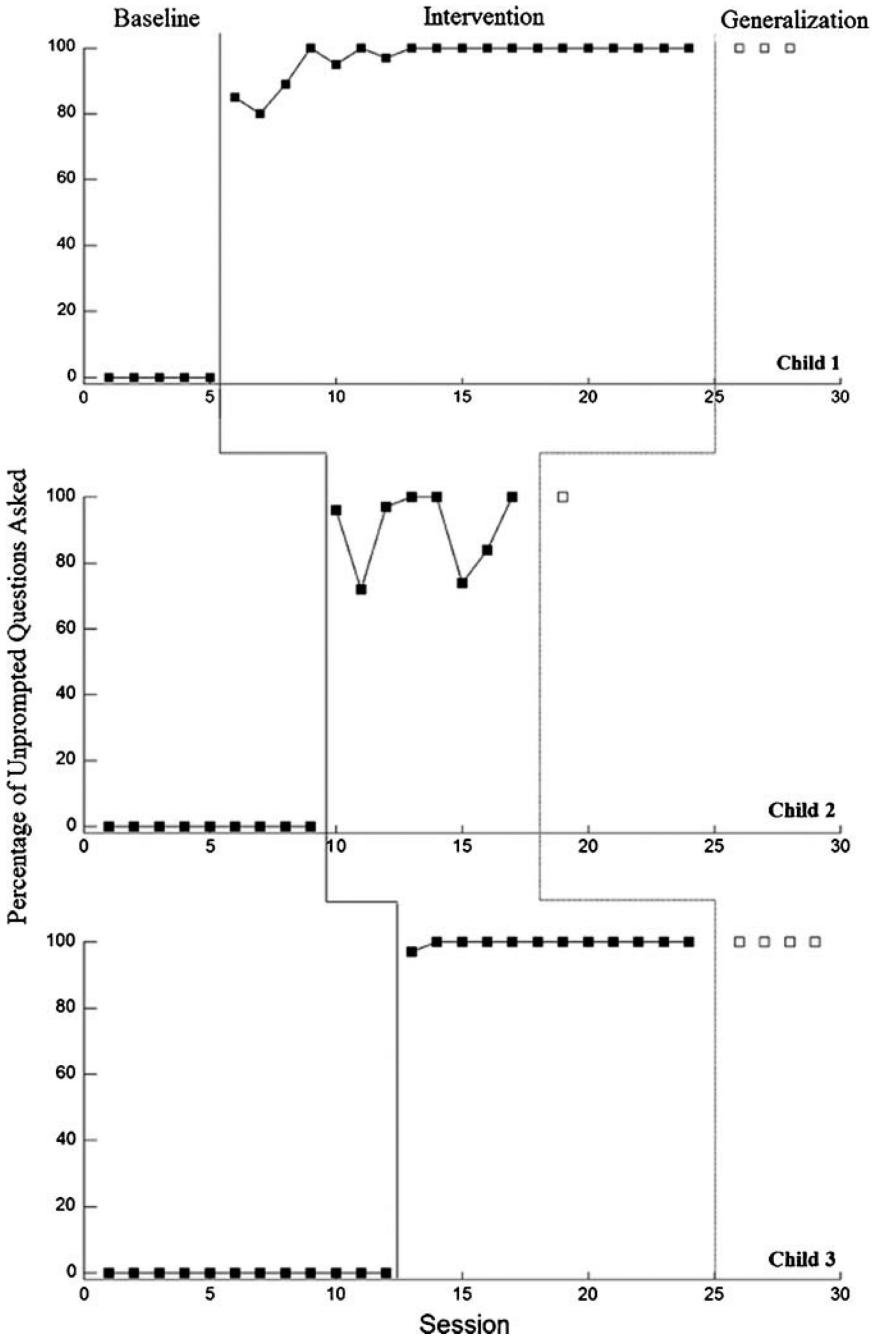

FIGURE 4.

Percentage of unprompted questions asked by three participants in baseline, intervention, and generalization sessions using a multiple-baseline, across-participants design. From "Question-asking and collateral language acquisition in children with autism," by Koegel, Koegel, Green-Hopkins, and Barnes (2010), *Journal of Autism and Developmental Disorders, 40,* p. 512. Copyright 2009 by the authors. Reprinted with permission.

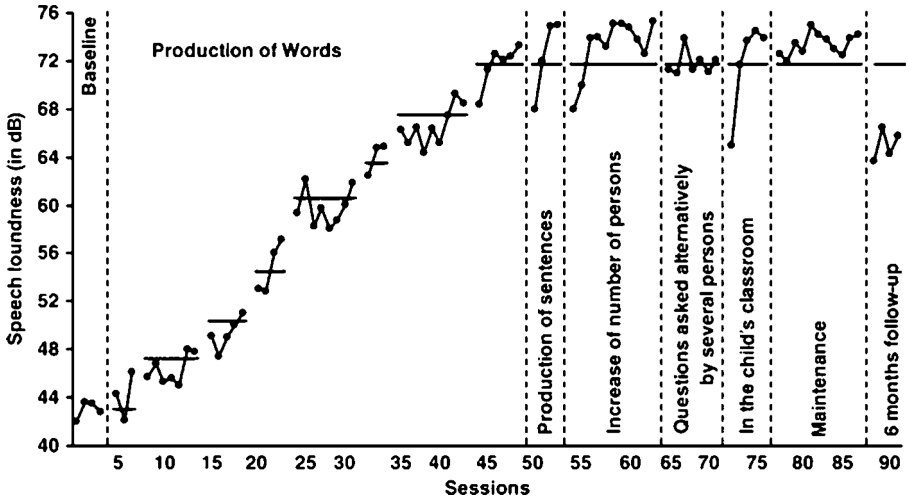

FIGURE 5.

Speech volume during a token reinforcement intervention and follow-up using a changing-criterion design. From "A controlled single-case treatment of severe long-term selective mutism in a child with mental retardation," by Facon, Sahiri, and Riviere, (2008), *Behavior Therapy*, *39*, p. 313. Copyright 2008 by Elsevier. Reprinted with permission.

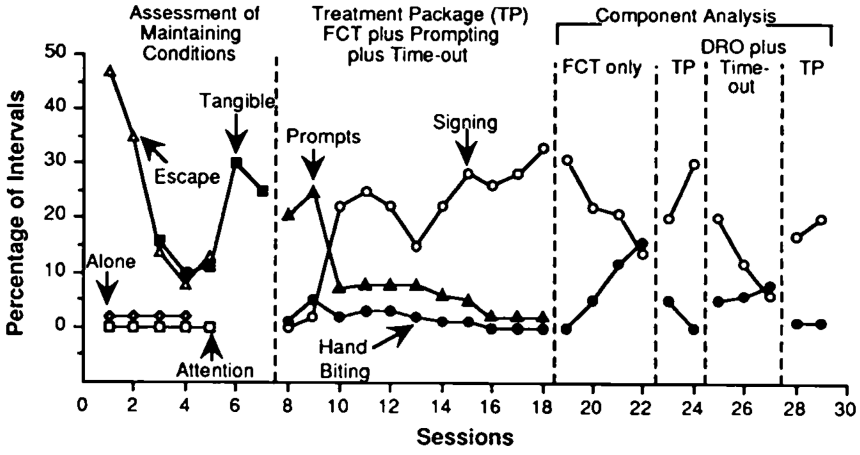

FIGURE 6.

Percent of intervals with challenging behavior and mands during functional analysis, intervention demonstration, and component analysis. From “A component analysis of functional communication training across three topographies of severe behavior problems,” by Wacker et al., 1990, *Journal of Applied Behavior Analysis, 23*, p. 424. Copyright 2008 by the Society for the Experimental Analysis of Behavior. Reprinted with permission.

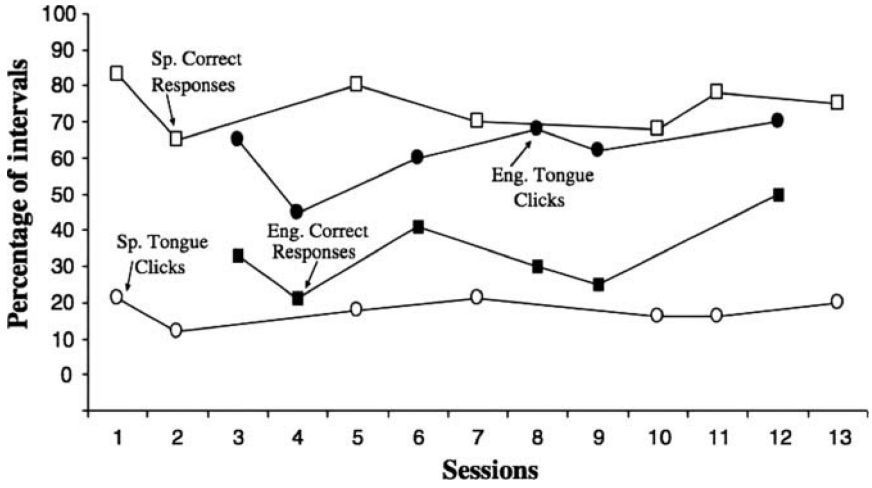

FIGURE 7.

Number of correct responses and tongue clicks during discrete trial training sessions in Spanish (Sp.) and English (Eng.) using an alternating treatments design. From “Effects of language instruction on response accuracy and challenging behavior in a child with autism,” by Lang et al., 2011, *Journal of Behavioral Education, 20,* p. 256. Copyright 2001 by Springer Science+Business Media, LLC. Reprinted with permission.

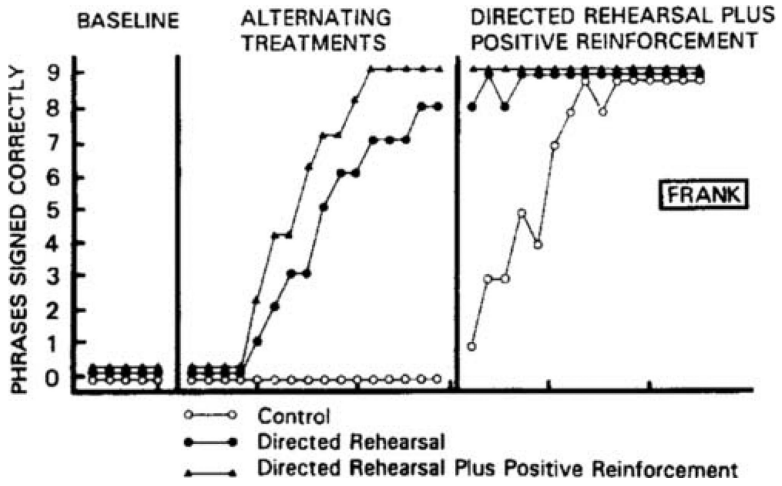

FIGURE 8.

Number of phrases signed correctly during directed rehearsal, directed rehearsal with positive reinforcement, and control sessions using an adapted alternating treatments design. From “Acquisition and generalization of manual signs by hearing-impaired adults with mental retardation,” by Conaghan, Singh, Moe, Landrum, and Ellis, 1992, *Journal of Behavioral Education, 2*, p. 192. Copyright 1992 by Human Sciences Press. Reprinted with permission.

TABLE 1

Summary of What Works Clearinghouse criteria for experimental designs.

Design element	Meets standards	Meets standards, but with reservations	Does not meet standards
Independent variable(s)	Actively manipulated by researcher	—	Researcher does not control changes to conditions
Dependent variable(s)	Measured systematically over time	—	No systematic measurement (e.g., anecdotal case study)
	Measured by more than one assessor	—	Only one assessor
	Includes interassessor agreement on at least 20% of data points in each phase	—	No interassessor agreement, only in some phases, or in less than 20% of data points
	Interassessor agreement meets minimal thresholds	—	Poor interassessor agreement
Length of phases	At least 5 data points per phase	3–4 points per phase	< 3 points per phase
Replication of effect	General:	—	< 3 replications
	3 attempts to demonstrate the effect at three points in time or with three phase repetitions		
	Reversal/withdrawal:	—	< 4 phases (e.g., AB, ABA, BAB)
	4 phases per case (e.g., ABAB)		
	Changing-criterion:	—	< 3 criteria
	3 criteria		
	Multiple-baseline/multiple-probe:	—	< 3 cases; < 6 phases
	6 phases across at least three cases		
	Alternating treatments:	Only 4 repetitions	2 treatments without baseline
	2 treatments compared to baseline or 3 treatments compared to each other		
	5 repetitions of each condition		< 4 repetitions

TABLE 2

Summary of single-subject experimental designs (SSEDs).

Design	Research questions	Advantages	Disadvantages
Pre-experimental (AB)	Does outcome X change from baseline levels with the introduction of intervention B?	• Quick and efficient to implement. • Appropriate for low-stakes decision making.	• Does not control for threats to internal validity; not an experimental design.
Withdrawal (ABA/ABAB)	Does outcome X covary with introduction and withdrawal of intervention B?	• Easy to implement, strong experimental control when effects are immediate and large.	• There are ethical considerations regarding withdrawing or reversing a potentially effective intervention. • Not all behaviors are "reversible."
Multiple-baseline/multiple-probe	Does outcome X change from baseline levels with the introduction of intervention B over multiple participants, responses, settings, etc.?	• Does not require withdrawal of intervention. • Appropriate for nonreversible behaviors.	• Ethical considerations regarding keeping individuals/behaviors in baseline conditions for a long period. • Requires multiple individuals, responses, settings, etc., that are comparable in order to replicate effects.
Changing-criterion	Do changes in the level of outcome X correspond to changes in the intervention criteria?	• Does not require reversal. • Appropriate for behaviors that can be changed gradually. • Useful for consequence-based interventions.	• Change must take place in graduated steps; not appropriate for behaviors that require immediate change. • Requires the use of incentive- or consequence-based interventions.
Multiple-treatment	What are the relative effects of interventions A and B (and C, D, etc.) on outcome X compared to each other and/or baseline levels?	• Can be extended to compare any number of interventions or variables. • Can extend a withdrawal study when effects of initial intervention are not as pronounced as expected. • Can be used to conduct component analyses of necessary and sufficient intervention components.	• Behaviors should be reversible to demonstrate relative effects. • Only comparisons between adjacent conditions are appropriate. • Can be time consuming and complicated to implement when the number of interventions being compared increases. • Results are susceptible to multiple treatment interference.
Alternating treatments	What are the relative effects of interventions A and B (and C, D, etc.) on outcome X compared with each other and/or baseline levels?	• Can be extended to compare any number of interventions or variables. • Can provide strong experimental evidence in relatively few sessions.	• Behaviors must be readily reversible to obtain differentiation between conditions. • Results are susceptible to multiple treatment interference.
Adapted alternating treatments	What are the relative effects of intervention A on outcome X and intervention B on outcome Y?	• Less prone to multiple treatment interference. • Can provide strong experimental evidence in relatively few sessions. • Does not require reversal.	• Set of behaviors or stimuli must be directly comparable for effects to be meaningful. • Potential generalization across behaviors must be considered.