Psychological Methods
2010, Vol. 15, No. 2, 124–144

1082-989X/10/$12.00 DOI: 10.1037/a0017736

Enhancing the Scientific Credibility of Single-Case Intervention Research: Randomization to the Rescue

Thomas R. Kratochwill
University of Wisconsin-Madison

Joel R. Levin
University of Arizona

In recent years, single-case designs have increasingly been used to establish an empirical basis for evidence-based interventions and techniques in a variety of disciplines, including psychology and education. Although traditional single-case designs have typically not met the criteria for a randomized controlled trial relative to conventional multiple-participant experimental designs, there are procedures that can be adopted to create a randomized experiment in this class of experimental design. Our two major purposes in writing this article were (a) to review the various types of single-case design that have been and can be used in psychological and educational intervention research and (b) to incorporate randomized experimental schemes into these designs, thereby improving them so that investigators can draw more valid conclusions from their research. For each traditional single-case design type reviewed, we provide illustrations of how various forms of randomization can be introduced into the basic design structure. We conclude by recommending that traditional single-case intervention designs be transformed into more scientifically credible randomized single-case intervention designs whenever the research conditions under consideration permit.

Keywords: single-case designs, intervention research, randomization schemes

Traditionally, a premium has been placed on the use of randomized experimental design methodology—and synonymously, in selected contexts, randomized controlled trials (RCTs)—for evaluations of the efficacy of psychological and educational interventions. Indeed, the term gold standard has often been applied to intervention research that adheres to the principle of randomization for imputing causal connections between interventions and outcomes (e.g., Mosteller & Boruch, 2002; Reyna, 2005; Shadish, Cook, & Campbell, 2002; Shavelson & Towne, 2002; Slavin, 2002). RCTs are typically experiments that minimize (although not necessarily eliminate) major internal-validity threats to drawing scientifically valid inferences from the data (Shadish et al., 2002). Multiple-participant experimental designs (referred to here as *group designs*) have played the primary role in the development of what are contemporaneously referred to as *scientifically credible* (e.g., Levin, 1994) and *evidence-based* (e.g., Hayes, Barlow, & Nelson-Gray, 1999) interventions. In such research, random assignment of participants (or analogous "units," discussed in a following section) to the different intervention (or intervention and control) conditions is the typical method for helping to assuage internal invalidity concerns.

Single-case (formerly, until about the mid 1980s, "single-subject") research designs have also been used to establish an empirical basis for evidence-based interventions and techniques in a variety of disciplines, and in recent years, there has been renewed interest in the role that single-case designs can play in establishing the scientific basis for various educational and psychological interventions (Horner et al., 2005; Kratochwill, 2007). Single-case designs have been adopted in several areas of scientific research, with the most prominent publication outlet being the *Journal of Applied Behavior Analysis* that initiated publication in 1968. Since that time, numerous journals publish single-case design studies in such fields as psychology (i.e., clinical psychology, counseling psychology, school psychology), social work, speech and language fields, and special education. The increasing role that single-case design plays in research in education has recently been emphasized through the establishment of the Single-Case Design Panel sponsored by the Institute of Education Sciences of the U.S. Department of Education. The specific charge to this panel is one of developing multiple-dimension coding standards for single-case designs, with the ultimate goal of (a) summarizing single-case studies in select substantive areas and (b) reporting results of literature reviews in the What Works Clearinghouse (see www.whatworks.ed.gov).

Despite many areas of application, traditional single-case designs have typically not met the criteria for an RCT and thus, relative to conventional multiple-unit (or group) designs, they are less likely to be included in literature reviews to establish a research domain's evidence. In fact, a perusal of journals that publish single-case designs exclusively (e.g., *Journal of Applied Behavior Analysis*) or frequently (e.g., *School Psychology Review*) reveals that randomization, the aforementioned hallmark of scien-

Thomas R. Kratochwill, Department of Educational Psychology, University of Wisconsin–Madison; Joel R. Levin, Department of Educational Psychology, College of Education, University of Arizona.

The preparation of this article was facilitated by a grant to Thomas R. Kratochwill and Joel R. Levin from the Institute of Education Sciences (IES Grant No. R324U060001). We are grateful to Jacquelyn Buckley of IES for her support and to Cathlin Foy for her assistance in constructing the figures.

Correspondence concerning this article should be addressed to Thomas R. Kratochwill, Department of Educational Psychology, 1025 W. Johnson St., University of Wisconsin, Madison, WI 53706. E-mail: tomkat@education.wisc.edu

tifically credible treatment or intervention research, is not applied in the basic formation of the designs. Adding randomization to the structure of single-case designs can augment this type of research in at least two important respects. First, randomization can strengthen the internal validity of these designs. As is discussed in the following section, researchers using single-case designs currently depend on some form of replication to argue against internal validity threats. Although replication represents a strong design option, randomization represents a higher level of methodological soundness in such applications. In this article, we introduce two single-case randomization schemes and two different randomization variations, simple and blocked. Second, including a randomization scheme in the design allows the researcher to apply various statistical tests based on randomization models, which can improve the statistical-conclusion validity of the research (for several randomization–test possibilities, see Borckardt et al., 2008; Edgington & Onghena, 2007; Koehler & Levin, 1998; Kratochwill & Levin, 1992; Levin & Wampold, 1999; and Wampold & Worsham, 1986).

The two major purposes of this article are (a) to review the various types of single-case research designs that can be used in intervention research and (b) to provide scientifically credible extensions of these designs—in particular, extensions incorporating randomized experimental schemes that allow investigators to draw more valid inferences from their research. Space limitations do not permit us to discuss common data-analysis strategies associated with these design variations here—in particular, visual–graphical analysis (e.g., Fisher, Kelley, & Lomas, 2003) and statistical analysis, including the methods mentioned in the immediately preceding paragraph. Several different visual–graphical and statistical approaches are detailed, respectively, in reviews currently being prepared by the present authors.

Single-Case Designs in Evidence-Based Intervention Research

Single-case research design has a long history of application throughout the social and educational sciences and has contributed greatly to the empirical basis for a variety of practices. Many of the current design structures used in single-case research have grown out of the experimental and applied behavioral analysis traditions of psychological research, but they have also been developed and applied in various other psychological and educational disciplines (Kratochwill, 1978; Kratochwill & Levin, 1992). As implied in the name, single-case designs have traditionally involved the use of a single participant in the experiment; yet, as elucidated in the following sections, various types of replication with more than one participant can be fit into this type of research enterprise. Over the years, several textbooks have featured various conceptual, methodological, and statistical aspects of single-case research design (see, for example, Kratochwill, 1992).

Characteristics of Single-Case Research Methodology

Single-case research designs can be regarded as a class of experimental research (as opposed to correlational, survey, or descriptive investigations). Like other experimental methodologies, single-case intervention designs seek to establish causal relationships between the independent (here, intervention) and dependent (outcome) variables. Several features or characteristics of single-case designs can be identified (Horner et al., 2005; Kratochwill, 1978; Kratochwill & Levin, 1992).

Units of Assignment and Analysis

As noted earlier, the term *single-case design* might suggest that one participant is the primary focus of the research study. This focus reflects the rich history of the use of these designs in the experimental and applied behavior analysis traditions, where an emphasis was placed on understanding an individual's behavior—particularly, on an individual's response to an intervention such as reinforcement or related operant variables (Kazdin, 1982; Kratochwill, 1978). As these designs expanded from their behavioral traditions, applications beyond the individual or a small number of participants became more commonplace. In single-case designs as discussed here, the "units" to whom an intervention is separately (and generally independently) administered can be a single participant, a dyad, a small group, a classroom, or even a large institutional population (e.g., school, community, hospital, or prison; for additional description and discussion, see Levin, O'Donnell, & Kratochwill, 2003). The units associated with participant assignment to intervention conditions (what are termed here the *methodological units*) and the units to which various inferential statistical techniques are applied (the *analytic units*) represent important considerations that we will later address in the context of the use of randomization in single-case (including group-administered) intervention designs (see also Baldwin, Murray, & Shadish, 2005; Cappella, Massetti, & Yampolsky, 2009; Levin, 1985, 2005; and Levin et al., 2003). In each of the design variations in single-case designs, one or more intervention conditions are compared with one or more baseline or nonintervention conditions, with the basis for drawing valid inferences consisting of either (a) a change in the unit(s) between baseline and intervention phases or (b) a differential change of intervention and control units between baseline and intervention phases.

Unit Descriptions

Another important feature of single-case research is the careful specification of contextual features of the experiment, including the participants, the setting, and the selection characteristics, among other features (Horner et al., 2005). When a single participant is involved in the study, typically very detailed descriptions are provided so that readers can assess the degree to which his or her characteristics are representative of the participant population and situational contexts to which the results can be generalized (i.e., population and ecological external-validity concerns; see Shadish et al., 2002). Thus, this is essentially a process of logical generalization in which the characteristics of the participant(s) and experimental context are compared with those that the researcher–practitioner wishes to generalize in some applied setting (Hayes et al., 1999). A thorough description of the experiment's participant and situational characteristics is also considered important in literature reviews when a body of evidence is being presented to support a particular evidence-based intervention or procedure (Kratochwill, 2007).

Outcome Measures

Single-case research designs involve the selection of one or more outcome measures that provide the basis for comparing intervention and nonintervention conditions from one phase of the study to the next (i.e., from a baseline to an intervention phase). Horner et al. (2005) described several features of outcome measures in single-case research that are important in this type of experimental investigation. First, outcome measures are operationally defined to allow valid measurement of the variable and replication of the assessment process in future research. Operational specification is required to understand what the outcome measures assess, and it represents an important construct validity issue in all experimental research.

Second, outcome measures in single-case research are assessed repeatedly within and across phases of the experiment in a time-series fashion. The purpose of this ongoing or repeated assessment is primarily to establish the unit's baseline or preintervention performance (similar to a pretest) and secondarily to allow comparison of these patterns of performance from one phase of the experiment to the next. This type of ongoing assessment of the outcome measure(s), from baseline to intervention phases, allows comparison of the participant's performance as the participant serves as his or her own control. Within this pattern of performance, various features of the data are given special consideration, including level (mean), trend (slope), and variability in the time series. Continuous measurement of these features between phases of the study provides an important basis for drawing valid inferences from the experiment. In single-case designs, compared with traditional group within-subject designs (also known as repeated-measures designs), the ratio of the number of units to the number of measures (here, time points at which measures or observations are taken) is typically much smaller.[1]

Third, the outcome measures are assessed for interobserver consistency throughout the experiment. Typically, independent observers establish standards or criteria for consistency while recording various features or the typography of the outcome measures. Indices of interobserver reliability or agreement (e.g., some type of correlation coefficient, percentage of agreement, Cohen's kappa) are used for reporting this kind of consistency information. Establishing standards of interobserver reliability or agreement on the outcome measure has been a long-standing tradition in behavioral analysis research, where single-case designs are frequently implemented.

Fourth, with respect to applied and clinical research in psychology and education, outcome measures are generally selected in consideration of their "social validity" (Wolf, 1978, p. 203). Although outcome measures represent an important construct validity issue in the experiment and, therefore, must be operationally specified, in most applied psychological and educational research, the societal importance of the outcomes is also a prime consideration. Social validation consists of at least two components, including social comparison and subjective evaluation (Kazdin, 1977, 1982). Social comparison involves comparing the participant in the research with peers who are regarded as functioning normally (e.g., average in academic skills or nondeviant in their behavior). Subjective evaluation involves the judgment by significant others that prior to an intervention the participant's performance or behavior was a problem or concern and that following the intervention the participant's performance or behavior has improved. For example, in consideration of the social validity of an intervention designed for a child who has been displaying aggressive behavior, reducing aggression would represent a socially important focus of the intervention; therefore, a clinically important outcome would be achieved if, following the intervention, the child's care providers (e.g., teachers, parents) were to judge the child's aggression to have been reduced. Thus, researchers in applied and clinical intervention research typically have dual criteria to take into account concerning the outcomes of the study: namely, the targeted outcome itself and the social consequences of the outcome. The social consequences measures can then be used to attest to the effectiveness of the intervention when the results are reported either descriptively, graphically, or statistically.

Researcher-Manipulated Variables

As with other experimental research, single-case research investigations are defined by systematic manipulation of some independent variable during the course of the experiment. Two important considerations emerge here. First, the construct validity of the manipulated variable is important. Interventions must be appropriately defined on the basis of meaningful constructs that represent the focus of the intervention effort. Second, an assessment of the intervention's implementation integrity (or "treatment fidelity"—see, for example, Gresham, 1997, and Hagermoser Sannetti & Kratochwill, 2005) is typically scheduled to determine whether the actual intervention was implemented as intended throughout the duration of the experiment. When both of these conditions are met, a single-case intervention study increases the validity of inferences concerning the researcher's interpretation of the effects observed and analyzed.

Baseline–Control Comparison Conditions

Single-case intervention designs always involve a systematic comparison of two or more experimental phases. One of these phases is generally a baseline, control, or preintervention phase, and the other is an intervention phase. However, preintervention phases are defined relative to what occurs during the intervention phases. In other words, an intervention can be compared with some already existing intervention or precondition in the research setting. Thus, although one option is to conduct the study with no active intervention (or control) condition operating during the baseline phase, a researcher might alternatively wish to compare

[1] Within-subjects design hypotheses have traditionally been tested statistically either by (a) parametric univariate repeated-measures analyses, which are often questionable in terms of the assumptions underlying them or (b) parametric multivariate repeated-measures analyses, which are questionable when the units-to-measures ratio is small and which cannot be conducted at all when the number of units is less than or equal to the number of measures (see, for example, Maxwell & Delaney, 2004). In more recent years, growth-curve analyses based on hierarchical linear modeling (HLM) have been applied. On the other hand, interrupted time-series designs (including single-case intervention designs) have generally been statistically analyzed—when they have not been analyzed by visual and graphical techniques—by a vast array of regression, analysis-of-variance, time-series, randomization, and analysis-modeling approaches.

phases containing two different interventions (e.g., pre-existing and new) within a single-case study, depending on the nature of the experimental arrangements. As will be seen in the following sections, in single-case designs involving more than one experimental unit, assignment of different units to different experimental conditions (e.g., Unit 1 to baseline followed by Intervention 1, and Unit 2 to baseline followed by Control or Intervention 2) is also possible.

Basic Single-Case Design Structures for Establishing Experimental Control

Like group intervention designs, traditional single-case intervention designs are structured to take into account major threats to internal validity in the study. However, unlike researchers using the most scientifically credible group-intervention designs, researchers using traditional single-case designs typically have not utilized randomization to reduce or eliminate internal validity threats. As an alternative, single-case design researchers address internal validity concerns through some type of replication during the course of the experiment. The replication criterion advanced by Horner et al. (2005, p. 168) represents a fundamental characteristic of single-case designs: "In most cases experimental control is demonstrated when the design documents three demonstrations of the experimental effect at three different points in time with a single participant (within-subject replication), or across different participants (inter-subject replication)." As these authors noted, an experimental effect is shown when the predicted changes in various outcome measures covary with the manipulated variable after the trend, level, and variability of the baseline (preintervention) series have been taken into account. However, it must be noted that there is no empirical basis for Horner et al.'s recommendation of three demonstrations; rather, this represents a conceptual norm among published research and textbooks that recommend methodological standards for single-case intervention designs.

The methodology for establishing experimental control that relies on replication generally falls into three major design types, as originally discussed by Hayes (1981).[2] The design types, reproduced in Table 1, are the basic building blocks for the construction of single-case designs. In this article, we expand on the building blocks by adapting various randomization tactics that will strengthen a single-case researcher's argument for drawing valid inferences from the study. We first review each of the Hayes design types. Then, within each design type, we provide suggestions for (and illustrations of) the incorporation of randomization into the basic design structure. Each design type is accompanied by an example from the published literature, along with a description of how some form of randomization could have been built into the original experiment. Our purpose in so doing is to illustrate how to transform traditional single-case intervention designs into more scientifically credible randomized single-case intervention designs.

Within-Series Single-Case Designs

In within-series designs, participant performance is measured within each condition of the investigation and compared between or among conditions. The most fundamental within-series intervention design is the two-conditions AB design, where the A condition is a baseline or preintervention series or phase and the B condition is an intervention series or phase. This design is sometimes called the *basic time-series design* or an *interrupted (or two-phase) time-series design* and is truly quasi-experimental in that (a) no type of randomization is used and (b) no replication of the baseline and intervention phases is scheduled in the design (Shadish et al., 2002). However, as we will note later, even in this most basic within-series design consisting of only one experimental unit receiving one baseline and one intervention series of observations or measures, the researcher can structure the study so that it includes a randomization component—thereby, enhancing the design's scientific credibility.

The most common form of the within-series design that meets the replication criterion advanced by Horner et al. (2005) is the ABAB design, which includes four alternating baseline and intervention phases (or four alternating intervention phases, BCBC), and hence there is an opportunity to produce a within-subject replication of the intervention effect. Specifically, in this four-phase design, intervention effects are assessed during the first B phase and then again during the second B or replication phase. Figure 1 illustrates the ABAB design with hypothetical data. The four-phase ABAB design was initially proposed (and is now universally accepted) as a more scientifically and clinically convincing single-case design relative to both the historically earlier implemented basic AB and single-reversal (or "return-to-baseline") ABA designs (see, for example, Hersen & Barlow, 1976). In addition to satisfying Horner et al.'s (2005) within-subject replication criterion, the ABAB design has numerous methodological merits. Moreover, the external validity of the design can be further enhanced by the inclusion of more than one case (i.e., a replicated ABAB design, discussed later in this article), thereby providing an opportunity to produce a between-subjects replication. The four-phase ABAB design currently represents the minimum within-series standard for single-case intervention researchers seeking to publish their work in top-tier academic journals (see, for example, Hayes et al., 1999, and Horner et al., 2005).

As was noted earlier, the just-discussed alternating-phase design can also be applied to two different interventions, B and C. For example, the researcher might compare Intervention B with Intervention C at several different time points or sessions in an alternating replicated series (i.e., BCBC . . . BC). The designation of the phase labels as B and C is somewhat arbitrary for, as we noted earlier, the "baseline" phase might actually consist of some known intervention that is already operating in the context of the experiment. The important issue in comparing two intervention conditions is whether the manipulated variable (Intervention B vs. Intervention C) is strong enough to produce clear, predicted, and different intervention effects, given the absence of a baseline condition. Some regard this test as a higher criterion of evidence for evidence-based interventions (Kazdin, 2004).

The within-series ABAB . . . AB design (or its ABCABC . . . ABC extension) is typically designated as a "simple phase-change" design structure. However, it is also possible for the

[2] We will adopt Hayes' (1981) design type designations as an organizational framework here, even though there are some fuzzy distinctions among them in contemporary single-case applications.

Table 1
Major Types of Single-Case Designs and Associated Characteristics

Design type/representative example	Characteristics
Within series Simple phase change—for example, AB; ABA; ABAB; BCBC Couples phase change—for example, B(B + C)B; C(B + C)C	In these designs, estimates of level, trend, and variability within a data series are assessed under similar conditions; the manipulated variable is introduced and concomitant changes in the outcome measure(s) are assessed in the level, trend, and variability between phases of the series.
Between series: Alternating intervention design	In these designs, estimates of level, trend, and variability in a data series are assessed in measures within specific conditions and across time. Changes/differences in the outcome measure(s) are assessed by comparing the series associated with different conditions.
Combined series: Multiple baseline, for example, across subjects, across behaviors, across situations	In these designs, comparisons are made both between and within a data series. Repetitions (replications) of a single simple phase change are scheduled, each with a new series and in which both the length and timing of the phase change differ across repetitions.

Note. A represents a baseline series; B and C represent two different intervention series. From "Single-Case Experimental Designs and Empirical Clinical Practice," by S. C. Hayes, 1981, *Journal of Consulting and Clinical Psychology, 49,* p. 208. Adapted with permission.

researcher to structure more "complex phase-change" strategies in this class of within-series designs. In all cases, the designs in this domain operate under the same logic; more complex phase-change structures allow the investigator to examine the effects of multiple interventions or compare interventions with each other. In this way, the researcher can manipulate various combinations of interventions that are tested against each of their individual components, as the following example illustrates.

Consider an alternating two-conditions within-series design in which a researcher compares one intervention, designated here as B, with an intervention "package" that consists of three components, B + (C + D). The researcher examines (through a within-series replication) the intervention package across phases of the study using the same basic design logic: B B + (C + D) B B + (C + D) . . . B B + (C + D) as that for the ABAB . . . AB design. The conditions compared in the experiment can encompass more components, depending on the nature of the intervention package under investigation—for example, B + (C + D) + (E + F) + G versus any of the individual components. Note that this design variation includes the same replication criterion when extended to multiple within-phase intervention components as when it is applied with only one intervention component within a phase.

The within-series design is actually a very common research strategy within psychology and education and has been used

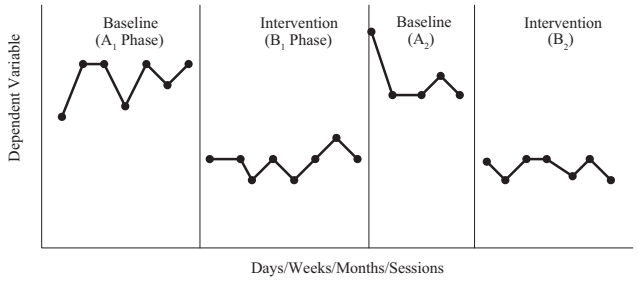

Figure 1. Hypothetical data for an ABAB design with one experimental unit.

numerous times with different participant populations and in different contexts, especially in applications of applied behavior analysis techniques and procedures. However, these within-series design structures have some potential shortcomings. A major issue is that they require a withdrawal of the intervention as part of the replication requirement (designated here as the B phase in an ABAB design). However, for a variety of reasons, withdrawal of the intervention may not result in the outcome measures returning to baseline levels. Depending on how this nonreturn to baseline is manifested in terms of the level and trend of the series, it may be more difficult to draw inferences about the magnitude and extent of the intervention effect or even to determine whether any intervention effect occurred. Because of this issue, the within-series design typically is not recommended under conditions in which participants would be expected to acquire certain skills or would learn new behaviors that would likely result in their not returning to baseline levels of performance. In other words, the design has some features that the researcher should be cognizant of and take into account when considering this option. Of even greater concern perhaps, a within-series design might not be appropriate under conditions in which withdrawal of the intervention is unethical or would provide physical or psychological discomfort for the participants in the investigation.

Another type of within-series design, mentioned only in passing here, is the *changing criterion design* (*CCD*), a within-series variation that has been used only infrequently outside its applied behavior analysis origins. In its most straightforward application, researchers use this design to attempt to rule out threats to a study's internal validity by providing opportunities for the outcome measure to covary with changing criteria that are scheduled in a series of predetermined steps or "subphases" within the study (Hall & Fox, 1977; Hartmann & Hall, 1976). The most basic form of the CCD begins with a baseline (A) phase, followed by a series of intervention (B) phases, with the intervention implemented continuously over time as changing criterion levels for "improved" outcome-measure performance are specified. Repeated replications of these stepwise changes and corresponding changes in the out-

come measure argue for the credibility of the intervention effect across time.

Between-Series Single-Case Designs

Between-series designs are structured to provide a comparison of two or more conditions (e.g., a baseline and intervention or two intervention series) in a more rapid fashion than is possible in conventional within-series designs. Two applications are the *alternating intervention design* (*AID*) and the *simultaneous intervention design* (*SID*). The AID is the most frequently used design option and allows researchers to expose a participant to different interventions administered in close proximity for equal periods of time (Barlow & Hayes, 1979; Hayes et al., 1999). Specifically, in the design, the researcher might establish a baseline and then alternate between two intervention series for a brief period of time. For example, one intervention could be administered in a morning session and a second intervention in an afternoon session over several days. The interventions are alternated systematically by counterbalancing or, as we will suggest later for its true experimental counterpart, by randomly assigning the two conditions to the study's different sessions. With the AID, the researcher can compare two or more conditions or interventions in a relatively brief period of time. This strategy allows researchers to avoid some of the major disadvantages of other multiple-intervention within-series designs, in which (as was discussed earlier) intervention withdrawal for an extended period of time is often necessary before a second intervention can be introduced. Figure 2 illustrates a two-intervention AID with hypothetical data.

The SID, a unique application in the between-series domain, involves presenting interventions to the units simultaneously (Kazdin & Hartmann, 1978). For example, a researcher may present two rewards to a participant simultaneously, with the option for the participant to select the more preferred reward. The simultaneous presentations are repeated over a designated time period and arranged in either counterbalanced or random order. The simultaneous presentation of interventions in the design structure, however, does not ensure that the participant is exposed to all interventions for equal time amounts or durations. Instead, the design guarantees that the different interventions are simultaneously and equally available in each session, with the participant able to select the particular intervention to which he or she prefers to be exposed. Thus, it is possible that the SID could provide a researcher with information on client responsiveness to interventions in which differential preferences are likely to exist. Like the AID, the SID is not restricted to two conditions or interventions, and so a researcher can make available several interventions and compare a participant's preferences for the different interventions. However, the conditions and implementation of such a multiple-intervention study can become quite complex, and therefore, for practical reasons, single-case research based on the SID and AID typically includes only two interventions. Illustrative applications of these designs are found in numerous places in the literature and are discussed in several sources (see, for example, Hayes et al., 1999).

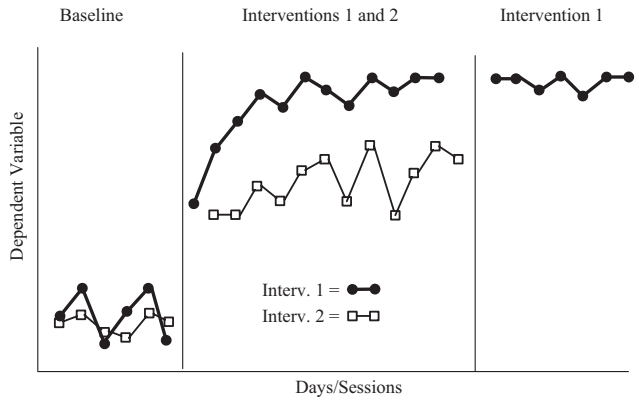

Figure 2. Hypothetical example of an alternating intervention design. During the baseline phase, a series of alternating "Intervention 1" and "Intervention 2" baseline assessments is made (e.g., during a morning and afternoon session over 6 successive days). During the intervention phase, the two interventions are administered in an alternating fashion (e.g., in 12 morning and 12 afternoon sessions) over several days, with the data plotted by intervention type. In the final phase, the more effective intervention (Intervention 1) is implemented.

Combined-Series Single-Case Designs

With combined-series designs, the researcher makes both within- and between-series comparisons to draw valid inferences from the data. The most common design in this domain is called the *multiple-baseline design* (*MBD*) and includes a simple within-phase element while replicating the intervention across participants (or other units), settings, or behaviors (or, more generally, other outcome measures). The internal validity of the design is strengthened through a staggering or sequential introduction of the interventions across time, with desired changes in the outcome measure occurring repeatedly and selectively with the successive intervention introductions (see Levin, 1992, pp. 216–217, for additional discussion of the favorable internal-validity characteristics associated with this design). The MBD is typically structured so that at least four replications are scheduled within the experiment, although applications of the design can range from two to several (i.e., more than four) replications, depending on the research questions, circumstances of the experiment, and practical and logistical issues. As will be discussed later, the extent to which the intervention's effect is similar across replications helps researchers to establish different aspects of the external validity (generalizability) of the intervention. Numerous examples of the design have been published in the literature and are discussed in greater detail in several sources (e.g., Hayes et al., 1999). Figure 3 illustrates the MBD across participants with hypothetical data.

Incorporating Randomization Into Single-Case Intervention Designs

In this section of the article, we present a case for incorporating randomization into single-case intervention designs, thereby strengthening the causal conclusions that can be drawn from them. We first provide an overview of the importance of randomization, describe the various types of randomization that can be considered for single-case designs, and then discuss specific randomization strategies that can be accommodated by the within-, between-, and combined-series designs that were presented previously. Examples

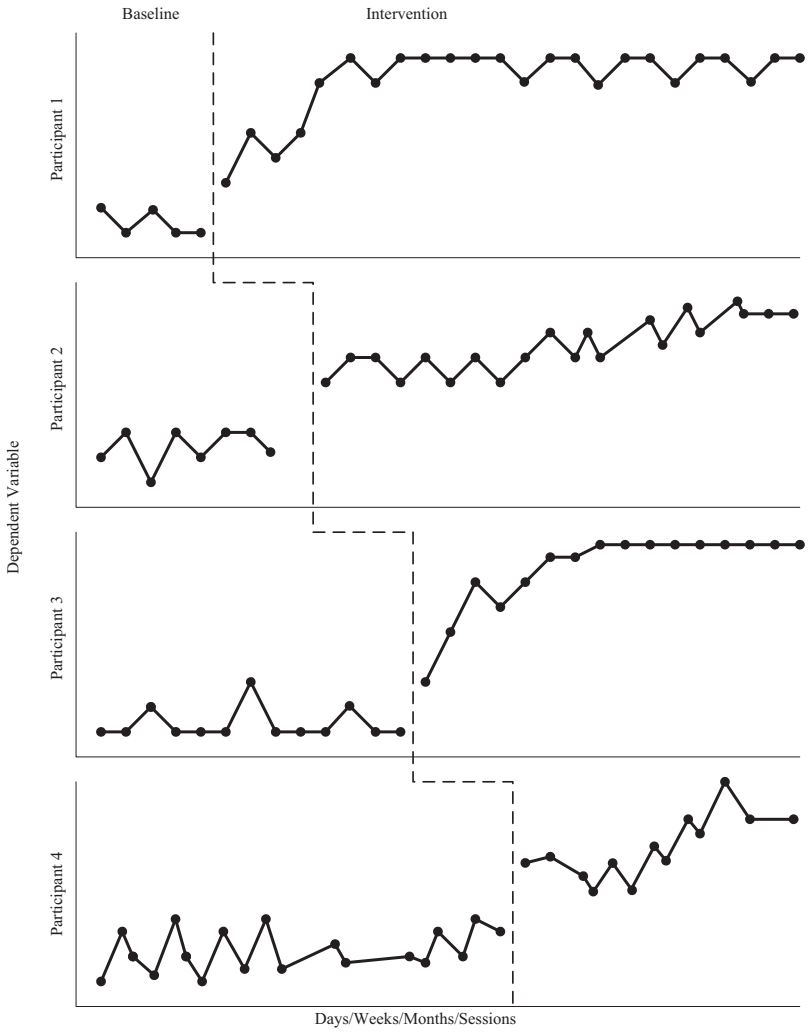

Figure 3. Hypothetical data for a multiple-baseline design across units, in which the intervention is introduced to four units at different points in time. Note the repeated and selective increase in outcome-measure performance in temporal correspondence with the intervention's introduction.

from the published single-case research design literature are presented to illustrate how some form of randomization can be built into these designs.

A Stage Model of Educational/Psychological Intervention Research: Applications to Single-Case Designs

In a previous chapter, Levin et al. (2003) offered a number of suggestions directed at improving the "awful reputation" of educational research (Kaestle, 1993, p. 23). Specifically, we recommended that intervention researchers should design and conduct research that:

> (1) makes explicit different research "stages," each of which is associated with its own assumptions, purposes, methodologies, and standards of evidence; (2) concerns itself with research credibility through high standards of internal validity; (3) concerns itself with research credibility through high standards of external validity and educational/societal importance; and most significantly (4) includes a critical stage that has heretofore been missing in the vast majority of intervention research, namely, a randomized "classroom trials" link (modeled after the "clinical trials" stage of medical research) between the initial development and limited testing of the intervention and the prescription and implementation of it. (Levin et al., 2003, pp. 569–570)

In our conceptual framework (developed by Levin & O'Donnell, 1999, and presented here as Figure 4), the *randomized classroom trials stage* refers to a broad range of educational and psychological randomized experiments, ranging from those with the experimental units composed of individual participants to those consisting of one or more classrooms, schools, or communities. Hereafter, we refer to this stage as the *randomized trials stage* of credible intervention research. We argue that the scientific credibility of traditional single-case designs would be greatly enhanced by including a randomized trial component in the evaluation of experimental interventions or programs.

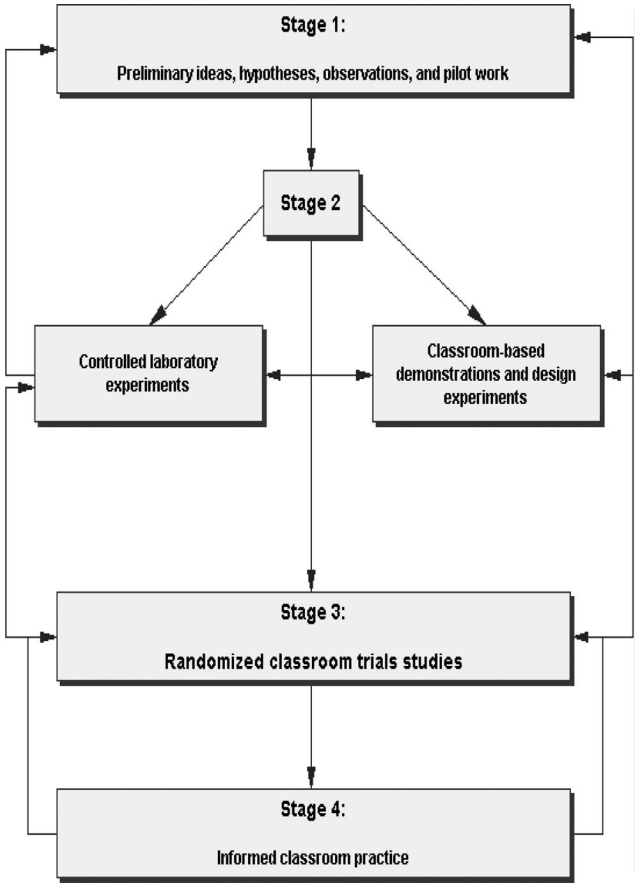

Figure 4. A stage model of educational intervention research. From "What to Do About Educational Research's Credibility Gaps?" by J. R. Levin & A. M. O'Donnell, 1999, *Issues in Education: Contributions from Educational Psychology, 5*, p. 205. Copyright 1999 by Information Age Publishing. Reprinted with permission.

As noted earlier, the methodological and analytic units are important desiderata in our case for extending randomization to single-case designs. The researcher must first determine the units upon which random assignment to the study's conditions or levels are based; the methodological units could refer to a single participant in a behavior-modification study, dyads, or small groups of students in a tutoring context, classrooms or schools in a study of alternative classroom management strategies, or schools districts where various institutional or systemic changes are being evaluated. Alternatively, and as will be seen in our discussion of specific designs, randomization might also be represented by the within- and between-series sequences in which the study's intervention conditions are implemented, as well as by the assignment of units to different "intervention start points" within the time series. Following a determination of the desired methodological units and the conduct of the experiment, the researcher must then adopt a statistical strategy for which the methodological and analytic units are congruent. In the present context, by *congruent*, we mean that studies in which interventions are assigned randomly and administered independently to individuals are analyzed with statistical methods that treat individuals as the units of analysis, whereas studies involving the assignment of interventions and administration of them to intact groups are analyzed with statistical methods that take the group structure (or cluster) into account. Discussion of specific "units-appropriate" (Levin, 2005, p. 13) single-case statistical strategies is beyond the scope of the present article but is the primary focus of another manuscript currently being prepared by the present authors.

Considerations in Incorporating Randomization Into Single-Case Designs

Incorporating randomization into single-case designs requires a more flexible type of thinking about experimental randomization, in that, unlike traditional group designs, single-case designs frequently do not have multiple independent units to be randomly assigned to the study's intervention conditions (but see later discussion). Specifically, as was just noted, what *can* be randomly assigned (instead of units to conditions) are the within-series sequencing of A and B phases, the specific time points at which each A and B phase commence, or both in the case of multiple-unit replication designs (see, for example, Edgington, 1975, 1992; Edgington & Onghena, 2007; Koehler & Levin, 1998; Levin & Wampold, 1999; Max & Onghena, 1999; Onghena & Edgington, 2005; and Todman & Dugard, 2001). In this article, we refer to the former randomization scheme as *randomized phase-order designs* and to the latter as *randomized phase start-point designs*. With these two schemes at hand, a single-case intervention researcher is able to incorporate randomization into a design structure that does not, on the surface, possess the internal validity criteria of a true experiment (e.g., such as in an AB design). In addition, in single-case intervention-comparison experiments where more than one independent unit is available, random assignment of units to intervention conditions can be combined with both of the randomization schemes just discussed. We revisit these randomization schemes and associated issues throughout this article.

Further clarification of how randomization can be applied in single-case designs is related to a presentation by Reichardt (2006) on "size-of-effect factors" in intervention research. He noted that an intervention effect is a function of four factors, consisting of recipient, setting, time, and outcome measures. The application of the earlier discussed MBD nicely illustrates how the effect of an intervention can vary according to these four factors.

The *recipient* refers to the entities or units (typically, but not necessarily, participants) that receive the intervention and to which the outcome measures are administered. In addition to an intervention effect being attributed to the intervention per se, it is also a function of the participants inasmuch as the effect can be moderated by the study's recipients or units. In MBD across participants that we discussed earlier, the researcher hopes that the effect of the intervention (and its magnitude) will be replicated in each series and therefore can be thought to generalize across recipients or units. The extent to which the intervention effect is similar across recipients helps to enhance the population external validity of the intervention (Bracht & Glass, 1968).

As Reichardt (2006, p. 2) details, the *setting* refers to the environment or situational context in which the intervention is initially implemented and subsequently assessed with respect to the various outcome measures of the study. The setting can refer to

both physical characteristics of the research investigation and functional characteristics that occur in the context of the intervention. In the application of the MBD where the intervention is implemented across settings, the intervention may have unique effects on the recipients or units, depending on the setting in which it is applied (e.g., home, school, community). The extent to which intervention effects are similar in the study's different settings helps to increase the intervention's *ecological* external validity (Bracht & Glass, 1968). Alternatively, the extent to which the effects differ across settings can represent a desirable *discriminant* validity aspect of the intervention (Campbell & Fiske, 1959; Levin, 1992).

The *time* dimension refers to "the chronological times at which both [an intervention] is implemented and an outcome [measure] is assessed, which thereby specifies the lag between these two times. The size of [an intervention] effect can differ both for different times in history and for different time lags" (Reichardt, 2006, p. 2). Again, a key internal-validity requirement of the MBD is the intervention's producing a *selective* effect, specifically an effect that results only when the intervention is applied to a recipient or unit at its personally targeted time period or start point. In the MBD across participants, the extent to which the intervention effect is similar for participants with differently staggered intervention start points helps to establish the intervention's ecological and *referent generality* external validity (Snow, 1974).

Finally, the size of an intervention effect is typically assessed across a number of preselected *outcome measures*. The size of the effect may be similar or different from one outcome measure to the next, depending on the measure's sensitivity, reliability, and construct validity (Cronbach & Meehl, 1955), as well as other factors. In the MBD application across behaviors (or more generally, across different outcome measures), either similar or different effects of the intervention can result. The extent to which an intervention produces similar effects across outcome measures helps to establish the referent generality external validity and *convergent validity* (Cronbach & Meehl, 1955) of the intervention. Conversely, the extent to which an intervention effect differs across outcome measures in prespecified ways can represent a positive discriminant validity credibility feature of the intervention (Levin, 1992).

Reichardt (2006, p. 2) labels these four categories the "size-of-effect factors" and advances the *principle of parallelism*, which states that "if a methodological option exists for any one of the four size-of-effect factors of the recipient, setting, time, and outcome variable, a parallel option exists for each of the other three factors as well." An important consideration based on Reichardt's work is that in an ideal comparison, an unambiguous determination of the intervention effect requires that everything other than the intervention remains the same at a given point in time. However, from a practical standpoint, the four study factors that Reichardt has specified (recipient, setting, time, and outcome measure) typically will vary with the intervention. Nevertheless, as Reichardt argued, it is possible—and desirable—to take these factors into account when interpreting size-of-effect outcome comparisons.

In particular, the prominent size-of-effect factor in the four types of comparison has a relationship to randomization, and each can vary either randomly or nonrandomly with the intervention conditions. That is, "[i]f the prominent size-of-effect factor varies randomly with the [intervention] conditions, the comparison is called a *randomized experiment*. If the prominent size-of-effect factor varies nonrandomly with the [intervention], the comparison is called a *nonrandomized experiment*" (Reichardt, 2006, p. 4). Reichardt further distinguished between two types of nonrandomized experiment: "those based on an explicit quantitative ordering and those not based on an explicit quantitative ordering" (p. 4). The 12 possible combinations of prominent size-of-effect factors (recipient, setting, time, and outcome measure) and type of experiment or assignment to intervention conditions (random; nonrandom, explicit quantitative ordering; and nonrandom, no explicit quantitative ordering) are presented in Table 2. There it can be observed that if the intervention conditions are assigned randomly to either recipients, times, settings, or outcome measures, then these comparisons represent a randomized experiment (referred to in the first column of Table 2). However, as Reichardt noted, varying an intervention randomly across a particular prominent size-of-effect factor will not necessarily guarantee that the intervention varies randomly across the other size-of-effect factors. For example, recipients may be randomly assigned to intervention conditions, but if there are nonrandom conditions-related differences on any of the other factors (times, settings, or outcome measures), the scientific integrity of the experiment will be compromised.

It may also be observed in Table 2 that Reichardt (2006) included interrupted time-series designs in the "nonrandom, explicit quantitative ordering" column. Reichardt (pp. 5–6) indicated that when time is a prominent size-of-effect factor, a time-series design is composed of days (or sessions) prior to and following some structured aspects of the design (e.g., an intervention). It is

Table 2
A Typology of Comparisons

Prominent size-of-effect factor	Assignment to intervention		
		Nonrandom	
	Random	Explicit quantitative ordering	No explicit quantitative ordering
Recipients	Randomized recipient design	Regression-discontinuity design	Nonequivalent recipient design
Times	Randomized time design	Interrupted time-series design	Nonequivalent time design
Settings	Randomized setting design	Discontinuity across settings design	Nonequivalent setting design
Outcome variables	Randomized outcome variable design	Discontinuity across outcome variables design	Nonequivalent variable design

Note. From "The Principle of Parallelism in the Design of Studies to Estimate Treatment Effects," by C. S. Reichardt, 2006, *Psychological Methods, 11*, p. 5. Reprinted with permission.

worth pointing out, however, that Reichardt's characterization of a "time-series design" is its most "basic" operationalization, namely, an interrupted time-series design (or what we have called here the AB design), with the intervention effect manifested as a break, discontinuity, or "interruption" in the pre- and postintervention regression lines at the point of intervention. Although Reichardt was correct in noting that interrupted time-series intervention designs—with their nonrandomized time comparisons based on an explicit quantitative—are most frequently implemented in the clinical and applied literature (e.g., behavior analysis research), we illustrate in the remainder of this article how randomization variations (in consideration of the four effect-size factors) can be incorporated into such designs to transform them into more scientifically credible randomized experiments.

Specifically, we will show how single-case interrupted time-series designs can be structured to satisfy one or more of the randomized experimental design criteria outlined in the first column of Table 2. In particular, and as is illustrated in the next section, both recipients/units and times can be structured randomly to allow a randomized trial to be conducted with the three classes of single-case design (within-, between-, and combined series). Other features of the design (i.e., settings and outcome measures) can and should remain constant so as not to compromise the internal validity of the study. For example, for the combined-series MBD across settings or behaviors to be regarded as a randomized experiment, the order or sequence in which the settings or behaviors are examined would need to be randomly determined.

Illustrations of Randomized Single-Case Intervention Designs

In this section, we illustrate how randomization can be implemented in single-case intervention designs with reference to the three design classes reviewed earlier: within-, between-, and combined-series design elements. We present a number of different randomization schemes (many of which are also described by Edgington, 1992, and Edgington & Onghena, 2007), followed by design-option variations and extensions within each of those schemes. For each design option, once again consideration must be given to the methodological units upon which randomization is based within the experiment (e.g., individuals, small groups, classrooms, schools), as well as to the other size-of-effect factors noted by Reichardt (2006). As was just mentioned, we will demonstrate that it is possible to structure single-case intervention designs in a way that satisfies the randomized experimental design criteria outlined in the first column of Table 2. In particular, any of the four size-of-effect factors can accommodate some appropriate type of randomization.

A few preliminary remarks are in order. First, we intend the following single-case intervention design randomization strategies to be representative, and not exhaustive, of those that could be implemented. Second, in the variations for which we provide randomization illustrations, we consider the possibility of applying some form of either *simple* or *blocked randomization*. Compared with nonrandomized single-case designs, both randomized variations are generally more scientifically credible. The blocked variations are even more methodologically sound and have the potential to provide statistical power advantages (Lall & Levin, 2004; Levin & Wampold, 1999), analogous to those associated with *randomized block designs* relative to *completely randomized designs* in the traditional experimental design literature (e.g., Keppel & Wickens, 2004; Kirk, 1995; Levin, 1997; Maxwell & Delaney, 2004). Third, several units-appropriate single-case statistical analysis strategies (not discussed here) have been developed on a design-by-design basis. Finally, the number of units in the design is represented by N, the number of within-conditions interventions (either intervention and control–baseline or alternative interventions) by c, the number of between-series intervention conditions (where applicable) by k, the number of time periods or sessions by t, and the number of potential intervention start points (where applicable) by s.

Within-Series Randomized Phase-Order Designs

Randomized ABAB . . . AB design. It will be remembered that a common single-case within-series design possesses an ABAB . . . AB design structure (see Figure 1). Incorporation of randomization into this design is relatively straightforward but requires that the series' condition replications (represented here in this design by sessions or time periods) are determined on a random basis (Edgington, 1992; Levin, Marascuilo, & Hubert, 1978; Onghena, 1992). This is analogous to the process of randomly assigning treatment–administration orders to participants in conventional group designs. The *randomized phase-order* within-series design illustrates an a priori randomized sequence in the ABAB . . . AB design with a single unit. For a simple randomization scheme, the A and B conditions (assumed in the following examples to be equally represented) are randomly assigned to the t time periods, with the consequence that there is no restriction in the number of A or B time periods that can follow one another consecutively. For a blocked randomization scheme, successive time periods are considered in blocks of two, with one A and one B session randomly assigned to each pair of time periods. This assignment procedure guarantees that conditions of the same type (A or B) cannot appear in more than two consecutive time periods. A hypothetical randomized single-unit ABAB . . . AB design is presented in Table 3.

The incorporation of randomization into within-series designs is relatively straightforward but adds some complexities to the research process. Depending on the experimental question and other practical realities of applied research, the following considerations need to be addressed: First, most behavioral and instructional intervention implementations require that an initial baseline or no-intervention (A) phase precedes the first intervention (B) phase (see also the following AID discussion). As a consequence, the randomization scheme and corresponding statistical analysis that are applied by the researcher must be adapted to satisfy that

Table 3
Randomized ABAB . . . AB Design With One Unit, Two Within-Series Conditions, and 10 Time Periods

Type of randomization	Design
Simple	ABBBABABAA
Blocked	ABABBAABBA

requirement. The randomized ABAB . . . AB design and its variations could be modified accordingly so that it commences with one or more A (baseline–warm-up–adaptation) phases, if that is deemed desirable or necessary by the researcher. In such cases, the experiment proper (i.e., the actual experimental sequence of randomized A and B phases) would follow an initial sequence of one or more mandatory A phases, with the stipulation that the researcher's interpretation of the results (and the statistical analysis of the data, when applicable) be modified to suit the experiment proper.[3]

As a published example of a traditional single-case ABAB intervention design with a true baseline condition, Thompson, Cotnoir-Bichelman, McKerchar, Tate, and Dancho (2007, Experiment 2) conducted a study in which two infants were taught, using delayed prompting and reinforcement techniques, to replace their typical crying and whining "demands" with simple infant-initiated signs. Replication of the effects occurred within the ABAB sequence, thereby meeting the criterion proposed by Horner et al. (2005). In accord with our immediately preceding discussion, the authors could have incorporated randomization into their design. Assuming that the researchers elected to include (a) an initial A phase (baseline, consisting of reinforcers provided whenever crying or whining occurred) before the signing intervention (B) was introduced and (b) two adjacent A and B phases, then the following randomized ABAB . . . AB design could have been constructed. First, require that one or more initial A periods be administered, represented here by A′. Then, randomly assign As and Bs until two adjacent A and B phases are produced. With a simple randomization scheme, the sequence might turn out to be something like A′BBABAAAB, whereas for a blocked randomization scheme, it might be A′BAABABAB. Either randomization scheme would likely have worked well in this experiment because the effect of the intervention (sign training) was relatively rapid, and short returns to baselines were achieved, thereby allowing more phase replications to occur. Other randomization possibilities could have been added to Thompson et al.'s (2007) research in that the authors had two infants available for each of the study's two experiments. In particular, and as is discussed in the following section on replicated randomized ABAB . . . AB design, in each experiment the two infants could have been assigned different random A and B phase sequences (see also Footnote 3).

Randomized ABCABC . . . ABC design. The use of randomization in the single-unit AB . . . AB design can be extended to the inclusion of three or more within-series conditions, such as when two interventions are compared with each other and with a baseline condition. In such cases, randomization could be applied to the three conditions in either a simple or blocked fashion. With the latter, the three conditions (A, B, and C) are randomized in blocks of three, which guarantees that a unit will be assigned no more than two consecutive time periods of the same intervention condition. This example is illustrated in Table 4.

Table 4
Randomized ABCABC . . . ABC Design With One Unit, Three Within-Series Conditions, and 15 Time Periods

Type of randomization	Design
Simple	ACABBBCCBCAACBA
Blocked	ACBCBAABCBACBAC

As a published example of a multiple-intervention design, Manuel, Sunseri, Olson, and Scolari (2007) conducted a study in which they developed interventions to increase students' selection of reusable dinnerware in a university cafeteria. Some 75 observation sessions were conducted, with an average of 251 students observed per session. The study consisted of nine phases, with A = baseline, and various intervention components involving B = increased counter space and signs promoting reusable tableware, C = environmental impact posters, D = employee prompting of cafeteria patrons to reuse dinnerware, and E = experimenter prompting with motivational signs. In particular, the authors implemented an ABA B + C AB B + D B + D + E B design to evaluate intervention effects with respect to two outcome measures (i.e., students' selection of reusable cups and reusable plates). Technically, the design did not adhere to the replication standards stated earlier. Nevertheless, our focus here is on the notion that the researchers could have incorporated a randomized order of intervention introduction into their design, insofar as the internal validity of the design would be strengthened by comparing the several within-series intervention conditions in an unbiased fashion (i.e., in a time-series sequence that is random, rather than one that is researcher determined). Moreover, because the study was conducted with a large number of students, the researchers might have been able to develop a replicated randomized intervention design (as discussed in the next section), in either its simple or blocked form, thereby increasing the internal validity of their experiment. As an aside, it is interesting to note that the authors invoked single-case methodology and terminology in presenting their results graphically, while at the same time calculating and reporting conventional effect sizes on the basis of outcome means and standard deviations (i.e., Cohen *d*s).

Replicated randomized ABAB . . . AB design. Another option with within-series designs is to apply randomization in a multiple-unit study. In this case, there is the same within-series randomization structure as was just illustrated except that more than one unit participates in the study (as was the case for the two infants in each of the earlier discussed Thompson et al., 2007, experiments). In Thompson et al.'s study, each unit could have been assigned randomly to its own within-series sequence of the design, either in a simple (unrestricted) or blocked (restricted) fashion. This replicated randomized within-series design could, for example, be structured as shown in Table 5. The design could encompass additional intervention conditions, units, or time periods, depending on the number of units available, the resources of the researcher, and research-specific logistical issues, among other factors. Generally speaking, the more replications that are included in the design (represented by either the number of units or the

[3] With these design modifications and stipulations in mind, the present form of randomization will likely prove to be more practicable in four-phase ABAB intervention designs when the A and B phases consist of two different intervention conditions than when they consist of a baseline and intervention condition, respectively. At the same time, the possibility of applying a different form of randomization for the ABAB design—namely, phase start-point randomization—has been incorporated into other within-series designs and will be a major topic discussed later in the article.

number of time periods), the more statistical power the researcher will have in the statistical analysis.[4]

Note that in these randomized ABAB . . . AB designs, the sequence in which conditions (A and B) are administered is randomized in either a simple or blocked fashion, in contrast to the systematically alternating administration of conditions in traditional ABAB . . . AB designs. Potential confounding of the intervention with Reichardt's (2006) time, settings, or outcome measures size-of-effect factors prevent the latter (traditional ABAB . . . AB designs) but not the former (randomized ABAB . . . AB designs) from being regarded as a randomized experiment. Similar comments apply to Reichardt's (p. 4) traditional within-subjects letter-pronunciation example, where for the two-condition experiment to be considered a scientifically valid "randomized" experiment, the specific letters would need to be presented in a random order (ideally, individually randomized on a unit-by-unit basis as well) to avoid a confounding with either time or outcome measures.

Randomized designs with a within-series factorial structure. Factorial within-series intervention designs can be constructed to answer research questions about both the separate and joint effects of different intervention types or interventions combined with other factors. As a published example of a single-case factorial design, Twardosz, Cataldo, and Risley (1974, Experiment 3) used three successive within-series designs (consisting of 19, 14, and 13 days, respectively) to investigate how two specific combinations of illumination (A_1 = light, A_2 = dark) and sound (B_1 = quiet, B_2 = noise) affected the sleeping patterns of 13 preschool children. In particular, the researchers were most interested in comparing a light, noisy combination with a dark, quiet combination, as shown in Table 6.

Twardosz et al. (1974) provided a rationale for investigating the two particular illumination–sound combinations that they did. If instead their interest had been in examining both the separate and joint contributions of illumination and sound, they could have adopted a randomized single-series 2 × 2 factorial design (Edgington, 1992). Suppose, for example, that 12 time periods are included. With a simple randomization scheme, an unrestricted sequence of an illumination condition (A_1 = light or A_2 = dark) combined with a sound conditions (B_1 = quiet or B_2 = noise) is allowed. In contrast, with a blocked randomization scheme, each possible combination of the A and B levels can be represented in four consecutive time periods, and so with 12 time periods, each AB combination would be represented three times. The design is illustrated in Table 7.

Table 5
Replicated Randomized ABAB . . . AB Design With Four Units, Two Within-Series Conditions, and 10 Time Periods

Type of randomization/design
Within-unit simple
Unit 1: ABBBABABAA
Unit 2: BBBAAAABAB
Unit 3: ABABABAABB
Unit 4: AABBBBAAAB
Within-unit blocked
Unit 1: ABBABABAAB
Unit 2: BABABAABAB
Unit 3: BAABABBAAB
Unit 4: ABABBAABBA

Table 6
Design of Twardosz et al.'s Illumination/Sound Study 3

	Day		
Child	1–19	20–33	34–46
1–13	A_1B_2	A_2B_1	A_1B_2

Note. N = 13 preschool children observed over 46 days. Table based on a study described in "Open Environment Design for Infant And Toddler Day Care" by S. Twardosz, M. F. Cataldo, & T. R. Risley, 1974, *Journal of Applied Behavior Analysis, 7,* 529–546.

Considerations in within-series randomized phase-order designs. Adoption of within-series designs presents two domains of challenges to researchers, namely (a) those associated with use of randomization in systematically structured designs and (b) those associated with a within-series replication specification (i.e., Horner et al.'s, 2005, replication criterion). The former issue was discussed previously in relation to intervention experiments that do and do not readily permit for randomized A (baseline) and B (intervention) phases and, in particular, with respect to the common requirement that such experiments generally must begin with a baseline phase. The latter issue is one that is well known in the research literature (e.g., Kazdin, 1982; Kratochwill, 1978) and requires use of outcome measures that reverse or return to baseline levels when the intervention is withdrawn during the return to baseline phase. For this reason, the researcher must preselect a design in which exposure to the intervention is not likely to cause such dramatic changes in behavior or acquisition of knowledge or skills that return to baseline phase levels would be difficult or unlikely. Also, there may be ethical considerations that must be addressed when an intervention is withdrawn, especially in cases in which participants or staff are at risk of injury or harm.

As was noted earlier, in behavior-change research contexts, the unit's partial or complete return to the baseline level (following changes attributable to the intervention) strengthens the credibility of imputing a causal connection between intervention and outcome. However, should such returns to baseline not occur, researchers may be limited in the number of within-series replications that they can schedule. That is, although the logic of the design depends on replication, there may be limited numbers of these replications obtained for an appropriate and sufficiently powerful statistical test to be conducted. Another potential problem is that the amount of time required to conduct the ideal intervention experiment may be excessive, given the desired number of replications that need to be included. Therefore, randomized within-series designs may work best when phases consist of short sessions consisting of the unit responding in a relatively rapid fashion to the intervention (as was the case in the Thompson et al., 2007, experiment and could be structured in the next design type to be presented).

[4] Greater statistical power with additional units or time periods will be achieved as long as the intervention effect size is assumed to remain constant across both of those factors (i.e., there is no interaction involving the intervention and either units or time periods). In the special class of single-case designs we present here for which the intervention start point is randomly determined, increased statistical power is also associated with a greater number of potential intervention start points (Lall & Levin., 2004).

Table 7
Randomized 2 × 2 Factorial Design With One Unit, Two Within-Series Levels of A (Illumination), Two Within-Series of B (Sound), and 12 Time Periods

Type of randomization	Design
Simple	A_1B_2 A_2B_2 A_2B_2 A_1B_1 A_1B_2 A_1B_1 A_2B_1 A_1B_2 A_2B_2 A_2B_1 A_2B_1 A_1B_1
Blocked	A_2B_1 A_2B_2 A_1B_1 A_1B_2 A_1B_1 A_2B_1 A_2B_2 A_1B_2 A_1B_2 A_2B_1 A_2B_2 A_1B_1

Between-Series Randomized Phase-Order Designs

Randomized alternating intervention designs. The earlier discussed AID is the most frequently adopted design in the between-series domain and will be the focus of our discussion here. In that design, the researcher begins with a baseline period to establish the performance level with which the intervention condition(s) will be compared (see Figure 2). It will be remembered that the AID involves a more rapid alternation of the phases relative to the within-series applications discussed earlier. With the incorporation of randomization into this design, the researcher basically follows the same procedures that were outlined for the randomized within-series ABAB . . . AB design, with one important difference: If A truly represents a baseline or control condition, then the time series must begin with an A phase (here, A′). That is, following an initially mandated baseline or control (A′) phase, the researcher then randomly assigns time periods of the design to compare the intervention (B) phase with A′.[5]

For example, consider an intervention-versus-baseline comparison in a randomized AID design with a single unit measured over 13 time periods (7 days consisting of seven morning and six afternoon sessions). In this experiment, a blocked randomization of the intervention could be applied either to just one dimension of the design (logically, time of day on a day-to-day basis) to produce a *partially counterbalanced* design or to both dimensions (across and within days) to produce a *completely counterbalanced* design. The latter double-blocking approach requires an even number of days excluding the initial A′ baseline session, and adopting that approach (as is illustrated in Table 8) would control for Reichardt's (2006) setting and outcome measure size-of-effect factor biases and would impart a scientifically credible status to the design.

Similar to the previously discussed randomized within-series applications that are extended to more than two-condition comparisons, the randomized AID (or SID) can be applied when the researcher wishes to compare three within-series conditions (either a baseline and two intervention conditions or three intervention conditions). In this case, randomization would be applied to the three conditions of the experiment (here, breakfast, lunch, and dinner sessions). With a double-blocking procedure, a multiple of 6 days (excluding the initial A′ baseline day) is required to produce a completely counterbalanced design, as is illustrated in Table 9. Note that two interventions can be compared relatively quickly within the structure of this design; that may circumvent our previously discussed concerns about extended replications in the within-series applications.

Table 8
Randomized Alternating Intervention Design With One Unit, Two Within-Series Conditions, and 13 Time Periods (Seven Mornings and Six Afternoons)

Type of randomization/ time period	Day 1	2	3	4	5	6	7
Simple							
Morning	A′	A	B	B	A	B	B
Afternoon		B	A	A	A	B	A
Blocked							
Morning	A′	B	A	A	B	B	A
Afternoon		A	B	B	A	A	B

Note. In the randomized versions of this design, the random sequencing of six A and six B time periods occurs following the required initial A′ time period.

Replicated randomized alternating intervention designs. It is also possible to extend the randomized AID (and SID) to replications across units, as was earlier illustrated with the randomized within-series designs. With this extension, the double-blocked randomization process (illustrated here) would be conducted on a unit-by-unit basis and would require an even number of days (excluding the initial A′ baseline day) to be completely counterbalanced. The extension of the randomized AID across units is illustrated in Table 10 with three units and a comparison of a baseline and intervention condition over 7 days, consisting of seven morning and six afternoon sessions.

As a published example of a replicated AID, Fisher, Kodak, and Moore (2007) compared three methods of teaching discriminations to two children with autism: trial and error (a control condition), least-to-most prompting, and identity matching. The trial-and-error method basically involved presenting the participant with a sample stimulus (a picture of someone) and asking him or her to point to the named person, with no feedback provided for correct and incorrect responses. In the least-to-most condition, correct responding resulted in access to a reward, whereas incorrect responding led to a modeled prompt with the correct picture. The identity-matching condition included the least-to-most condition procedures but also an additional task in which the experimenter held a picture that was identical to the correct comparison stimulus and then named the stimulus. In a 34-session AID format, the researchers compared the percentage of correct spoken-word-to-picture relations by children in the three intervention conditions. A visual–graphical analysis revealed that the identity-matching condition was superior to the two other conditions for both children. Although the authors did not describe their intervention-order assignment method in the Method section, an examination of the time-series graph reveals that (a) for the first child, what appears to be a partially block-randomized order was employed (i.e., three of the six possible block-randomized orders were systematically se-

[5] The same comments apply to the earlier discussed SID, for which the randomization rationale and procedures follow those presented here.

lected and randomized throughout), and (b) for the second child, either a haphazard or simple randomized order of the interventions was adopted, resulting in 12 trial-and-error, 10 least-to-most, and 12 identity-matching administrations over the study's 34 sessions.

Considerations in between-series randomized phase-order designs. In the AID and SID, there are some considerations that are an important part of the design and that randomization may not address. One such consideration is the design's potential for unwanted "carryover effects" from one intervention condition to the other. That is, an intervention introduced in one session may "carry over" to a control or alternative intervention in the next session, thereby making it likely that an experimental effect will be underestimated or, in some cases, even detected. Randomizing the order of intervention conditions will not necessarily eliminate this problem. Carryover effects might be assumed to be more prevalent in the AID and SID between-series designs (relative to many within-series ABAB . . . AB–type applications) in that with the former the different intervention conditions generally are administered in closer proximity and with more rapid switches from one intervention to the next. The foregoing statement need not apply to all between- and within-series applications. Within-series experiments also become increasingly complex to design and administer with three or more intervention conditions—as might be appreciated to some extent from our AID example with three different intervention conditions implemented at three different times of day over 6 days.

Combined-Series Randomized Phase-Order Designs

Randomized multiple-baseline design (across units, settings, or behaviors). The domain of combined-series single-case intervention designs comprises three design types. These consist of variations of the MBD and include replication across either participants (or other units), settings, or behaviors (or outcome measures). As in Figure 3, here we focus on an MBD application across participants or units, although the design can be similarly applied to multiple settings or multiple behaviors (but see our cautionary comments later). This MBD across units is the strongest application of the design (and arguably the strongest of all single-case designs discussed in this article) from an internal-validity perspective, as was indicated in our earlier discussion. The design requires replication across selected units, with repetition of the intervention sequentially introduced across phases of the design.

The most straightforward application of randomization in the MBD across units involves randomly assigning each unit to the design's staggered intervention start points (see, for example, Wampold & Worsham, 1986, who initially proposed the randomized MBD and an associated statistical analysis). That is, each unit begins its exposure to the sequential introduction of the intervention in a random order. It is interesting that traditional applications of the MBD make no mention of assigning the unit replicates randomly to the staggered sequences. Yet, doing so strengthens the design's internal validity and, with the Wampold–Worsham analysis, its statistical conclusion validity as well. Table 11 illustrates

Table 9
Randomized Alternating Intervention Design With One Unit, Three Within-Series Conditions, and 19 Time Periods (Seven Breakfasts, Six Lunches, and Six Dinners)

Type of randomization/ time period	Day 1	2	3	4	5	6	7
Simple							
Breakfast	A′	B	C	B	A	C	B
Lunch		C	A	C	B	A	A
Dinner		C	A	B	B	C	A
Blocked							
Breakfast	A′	C	A	B	A	B	C
Lunch		B	C	A	B	C	A
Dinner		A	B	C	C	A	B

Note. In the randomized versions of this design, the random sequencing of six A, six B, and six C time periods occurs following the required initial A′ time period.

Table 10
Randomized Alternating Intervention Design With Three Units, Two Within-Series Conditions, and 11 Time Periods (Seven Mornings and Six Afternoons)

Type of randomization	Day 1	2	3	4	5	6	7
Within-unit simple							
Unit 1							
Morning	A′	B	B	B	A	A	B
Afternoon		B	A	B	A	A	A
Unit 2							
Morning	A′	A	B	B	B	A	B
Afternoon		B	A	A	A	B	A
Unit 3							
Morning	A′	A	A	B	B	B	A
Afternoon		A	B	A	B	B	A
Within-unit blocked							
Unit 1							
Morning	A′	B	A	A	B	B	A
Afternoon		A	B	B	A	A	B
Unit 2							
Morning	A′	A	A	B	A	B	B
Afternoon		B	B	A	B	A	A
Unit 3							
Morning	A′	B	B	A	B	A	A
Afternoon		A	A	B	A	B	B

Note. In the randomized versions of this design, the random sequencing of six A and six B time periods occurs following the required initial A′ time period.

Table 11
Randomized Multiple-Baseline Design With Five Randomized Units, Two Within-Series Conditions, 10 Time Periods, and a Staggered Intervention Introduction of One Time Period

Unit no.	Design
3	AAABBBBBBB
5	AAAABBBBBB
2	AAAAABBBBB
4	AAAAAABBBB
1	AAAAAAABBB

Table 12
Randomized Multiple-Baseline Design With Four Randomized Settings or Behaviors, Two Within-Series Conditions, 13 Time Periods, and a Staggered Intervention Introduction of Two Time Periods

Setting or behavior no.	Design
2	AAABBBBBBBBBB
1	AAAAABBBBBBBB
4	AAAAAAABBBBBB
3	AAAAAAAAABBBB

the design for five units and 10 time periods, with the intervention start point randomly assigned to units, beginning at Time Period 4 with a staggered introduction of the intervention occurring one time period after each preceding one.

As was just indicated, the incorporation of randomization into the MBD across settings or behaviors (or other measures) is also straightforward and involves randomly assigning the settings or behaviors to the predetermined staggered intervention start points. This procedure is illustrated in Table 12 for four settings or behaviors and 13 time periods, with the intervention start point randomly assigned to settings or behaviors, beginning at Time Period 4 with a stagger of two time periods thereafter. As a published example of an MBD, Reeve, Reeve, Townsend, and Poulson (2007) introduced a multicomponent package to four children with autism to determine whether the children could acquire appropriate responses for helping adults on a variety of different tasks (e.g., locating objects, putting items away, carrying objects). Following a baseline phase in which the children exhibited no correct helping responses, the intervention was implemented in sequential MBD fashion for each of the four children. In this example, the children could have been randomly assigned to the four staggered intervention start points—and, in fact, they may have been, although the authors did not specify the nature of their assignment process.

Considerations in combined-series randomized phase-order designs. Application of randomization to combined series designs is straightforward with respect to assignment of units to intervention start points. Randomization applied to the across-units version of this design class looks much like a traditional group experiment in which units are assigned randomly to experimental conditions. In fact, the design can be structured analogously to a traditional group experiment with random assignment of participants to conditions (as was indicated for the Reeve et al., 2007, study). In this regard, the randomized MBD across units represents the strongest inference design in the class of combined-series designs and perhaps even in the entire class of single-case designs. The application of the randomized MBD across settings or behaviors within a single unit is not as strong as the randomized MBD across units, namely because the latter application provides multiple sources of intervention-efficacy evidence that are independent, whereas the former applications do not. Given this independence issue, the randomized MBD across units results in an experiment that, relative to its across-settings and across-measures counterparts, is both more scientifically credible and has greater population external validity because it includes a more solid replication component.

Randomized Phase Start-Point Designs

In addition to the preceding randomized phase-order designs that were just discussed, single-case intervention researchers can incorporate a different form of randomization into their experiments, namely one in which they randomly determine the specific time points at which the various phases of the time series begin. This novel form of randomization similarly enhances a single-case intervention study's scientific credibility. In this section, we discuss several design variations that capitalize on a randomized phase start-point scheme.

AB design. As we noted earlier in this article, the most basic form of the within-series design is the AB design in which the A phase is a baseline or pretest series and the B is the intervention phase. In this interrupted time-series design, the researcher can structure the study so that the point of intervention, or *intervention start point*, is determined on a random basis (Edgington, 1975, 1992). In Edgington's inventive incorporation of randomization into an interrupted time-series design, the researcher initially specifies the total number of A and B time periods from which data are to be collected (for example, 20). Then, a range of potential intervention start points is determined on the basis of the researcher's specification of the minimum number of A (baseline–control) and intervention (B) time periods that he or she considers to be acceptable for the study (e.g., 4 and 3, respectively). Given the present two specifications, this range would be between Time Periods 5 and 17 inclusive, resulting in 13 potential intervention start points. A single (actual) intervention start point is then randomly selected from the 13 start points available. The design, with the actual randomly selected intervention start point being Time Period 9, can be illustrated as in Table 13.

Edgington's (1975) randomized phase start-point approach not only serves to bestow a "true experiment" status on the traditional AB (interrupted time-series) design, but as Edgington showed, it also lends itself directly to a units-appropriate statistical analysis of the resulting outcomes (e.g., Edgington, 1992; Edgington & Onghena, 2007; Levin & Wampold, 1999). Moreover, in single-case intervention experiments containing more than just one A and one B phase (i.e., ABAB . . . AB designs), the logic of Edgington's basic approach can be extended in a straightforward fashion to

Table 13
AB Design With One Unit, Two Within-Series Conditions, 20 Time Periods, and 13 Potential Intervention Start Points

	Time period																			
Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Unit 1	A	A	A	A	A	A	A	A	B[a]	B	B	B	B	B	B	B	B	B	B	B

Note. Potential start points are between Time Periods 5 and 17 inclusive.
[a] Intervention start point was randomly selected.

Table 14
All Phase-Sequence Possibilities for the Four-Phase ABAB Design

Phase sequence	Randomization scheme	Effects that can and cannot be purely associated with a specific sequence		
		Initial intervention effect?	Return-to-baseline effect?	Replicated intervention effect?
ABAB	1, 2	Yes	Yes	Yes
ABBA	1, 2	Yes	Yes	No
AABB	1	Yes	No	No
BABA	1, 2	No	No	No
BAAB	1, 2	No	No	No
BBAA	1	No	No	No

Note. Effect assessments are based on a four-phase ABAB design with a true baseline condition. With a simple phase-sequence randomization scheme (1), all six phase sequences are possible; whereas with a blocked phase-sequence randomization scheme (2), only four phase sequences are possible.

encompass multiple randomized phase start points (Edgington & Onghena, 2007).

Special consideration of the randomized four-phase ABAB design. In an earlier section, we discussed a randomization scheme for the four-phase ABAB design, namely randomizing the order in which the four phases are administered (e.g., BAAB) in what we termed a randomized phase-sequence design. Although that approach is methodologically appropriate and practically defensible in situations where the A and B phases represent two alternative intervention conditions—including a "standard" (B) and a "new" (C) method or technique—we noted that in situations where the A phases consist of a true baseline condition, it would be difficult to justify not beginning the experiment with a baseline (A) phase. The rationale is analogous to that for a conventional pretetest–intervention–posttest design (i.e., Campbell & Stanley's, 1966; original O_1XO_2 design) where a pretest assessment (O_1) logically precedes a postintervention assessment (O_2). For related discussions, see Todman & Dugard (2001, pp. 202–205) and Edgington & Onghena (2007, pp. 239–240).

In addition, clinical or behavioral single-case researchers generally demand that in the ABAB design (represented here as $A_1B_1A_2B_2$), following a demonstration of an "initial intervention" effect (A_1B_1), both a "return-to-baseline" effect (B_1A_2) and a "replicated intervention" effect (A_2B_2) be shown. For researchers wishing to assess these three effects, not adhering to the universally applied ABAB sequence would be a source of consternation. As an example, if the sequence BBAA were to result from a random selection process, none of the desired effects could be assessed and then not in a "pure" uncontaminated fashion. For example, even though a return to baseline (BA) is nested within this sequence, the "return" is not really a return because there is no initial baseline phase. In addition, that effect would be subject to the influence of other extraneous variables, intervention carryover being a major one, along with the usual time-tied confounders such as Campbell and Stanley's (1966) *maturation* (e.g., practice and fatigue). Note that we are applying the term *pure* strictly in relation to the present single-case time-series experimental context. That is, although we recognize that other unwanted variables—intervention novelty, as a primary one—can seriously contaminate effect assessment in a conventional pretest–posttest design, the multiple-observations feature of a time-series experiment would help to mitigate such concerns here (e.g., Shadish et al., 2002).

A visual, and more complete, representation of the interpretative challenges associated with a randomized phase-sequence scheme for the four-phase ABAB design with a true baseline condition is presented in Table 14. There it may be seen that the three sequences that commence with an intervention phase (B) rather than a baseline phase (i.e., the three final phase-sequence possibilities in Table 14) do not permit a pure assessment of any of the three desired effects.

With the preceding rationale and concerns, then, it would appear that a fixed ABAB administration sequence is the "order of the day" and that any permutation of that sequence through randomization would be eschewed by many single-case researchers. For that reason, a multiple (here, three-part) randomized phase start-point model approach, as an extension of that discussed for the AB design, could be considered (see Edgington & Onghena, 2007). With this approach, before the $A_1B_1A_2B_2$ experiment begins, the researcher randomly determines (a) the initial intervention start point from a set of acceptable start points (i.e., the A_1 to B_1 phase-transition point), (b) the return-to-baseline start point from a similarly acceptable set (the B_1 to A_2 phase-transition point), and (c) the replicated intervention start point from an acceptable set (the A_2 to B_2 phase-transition point). This three-part randomized phase start-point approach meshes well methodologically with the Edgington (1975) original randomized phase start-point model, while at the same time responding to the "fixed intervention order" concern of many single-case researchers. Finally, as was true for the AB design, this ABAB randomized phase start-point approach also lends itself to a units-appropriate statistical analysis, the properties of which the present authors are currently exploring.[6]

Replicated AB design. The replicated AB design with randomized phase start points and its associated analysis were initially proposed by Marascuilo and Busk (1988) as an extension of Edgington's (1975) AB randomization model for a single experimental unit. In this extension, each unit's intervention start point is determined separately and on a random basis from the predesignated set of permissible start points. This design bears a resemblance to the previously discussed randomized MBD across units

[6] In Edgington and Onghena's (2007, pp. 245–246) randomization test for ABAB . . . AB-type designs, the mean of the combined A phases is compared with that of the combined B phases. That tests a different statistical hypothesis than those targeting the three individual effects that we have specified here as typically being of concern in clinical or behavioral single-case applications (viz., initial intervention effect, return-to-baseline effect, and replicated intervention effect).

Table 15
Replicated AB Design With Three Units, Two Within-Series Conditions, 20 Time Periods, and 13 Potential Intervention Start Points

	Time period																			
Unit no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	A	A	A	A	A	A	A	A	B[a]	B	B	B	B	B	B	B	B	B	B	B
2	A	A	A	A	A	A	A	A	A	A	A	A	A	B[a]	B	B	B	B	B	B
3	A	A	A	A	A	B[a]	B	B	B	B	B	B	B	B	B	B	B	B	B	B

Note. Potential intervention start points are between Time Periods 5 and 17 inclusive.
[a] Intervention start point for each unit was randomly selected.

(and see later discussion). Note, however, that in contrast to that design with its systematically staggered introduction of the intervention, in this replicated AB application, it would be possible for the intervention to be introduced in a less staggered (or even in an overlapping) fashion, which would weaken the internal and discriminant validity characteristics of the experiment. The design is illustrated in Table 15.

Multiple-baseline design. A more systematically structured version of Marascuilo and Busk's (1988) replicated AB design with separately randomized phase start points for each experimental unit is Koehler and Levin's (1998) "regulated randomization" MBD. That design allows for more start-point possibilities than the previously discussed MBD, while at the same time maintaining the systematically staggered intervention introduction that is not incorporated into the Marascuilo–Busk approach. In particular, and as we have been discussing for the Edgington class of designs, researchers add a second randomization component to that of the earlier discussed randomized MBD by pre-experimentally designating one or more potential intervention start points for each unit, with each actual start point determined on a random basis. With this second randomization component, both the Marascuilo–Busk and Koehler–Levin procedures may be regarded as between-units generalizations or replications of the randomized phase start-point model. This design is illustrated in a 15 time-period example of Table 16, where researcher and research exigencies led to three a priori designated potential intervention start points for Units 1 and 3 and two potential intervention start points for Units 2 and 4.[7] Actual research examples of this design's application, and its associated randomization statistical analysis, are provided by Koehler and Levin (1998) and McKie (1998).

Comparative AB design. Extending the just-discussed randomized phase start-point models, Levin and Wampold (1999) developed the comparative AB design and its associated analysis to accommodate a scientifically credible comparison of two different between-series interventions (or of an intervention and control condition) in a single-case AB design. For example, with one unit randomly assigned to Intervention X (B) and another unit to Intervention Y (C), the design allows for a direct between-unit comparison of the A–B and A–C changes produced by the two interventions, akin to the interaction of a between-subjects factor (e.g., treatment) and a within-subjects factor (e.g., time), as with a split-plot arrangement in the conventional experimental design literature (see, for example, Maxwell & Delaney, 2004). Levin and Wampold proposed two different intervention start-point randomization approaches, the independent intervention start-point and the simultaneous intervention start-point models (corresponding to our present simple and blocked randomization approaches, respectively). As we have previously discussed here for other AB-type designs, the latter (simultaneous intervention start-point) model has the methodological advantage of controlling for potential confounding from time or measures. The two models are illustrated in Table 17.

Levin and Wampold (1999) developed two different statistical tests to accompany these models, one called the test of the *general intervention effect* and the other the test of the *comparative intervention effect*. The former test assesses the within-series A–B change averaged across the two experimental units (equivalent to an intervention "main effect") and amounts to the same randomization test proposed by Marascuilo and Busk (1988). As was just noted, the latter test (which is associated with the comparative AB design discussed in this and the next section) compares the A–B changes in the two intervention conditions, X and Y.[8]

Replicated comparative AB design. The replicated version of the just-discussed comparative AB design with randomized phase start points and its associated analysis were initially proposed by Levin and Wampold (1999); this design is illustrated in Table 18. In addition, the simultaneous intervention model has recently been extended to allow for blocked random assignment of units (with a corresponding randomization statistical test) to more than two intervention conditions (Lall, Levin, & Kratochwill, 2009). Units may be blocked either arbitrarily (e.g., in order of their enlistment in the experiment) or on the basis of a relevant pre-experimental variable (e.g., a variable related to the outcome being measured in the experiment). As compared with a simple randomization scheme, greater experimental control is gained with blocked randomization.

Concluding Remarks and Limitations of Randomized Single-Case Intervention Designs

In this article, we have discussed how various interrupted time-series investigations (as represented here by single-case intervention designs) can be transformed from Reichardt's (2006) nonrandomized experiments to randomized experiments by incorporating randomization into the design structure. Each of the designs re-

[7] Although not illustrated here, this design also allows for any unit to be randomly assigned a single pre-designated intervention start point (Koehler & Levin, 1998).

[8] Elsewhere (Lall & Levin, 2004) is discussed a caveat associated with the statistical test of Levin and Wampold's independent start-point model's comparative intervention effect, which is not an issue in the simultaneous start-point model.

Table 16
Multiple-Baseline Design With Four Randomized Units, Two Within-Series Conditions, 15 Time Periods, 3, 3, 2, and 2 Potential Intervention Start Points for Units 1, 3, 2, and 4, Respectively, and a Staggered Intervention Introduction of at Least One Time Period

	Time period														
Unit no.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	A	A	A	**B**	**B**	**B**	B	B	B	B	B	B	B	B	B
3	A	A	A	A	A	A	**B**	**B**	**B**	B	B	B	B	B	B
2	A	A	A	A	A	A	A	A	A	**B**	**B**	B	B	B	B
4	A	A	A	A	A	A	A	A	A	A	A	**B**	**B**	B	B

Note. For each unit, the intervention start point is one of the randomly predetermined bolded **B**s.

quires various practical, logistical, and conceptual tradeoffs, depending on what type of randomization scheme is implemented in the design. Generally speaking, the simple randomization schemes presented here strengthen the internal validity characteristics of single-case designs relative to traditional nonrandomized single-case designs, whereas the blocked randomization schemes greatly strengthen those internal validity characteristics. Thus, on an internal-validity continuum, Reichardt's nonrandomized time-series experiments would anchor the weak side, and the present block-randomized time-series experiments (including block-randomized on multiple factors where applicable) would anchor the strong side.

In addition, the type of randomization scheme implemented dictates the kinds of units-appropriate statistical analyses that should be conducted. Even in the absence of a suitable statistical test, a single-case intervention design's inclusion of appropriate forms of randomization enhances the experiment's scientific credibility (as reflected by internal-validity considerations). Randomization therefore represents a highly recommended component of single-case and related time-series designs, designs in which randomization is not considered as a part of the traditional modus operandi (see also Levin at al., 2003).

There are, however, limitations associated with the use of randomization in single-case intervention designs. First, a randomized phase-order scheme does not guarantee that a given number of A versus B comparisons will occur within a fixed number of time periods. For example, in the traditional four-phase ABAB design, there is a guaranteed initial intervention versus baseline (B vs. A) comparison followed by a guaranteed second intervention versus baseline comparison. Thus, the design satisfies the replication standard prescribed by Horner et al. (2005) within a four time-period experiment. With both the simple and blocked phase-sequence randomization schemes, on the other hand, more (and perhaps many more) time periods might need to built into the experiment to produce two adjacent AB phases for an initial and a replication comparison. With reference to the 10 time-period ABAB . . . AB design of Table 3, for instance, there are five adjacent B versus A comparison opportunities in the traditional nonrandomized design but only three in the two randomized design variations that happened to be produced for the Table 3 sequence. For the two randomized phase-order variations, therefore, the cost of the research would necessarily increase in that additional time, measurement, and other resources might be required to conduct the experiment. However, it is possible for single-case researchers to overcome these concerns by adopting a different type of randomization scheme, namely, the randomized phase start-point model that was discussed here (see also Edgington & Onghena, 2007).

Second, one of the features of traditional single-case designs that is often regarded as central to their use is the flexibility accorded to the researcher concerning when to change phases, which includes (among other aspects) extending measurement until some degree of outcome stability has occurred (Hayes et al., 1999). That is, some (though not all) traditional single-case intervention studies may be likened to "design experiments" in the learning sciences field, where the design unfolds as candidate interventions are auditioned and data are forthcoming (see, for example, Levin et al., 2003). Randomization may eliminate some

Table 17
Comparative AB Design With Two Units Representing Two Different Between-Series Conditions (Intervention X and Intervention Y), Two Within-Series Conditions, 20 Time Periods, and 13 Potential Intervention Start Points

Type of start-point model	Time period																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Independent																				
Intervention X: Unit 1	A	A	A	A	A	A	A	A	B[a]	B	B	B	B	B	B	B	B	B	B	B
Intervention Y: Unit 2	A	A	A	A	A	B[a]	B	B	B	B	B	B	B	B	B	B	B	B	B	B
Simultaneous																				
Unit 1x	A	A	A	A	A	A	A	A	A	A	A	B[b]	B	B	B	B	B	B	B	B
Unit 1y	A	A	A	A	A	A	A	A	A	A	A	B[b]	B	B	B	B	B	B	B	B

Note. Potential intervention start points are between Time Periods 5 and 17 inclusive.
[a] Randomly selected intervention start point for each unit. [b] Randomly selected intervention start point for the pair of units.

Table 18
Replicated Comparative AB Design With Four Units Representing Two Different Between-Series Conditions (Intervention X and Intervention Y), Two Within-Series Conditions, 20 Time Periods, and 13 Potential Intervention Start Points

Type of start-point model	Time period																			
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Independent																				
Intervention X																				
Unit 1	A	A	A	A	A	A	A	A	B[a]	B	B	B	B	B	B	B	B	B	B	B
Unit 2	A	A	A	A	A	B[a]	B	B	B	B	B	B	B	B	B	B	B	B	B	B
Intervention Y																				
Unit 3	A	A	A	A	A	A	A	A	A	A	A	A	B[a]	B	B	B	B	B	B	B
Unit 4	A	A	A	A	A	B	B[a]	B	B	B	B	B	B	B	B	B	B	B	B	B
Simultaneous																				
Unit 1x	A	A	A	A	A	A	A	A	A	A	A	B[b]	B	B	B	B	B	B	B	B
Unit 1y	A	A	A	A	A	A	A	A	A	A	A	B[b]	B	B	B	B	B	B	B	B
Unit 2x	A	A	A	A	A	A	A	A	A	B[b]	B	B	B	B	B	B	B	B	B	B
Unit 2y	A	A	A	A	A	A	A	A	A	B[b]	B	B	B	B	B	B	B	B	B	B

Note. Potential intervention start points are between Time Periods 5 and 17 inclusive.
[a] Randomly selected intervention start point for each unit. [b] Randomly selected intervention start point for the pair of units.

of this phase-determination flexibility in that it is an a priori design consideration or specification. Although one cannot endorse a design-experiment format when conducting scientifically credible intervention experiments (see, for example, our previous discussion of Levin et al.'s stage model in Figure 2), we recognize that this feature might be advantageous in applied and clinical single-case research situations, often for the same reasons that it might not be possible or convenient to implement randomized group studies in those settings.

Third, randomized designs may limit the kinds of statistical analysis that are applied to the data, thereby possibly reducing the statistical conclusion validity of the study. The most direct and straightforward application of statistical tests to the specific designs we have featured in this article fall into the category of nonparametric randomization (or permutation) tests (e.g., Edgington & Onghena, 2007; Levin & Wampold, 1999; Todman & Dugard, 2001). However, because such tests require a certain sufficient number of phases to have adequate statistical power for detecting intervention effects, it is possible that single-case researchers will put these design-and-analysis concerns ahead of substantive, clinical, and practical ones—a case of the "tail wagging the dog."

Despite the potential issues and limitations in the design of randomized single-case experiments, we have offered researchers several options that can increase the validity of this class of design and thereby enhance the scientific credibility of intervention-research findings in applied and clinical fields. Ultimately, what these randomized single-case intervention designs have to offer will become evident in both the greater scientific credibility of research findings and the integrative summaries of our knowledge base for evidence-based interventions across areas of research in psychology and related fields.

References

Baldwin, S. A., Murray, D. M., & Shadish, W. R. (2005). Empirically supported treatments or Type I errors? Problems with the analysis of data from group-administered treatments. *Journal of Consulting and Clinical Psychology, 73,* 267–285.

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

Borckhardt, J. J., Nash, M. R., Murphy, M. D., Moore, M., Shaw, D., & O'Neil, P. (2008). Clinical practice as natural laboratory for psychotherapy research: A guide to case-based time-series analysis. *American Psychologist, 63,* 77–95.

Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal, 5,* 437–474.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait–multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

Cappella, E., Massetti, G. M., & Yampolsky, S. (2009). Rigorous, responsive, and responsible: Experimental designs in school intervention research. In L. M. Dinella (Ed.), *Conducting science-based psychology research in schools* (pp. 51–78). Washington, DC: American Psychological Association.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Edgington, E. S. (1975). Randomization tests for one-subject operant experiments. *Journal of Psychology, 90,* 57–68.

Edgington, E. S. (1992). Nonparametric tests for single-case experiments. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis* (pp. 133–157). Hillsdale, NJ: Erlbaum.

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

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

Fisher, W. W., Kodak, T., & Moore, J. W. (2007). Embedding an identity-matching task within a prompting hierarchy to facilitate acquisition of conditional discriminations in children with autism. *Journal of Applied Behavior Analysis, 40,* 489–499.

Gresham, F. M. (1997). Intervention integrity in single-subject research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 93–117). Mahwah, NJ: Erlbaum.

Hagermoser Sanetti, L., & Kratochwill, T. R. (2005). Treatment integrity assessment within a problem-solving model. In R. Brown-Chidsey (Ed.), *Assessment for intervention: A problem solving approach* (pp. 304–325). New York, NY: Guilford Press.

Hall, R. V., & Fox, R. G. (1977). Changing-criterion designs: An alternate applied behavior analysis procedure. In C. C. Etzel, J. M. LeBlanc, & D. M. Baer (Eds.), *New developments in behavioral research: Theory, method, and application.* Hillsdale, NJ: Erlbaum.

Hartmann, D. P., & Hall, R. V. (1976). A discussion of the changing criterion design. *Journal of Applied Behavioral Analysis, 9,* 527–532.

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

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

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

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

Kaestle, C. F. (1993). The awful reputation of education research. *Educational Researcher, 22,* 23–31.

Kazdin, A. E. (1977). Assessing the clinical or applied significance of behavior change through social validation. *Behavior Modification, 1,* 427–452.

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

Kazdin, A. E. (2004). Evidence-based treatments: Challenges and priorities for practice and research. In B. J. Burns and K. E. Hoagwood (Eds.). *Child and Adolescent Psychiatric Clinics of North America, 13,* 923–940.

Kazdin, A. E., & Hartmann, D. P. (1978). The simultaneous-treatment design. *Behavior Therapy, 8,* 682–693.

Keppel, G., & Wickens, T. D. (2004). *Design and analysis: A researcher's handbook* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove, CA: Brooks/Cole.

Koehler, M. J., & Levin, J. R. (1998). Regulated randomization: A potentially sharper analytical tool for the multiple-baseline design. *Psychological Methods, 3,* 206–217.

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

Kratochwill, T. R. (2007). Preparing psychologists for evidenced-based school practice: Lessons learned and challenges ahead. *American Psychologist, 62,* 829–843.

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

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

Lall, V. F., & Levin, J. R. (2004). An empirical investigation of the statistical properties of generalized single-case randomization tests. *Journal of School Psychology, 42,* 61–86.

Lall, V. F., Levin, J. R., & Kratochwill, T. R. (2009, April) *Additional investigations of a randomization test for assessing single-case intervention effects.* Paper presented at the annual meeting of the American Educational Research Association, San Diego, CA.

Levin, J. R. (1985). Some methodological and statistical "bugs" in research on children's learning. In M. Pressley & C. J. Brainerd (Eds.), *Cognitive learning and memory in children* (pp. 205–233). New York, NY: Springer–Verlag.

Levin, J. R. (1992). Single-case research design and analysis: Comments and concerns. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 213–224). Hillsdale, NJ: Erlbaum.

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

Levin, J. R. (1997). Overcoming feelings of powerlessness in "aging" researchers: A primer on statistical power in analysis of variance designs. *Psychology and Aging, 12,* 84–106.

Levin, J. R. (2005). Randomized classroom trials on trial. In G. D. Phye, D. H. Robinson, & J. R. Levin (Eds.), *Empirical methods for evaluating educational interventions* (pp. 3–27). San Diego, CA: Elsevier Academic Press.

Levin, J. R., Marascuilo, L. A., & Hubert, L. J. (1978). N = nonparametric randomization tests. In T. R. Kratochwill (Ed.), *Single subject research: Strategies for evaluating change* (pp. 167–196). New York, NY: Academic Press.

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

Levin, J. R., & O'Donnell, A. M. (1999). What to do about educational research's credibility gaps? *Issues in Education: Contributions from Educational Psychology, 5,* 177–229.

Levin, J. R., & Wampold, B. E. (1999). Generalized single-case randomization tests: Flexible analyses for a variety of situations. *School Psychology Quarterly, 14,* 59–93.

Manuel, J. C., Sunseri, M. A., Olson, R., & Scolari, M. (2007). A diagnostic approach to increase reusable dinnerware selection in a cafeteria. *Journal of Applied Behavior Analysis, 40,* 301–310.

Marascuilo, L. A., & Busk, P. L. (1988). Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behavioral Assessment, 10,* 1–28.

Max, L., & Onghena, P. (1999). Some issues in the statistical analysis of completely randomized and repeated measures designs for speech, language, and hearing research. *Journal of Speech, Language, and Hearing Research, 42,* 261–270.

Maxwell, S. E., & Delaney, H. D. (2004). Designing experiments and analyzing data: A model comparison perspective (2nd ed.). Mahwah, NJ: Erlbaum.

McKie, A. (1998). *Effectiveness of a neoprene hand splint on grasp in young children with cerebral palsy* (Unpublished master's thesis). University of Wisconsin, Madison.

Mosteller, F., & Boruch, R. (Eds.). (2002). *Evidence matters: Randomized trials in education research.* Washington, DC: Brookings Institute.

Onghena, P. (1992). Randomization tests for extensions and variation of ABAB single-case experimental designs: A rejoinder. *Behavioral Assessment, 14,* 153–171.

Onghena, P., & Edgington, E. S. (2005). Customization of pain treatments: Single-case design and analysis. *Clinical Journal of Pain, 21,* 56–68.

Reeve, S. A., Reeve, K. F., Townsend, D. B., & Poulson, C. L. (2007). Establishing a generalized repertoire of helping behavior in children with autism. *Journal of Applied Behavior Analysis, 40,* 123–136.

Reichardt, C. S. (2006). The principle of parallelism in the design of studies to estimate treatment effects. *Psychological Methods, 11,* 1–18.

Reyna, V. F. (2005). The *No Child Left Behind Act,* scientific research, and federal education policy: A view from Washington, DC. In G. D. Phye, D. H. Robinson, & J. R. Levin (Eds.), *Empirical methods for evaluating educational interventions* (pp. 29–52). San Diego, CA: Elsevier Academic Press.

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

Shavelson, R. J., & Towne, L. (2002). *Scientific research in education.* Washington, DC: National Academy Press.

Slavin, R. E. (2002). Evidence-based education policies: Transforming educational research and practice. *Educational Researcher, 31,* 15–21.

Snow, R. E. (1974). Representative and quasi-representative designs for research on teaching. *Review of Educational Research, 44,* 265–291.

Thompson, R. H., Cotnoir-Bichelman, N. M., McKerchar, P. M., Tate, T. L., & Dancho, K. A. (2007). Enhancing early communication through infant sign training. *Journal of Applied Behavior Analysis, 40,* 15–23.

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

Twardosz, S., Cataldo, M. F., & Risley, T. R. (1974). Open environment design for infant and toddler day care. *Journal of Applied Behavior Analysis, 7,* 529–546.

Wampold, B. E., & Worsham, N. L. (1986). Randomization tests for multiple-baseline designs. *Behavioral Assessment, 8,* 135–143.

Wolf, M. M. (1978). Social validity: The case of subjective measurement or how applied behavior analysis is finding its heart. *Journal of Applied Behavior Analysis, 11,* 203–214.

Received August 28, 2008
Revision received July 21, 2009
Accepted August 13, 2009 ■