Applying the Taxonomy of Validity Threats from Mainstream Research Design to Single-Case Experiments in Applied Behavior Analysis

Anna Ingeborg Petursdottir¹ · James E. Carr²

Published online: 20 September 2018
© Association for Behavior Analysis International 2018

Abstract

Mainstream research design in the social and behavioral sciences has often been conceptualized using a taxonomy of threats to experimental validity first articulated by Campbell and his colleagues (Campbell & Stanley, 1966; Cook & Campbell, 1979). The most recent update of this framework was published by Shadish, Cook, and Campbell (2002), in which the authors describe different types of validity and numerous threats to each primarily in terms of group-design experiments. In the present article, we apply Shadish et al.'s analysis of threats to internal, external, statistical conclusion, and construct validity to single-case experimental research as it is typically conducted in applied behavior analysis. In doing so, we hope to provide researchers and educators in the field with a translation of the validity-threats taxonomy into terms and considerations relevant to the design and interpretation of applied behavior-analytic research for the purposes of more careful research design and the ability to communicate our designs to individuals outside of behavior analysis, using their own vocabulary.

Keywords Single-case experimental designs · Validity · Validity threats

Behavior analysts do things differently than members of other sciences and professions who are concerned with human behavior. These differences include the primary emphasis on measurable behavior, a focus on the environment as the cause of behavior, the use of repeated measures of behavior, and the use of single cases as the basis of experimental research, among others. Because of the uniqueness of behavior-analytic methods, it is often necessary to explain them to others (e.g., members of other sciences and professions, funders, regulators). Having some awareness of the mainstream vocabulary in a particular domain can potentially aid in these communicative efforts.

With respect to behavior-analytic research, although the utility of single-case experimental designs has received increased appreciation in recent years (e.g., Tate et al., 2016), this relatively uncommon research approach is still frequently marginalized in comparison to other approaches, such as

randomized group designs (e.g., Cochrane Consumer Network, n.d.). The language of mainstream experimental design—including its logic and structure—is rooted in the taxonomy of validity threats originally proposed by Campbell and Stanley (1966), and later extended by Cook and Campbell (1979). This taxonomy is a widely used framework for designing and evaluating experimental research in the behavioral and social sciences. The most recent version of the validity-threats framework was published by Shadish et al. (2002). The authors present their work as a theory of generalized causal inference “grounded in the actual practice of scientists” (p. 24). Consistent with Cook and Campbell's (1979) earlier analyses, Shadish et al. (2002) discuss four types of validity that need to be considered when designing experiments and interpreting experimental data: *statistical conclusion validity*, *internal validity*, *construct validity*, and *external validity*. The distinction between validity types, however, differs somewhat from earlier versions, and their treatment of construct and external validity was substantially revised.

Unfortunately, considerations related to single-case experimental designs are not directly addressed by Shadish et al. (2002), whose taxonomy is still framed around group-design experiments. Neither are such considerations commonly addressed elsewhere in the literature, because even texts that are primarily devoted to single-case experimentation do not

✉ Anna Ingeborg Petursdottir
a.petursdottir@tcu.edu

¹ Department of Psychology, Texas Christian University, TCU Box 298920, Fort Worth, TX 76129, USA

² Behavior Analyst Certification Board, Littleton, CO, USA

typically make extensive use of the validity-threats framework in their discussion of experimental control and generality (e.g., Bailey & Burch, 2002; Barlow, Nock, & Hersen, 2009; Hayes, Barlow, & Nelson-Gray, 1999; Sidman, 1960). An exception may be found in Kazdin (2011), which contains an informative chapter describing numerous validity threats as they relate to experimentation in general, followed by a thoughtful analysis of the extent to which some of the major internal validity threats are of concern in the design of a single-case experiment. We should note, however, that Kazdin (2011) has renamed many of the validity threats, which might limit communication with mainstream researchers about single-case designs.

In the present article, we apply Shadish et al.'s (2002) revised framework to experimental research in applied behavior analysis.¹ We believe that this type of information is of value to researchers and educators in applied behavior analysis and thus seek to extend it to the entire list of validity threats as it appears in its most recent version (Shadish et al., 2002). We believe that this translational effort is useful for at least three reasons. First, educators who are teaching advanced behavior-analytic research methods can make use of the validity-threats taxonomy to illustrate how the proper use of single-case designs and execution of a line of research can satisfy mainstream experimental research concerns (functionally, if not structurally). Second, behavior-analytic researchers or practitioners who need to communicate with others in an effort to legitimize single-case experimental research might find the current analysis useful as a resource. Third, systematically analyzing single-case designs using the dominant experimental-analytic framework may also be of value in efforts to disseminate knowledge of single-case designs outside of the behavioral literature (e.g., Dixon, 2002; Morgan & Morgan, 2001).

In the following sections, we describe how each of the threats to the four types of validity applies to single-case experimental research. We emphasize single-case experimentation as it is most commonly practiced in applied behavior analysis—that is, in conjunction with direct observation and

methods of data analysis via visual inspection. When appropriate, we reference specific classes of single-case designs, in particular, reversal designs, multiple-baseline designs, and alternating-treatments designs. Table 1 presents an overview of all of the validity threats described by Shadish et al. (2002), as well as our analysis of the extent to which each is of relevance to the design and interpretation of a single-case experiment.

Four Types of Validity

Statistical Conclusion Validity

Shadish et al. (2002) define statistical conclusion validity as the validity of the conclusion that the dependent variable covaries with the independent variable, as well as that of any conclusions regarding the degree of their covariation. In other words, it concerns the question of whether an effect was observed in the experiment, regardless of how that effect was produced. Group-design experiments typically employ inferential statistics in order to answer this question, and thus a part of Shadish et al.'s discussion focuses on the proper use of statistical tests in order to enhance statistical conclusion validity. Researchers in applied behavior analysis, by contrast, primarily utilize visual inspection and descriptive statistics in order to determine whether the target behavior covaries with treatment. Numerous statistical tests are available for analyzing data from single-case designs (Gorman & Allison, 1997), and some behavior analysts have advocated their use (e.g., Crosbie, 1999). However, the usefulness of inferential statistics in behavior analysis has long been questioned (Michael, 1974; Sidman, 1960) and continues to be debated (Branch, 1999; Hopkins, Cole, & Mason, 1998; Johnston & Pennypacker, 2009; Parsonson & Baer, 1992; Perone, 1999). At present, the use of inferential statistics appears to be infrequent in the behavior-analytic literature (Hopkins et al., 1998). We therefore discuss statistical conclusion validity primarily as it concerns the validity of conclusions drawn via visual inspection.

Shadish et al. (2002) list nine threats to statistical conclusion validity: (a) *low statistical power*, (b) *violated assumptions of statistical tests*, (c) *“fishing” and the error rate problem*, (d) *unreliability of measures*, (e) *restriction of range*, (f) *unreliability of treatment implementation*, (g) *extraneous variance in the experimental setting*, (h) *heterogeneity of units*, and (i) *inaccurate effect-size estimation*. The first three threats apply mainly to situations in which inferential statistics are used to determine the presence or absence of an effect and are not easily applicable to other methods of data analysis. In addition, the threat resulting from inaccurate effect-size estimation applies mainly to situations in which the researcher wishes to make a statement about the size of the observed

¹ Shadish et al. (2002) discuss validity threats exclusively in the context of group-design experiments, whereas in applied behavior analysis, single-case experiments predominate. As the authors note, a single-case experimental design may be considered an instance of an interrupted time-series design applied to one individual. Many of the considerations that they describe in the context of time-series designs thus apply to single-case designs as well. However, the authors discuss interrupted time-series designs primarily as they have been applied to large numbers of individuals at the level of large-scale community interventions, often retrospectively and in conjunction with archival data-collection methods. In applied behavior analysis, in contrast, the use of single-case designs is most frequently combined with direct-observation methods of data collection and continuous analysis of data. Further, the time-series designs described by Shadish et al. do not include designs analogous to the alternating-treatments design (i.e., multielement designs; Barlow & Hayes, 1979), which differs in logic from other common single-case designs frequently used in behavior analysis.

Table 1 Validity threats listed by Shadish et al. (2002) and their relevance to the design and interpretation of single-case experiments utilizing direct-observation and visual-inspection methods

Validity type	High potential relevance	Low or limited relevance
Statistical conclusion validity	<ul style="list-style-type: none"> • Unreliability of measures • Restriction of range • Unreliability of treatment implementation • Extraneous variance in the experimental setting 	<ul style="list-style-type: none"> • Low statistical power • Violated assumptions of statistical tests • “Fishing” and the error rate problem • Heterogeneity of units • Inaccurate effect-size estimation
Internal validity	<ul style="list-style-type: none"> • History • Maturation • Testing • Instrumentation • Additive and interactive effects 	<ul style="list-style-type: none"> • Ambiguous temporal precedence • Selection • Regression artifacts • Attrition
Construct validity	<ul style="list-style-type: none"> • Inadequate explication of constructs • Construct confounding • Mono-operation bias • Mono-method bias • Confounding constructs with levels of constructs • Reactivity to the experimental situation • Novelty and disruption effects • Experimenter expectancies • Treatment diffusion 	<ul style="list-style-type: none"> • Treatment-sensitive factorial structure • Reactive self-report changes • Compensatory equalization • Compensatory rivalry • Resentful demoralization
External validity	<ul style="list-style-type: none"> • Interaction of the causal relationship with units (narrow to broad) • Interaction of the causal relationship over treatment variations (narrow to broad) • Interaction of the causal relationship with outcomes (narrow to broad) • Interaction of the causal relationship with settings (narrow to broad) • Context-dependent mediation 	<ul style="list-style-type: none"> • Interaction of the causal relationship with units (broad to narrow) • Interaction of the causal relationship over treatment variations (broad to narrow) • Interaction of the causal relationship with outcomes (broad to narrow) • Interaction of the causal relationship with settings (broad to narrow)

effect. Effect-size estimates occasionally appear in the applied behavior analysis literature (e.g., Ryan & Hemmes, 2005), and their use has been recommended for the purposes of conducting meta-analyses of single-case experiments (Busk & Serlin, 1992). However, they are at present infrequently reported and thus will not be discussed here. The remaining threats relate to measurement rather than statistical tests and thus apply to single-case research in much the same manner that they apply to group designs. Those threats are discussed in the following sections.

Unreliability of Measures Shadish et al. (2002) point out that unreliable measures of either the dependent or the independent variable affect the researcher’s ability to detect an effect. The authors discuss this problem mainly in the context of its effects on the outcomes of statistical analyses, such as analyses of covariance. However, in single-case experiments accompanied by visual inspection analysis, error variance due to unreliability of measures may affect statistical conclusion validity by obscuring changes in level, trend, and variability that control decision making by the visual inspector. Researchers in applied behavior analysis typically place a strong emphasis on preventing and detecting this threat by providing

unambiguous operational definitions to observers and assessing interobserver agreement continuously throughout the study.

Restriction of Range According to Shadish et al. (2002), statistical conclusion validity may be threatened when the range of possible values of either the dependent or the independent variable is restricted, as this may prevent the detection of an effect. Dependent variables may be restricted by either floor effects or ceiling effects. In single-case experiments that utilize direct-observation methods, dependent measures often consist of the frequency, rate, duration, latency, or accuracy of responding. Any of these dimensions may be subject to ceiling or floor effects—for example, if a person is physically incapable of emitting a response at higher rates or lower latencies than in baseline, if baseline rates are very low for behavior targeted for reduction, or if baseline accuracy approaches 100% for behavior that is targeted for acquisition. Such problems may be prevented through preexperimental observation and careful selection of the target behavior. Restriction of range may also be related to the scaling of the dependent variable; for example, the plotting of a dichotomous measure on the y-axis of a graph may hinder the detection of an effect via visual inspection (e.g., Charlop-Christy &

Daneshvar, 2003). Some authors have addressed this potential limitation by switching from a traditional line graph to a cumulative record to minimize the variability associated with dichotomous measures (e.g., Lechago, Carr, Grow, Love, & Almason, 2010).

With respect to the independent variable, restriction of range may occur, according to Shadish et al. (2002), when two treatments that are being compared are highly similar and thus appear to produce similar effects when typical methods of data analysis are used. In single-case experiments, this consideration may be of particular importance in alternating-treatments designs, in which two or more treatments are considered to produce differential effects if the data paths depicting responding under each treatment condition are visually distinct. If both treatments result in either rapid elimination of behavior targeted for reduction or rapid mastery of target behavior, data paths may not be visually separate even when the two treatments are differentially effective. In order to detect differential effects, it may sometimes be possible to conduct more sensitive data analyses by plotting within-session data instead of session-by-session summaries (e.g., Vollmer, Iwata, Zarcone, Smith, & Mazaleski, 1993), although differential effects detected by this method may not always be of clinical importance. If differential effects are still not detected, the primary recommendation made by Shadish et al. (2002) is to alter treatment parameters in order to reduce the similarity of treatments. As an example, researchers conducting a parametric study on the effects of reinforcer magnitude might select very distinct magnitudes for study. Kazdin (1982) referred to this tactic as “functional manipulation.”

Unreliability of Treatment Implementation Shadish et al. (2002) speak of unreliability of treatment implementation if implementation differs from one site to another or from one person to another within a site. In single-case research, if different persons implement treatment in different sessions, variations in treatment implementation will introduce variability in the treatment phase, which may obscure the treatment effect. Such problems may be best avoided by thorough training of behavior-change agents and continuous assessment of treatment integrity (Peterson, Homer, & Wonderlich, 1982; Vollmer, Sloman, & St. Peter Pipkin, 2008).

Extraneous Variance in the Experimental Setting As with unreliability of measures and treatment implementation, any uncontrolled factors in the experimental setting may produce error variance that interferes with the researcher’s ability to detect an effect. In behavior analysis, the traditional approach to preventing and eliminating such problems is to attempt to identify and control unwanted sources of variability (Sidman, 1960). In applied research, many sources of variability may be outside of the control of the researcher. However, steps that

can be taken to reduce extraneous variance include conducting sessions at the same time every day in the same physical environment and in the presence of the same individuals. If responding is still variable, continuing data collection until the range of variability is constant within a phase will help the researcher determine whether the pattern of responding observed in a subsequent phase differs from that predicted by the previous phase (Johnston & Pennypacker, 2009). Also, Shadish et al.’s (2002) recommendation to measure and include in data analysis potential sources of variance that are outside of the experimenter’s control applies to single-case designs, as well as to other research.

Heterogeneity of Units A final source of unwanted variability considered by Shadish et al. (2002) is heterogeneity of units, which in most cases refers to the heterogeneity of participants whose data are aggregated in a group design. In behavior analysis, single-case designs are sometimes applied to the collective performance of a defined group of persons (e.g., Conyers et al., 2004; Ryan, Ormond, Imwold, & Rotunda, 2002). However, because each data point represents aggregate performance, the diversity of individual responding within the group does not enter into the analysis of change following treatment onset. Thus, heterogeneity of units may be a relevant threat only if the composition of the group changes from one observation period to another. Changes in group composition may occur with community interventions when participants are unknown to the experimenter—for example, drivers traveling by particular intersections (e.g., Van Houten & Malenfant, 2004), or in settings such as classrooms or organizations, in which data may be influenced from 1 day to another by absences of part of the group (e.g., Bateman & Ludwig, 2003). Such influences may best be minimized by ensuring that data are collected under similar conditions each time—for example, at the same time of the day every day and under similar conditions.²

Internal Validity

Shadish et al. (2002) describe internal validity as concerning “inferences about whether observed covariation between A and B reflects a causal relationship from A to B in the form in which the variables were manipulated or measured” (p. 53). They consider eight variables that may threaten the validity of such an inference: (a) *ambiguous temporal precedence*, (b)

² As discussed by Shadish et al. (2002), the term “units” does not necessarily refer to persons but may also refer, for example, to observation times, which may be a relevant application to single-case designs. Heterogeneity of units may then be a threat to statistical conclusion validity anytime variable observation times introduce variability into the data. However, insofar as such variability is considered to be a function of uncontrolled extraneous factors, the presence of which is correlated with, for example, the time of day or day of the week, there may be no need to discuss it separately.

selection, (c) history, (d) maturation, (e) regression artifacts, (f) attrition, (g) testing, and (h) instrumentation.

Ambiguous Temporal Precedence Shadish et al. (2002) add this threat to the list of internal validity threats covered in previous work (Campbell & Stanley, 1966; Cook & Campbell, 1979) because a causal inference necessitates that the presumed cause precedes the presumed effect. According to Shadish et al. (2002), ambiguity of precedence is mainly of concern in correlational studies, as opposed to experiments in which the presumed cause is deliberately manipulated and, thus, its onset is known and controlled by the experimenter. In a single-case design, temporal precedence could become ambiguous if the design were applied to study the effects of naturally occurring events. However, with their reliance on frequent and repeated measures, single-case designs appear to be in a unique position to detect the possibility of reverse temporal precedence, which should result in a baseline trend or change in level, as long as the measures are frequent enough.

Selection Shadish et al. (2002) define selection as the possibility that preexisting differences between groups of participants exposed to different conditions account for an observed effect. As such, selection is not of concern in single-case experiments, as comparisons are made within, rather than between, participants (Kazdin, 2011).

History Shadish et al. (2002) define history as “all events that occur between the beginning of the treatment and the posttest that could have produced the observed outcome in the absence of that treatment” (p. 56). In single-case experiments, in which the dependent variable is measured multiple times during each condition, such events may occur between any two measurements and will generally appear as variability within phases, which is not a problem if phases are conducted until the data are stable. However, history may be a plausible explanation of a change in the dependent variable that is delayed from treatment onset, rather than occurring immediately. In addition, the possibility exists that an unusually influential event may coincide with a phase change, in which case history remains a potential threat to internal validity. A single-case design consisting of only one baseline and one treatment phase (i.e., an A-B design) therefore does not effectively rule out history as a threat to internal validity. In single-case experimentation, the primary strategy employed to rule out history effects is to demonstrate repeatedly that a change in the independent variable is followed by a change in the dependent variable. The more demonstrations, the less likely it is that each condition change was accompanied by an influential extraneous event. In reversal designs, history thus becomes an implausible explanation of an initial effect if behavior returns to baseline level following withdrawal of treatment, and yet more

plausible if reintroduction of treatment again results in behavior change. In multiple-baseline designs, history is ruled out if a treatment effect is replicated across participants, behaviors, or settings, whereas no change is observed prior to treatment implementation. It should be noted, however, that in multiple-baseline designs across participants, history in the form of shared environmental variables is effectively ruled out only when data are collected concurrently for all participants (Harris & Jenson, 1985). Finally, in alternating-treatments designs, history effects should be apparent in all conditions. History is therefore unlikely to be responsible for a consistent difference between two or more conditions when the data paths that represent them are visually distinct.

Maturation Maturation as a threat to internal validity is similar to history, except that rather than consisting of discrete environmental events, it consists of changes in participant behavior that occur gradually over time, even in the absence of treatment (Shadish et al., 2002). In single-case designs, the behavior of an individual is measured repeatedly over time, and as a result the data may be susceptible to maturational influences. The likelihood of such influences may differ depending on participant characteristics; for example, it may be higher with infants than with adults. When maturational influences occur, they should generally be gradual and evident as trends within phases (Kazdin, 1982), which should not be a problem as long as conditions are conducted to stability. However, the possibility exists that the onset of maturational influences may coincide with the beginning of treatment. Most likely, this would be indicated by a gradual rather than a sudden change in the data, but it is also possible that gradual changes may appear sudden on a graph. For example, a trend may be obscured if a longer-than-usual time elapses between the last observation before a phase change and the first observation after a phase change. Also, some maturational influences might actually have an abrupt rather than a gradual onset (Campbell & Stanley, 1966). In single-case experiments, therefore, detecting or ruling out maturational influences requires similar design considerations as those applied against history threats.

Regression Artifacts Statistical regression, as originally defined by Campbell and Stanley (1966), refers to the tendency for extreme scores on one observation to be closer to the mean on the next observation, because extreme scores are likely to contain greater measurement error (considered to be random) than scores closer to the mean. Statistical regression may threaten the internal validity of an experiment by mimicking a treatment effect, especially if the researcher selects participants for study based on their initial extreme scores on the dependent variable—for example, participants exhibiting severe depressive behavior. In single-case designs, it is possible that individuals are sometimes selected for study based on

extreme scores on a single measurement of the dependent variable. However, because the dependent variable is subsequently measured multiple times, statistical regression as originally described by Campbell and Stanley is generally not a threat to internal validity. The effects on the dependent variable, if any, should be evident in baseline as unusually high or low data points that are statistically likely to be followed by data points closer to the mean level. That is, regression toward the mean may contribute to within-phase variability but cannot be conceived of as a plausible explanation of a marked change in level across phases (Kazdin, 2011).

Shadish et al. (2002) also use the term *regression* to refer to *spontaneous remission*, a phenomenon unrelated to measurement error. Spontaneous remission may threaten internal validity when a client seeks treatment under particularly distressing circumstances, with behavior returning to the individual's mean level over time, regardless of whether treatment is implemented. Such changes probably tend to occur as a result of history and/or maturational influences (Kazdin, 2003) and thus require similar design considerations as history and maturation.

Attrition Attrition refers to the loss of participants during an experiment, which may threaten the internal validity of a group-design experiment if it results in differences between participants exposed to different conditions (Shadish et al., 2002). When single-case designs are used to evaluate the behavior of individuals, attrition is not typically relevant as an internal validity threat. However, it may be of concern when single-case designs are applied to the collective performance of a group (Kazdin, 2011)—for example, if a program designed to improve productivity in a corporate setting were applied to two different departments in a multiple-baseline design. An obtained effect might then be a result of differential attrition of high or low performers in each department. The differential attrition might result directly from the implementation of treatment and thus coincide with it; for example, low performers might drop out following the implementation of incentives contingent on high performance. Researchers using single-case designs in this manner should provide data on attrition in order to convincingly demonstrate that attrition has not affected their results.

Testing Shadish et al. (2002) define testing effects as the influence of exposure to a test or observation upon performance on a subsequent test or during subsequent observations. Single-case experiments necessarily involve repeated measurement of the same behavior, and thus testing is of possible concern. However, when the dependent variable is measured multiple times, testing effects should generally be evident early on (Kazdin, 1982). A stable baseline should therefore be sufficient to rule out testing influences. However, if it is the case that any specific amount of assessment can result in a

testing effect (Hayes et al., 1999), it may coincide with a phase change at any time, thus requiring similar design considerations as other internal validity threats, such as history.

Instrumentation Instrumentation refers to changes in the measurement system during an experiment that may be responsible for an observed effect (Shadish et al., 2002). In single-case designs, instrumentation is of concern insofar as measurement changes may coincide with phase changes. This threat is best controlled at the data-collection level by ensuring that no significant changes are made to the measurement system during the study. However, the use of human observers as part of a direct-observation data-collection system may result in unintended changes. *Observer drift* refers to subtle changes over time in observers' application of the operational definition (Kazdin, 2011), and *observer bias* consists of systematic errors made as a result of the observer's expectancies or prejudices (Barlow et al., 2009). Observer drift is unlikely to coincide with a change in the independent variable more than once during the course of an experiment and therefore may be ruled out through the proper use of reversal designs, multiple-baseline designs, and alternating-treatments designs. It can also be reduced by periodically retraining and recalibrating observers during the study (Barlow et al., 2009; Kazdin, 2011). Observer bias, on the other hand, is more likely to coincide with phase changes than to occur at other times and cannot as such be ruled out at the design level. Using observers who are blind to the experimental condition and/or direction of expected behavior change can minimize bias; however, the use of blind observers is not always practical or possible. Other strategies to prevent bias include educating observers about bias and providing precise operational definitions (Barlow et al., 2009). It is also important to avoid feedback to data collectors (O'Leary, Kent, & Kanowitz, 1975). Finally, data-collection systems that employ human observers should always incorporate assessment of interobserver agreement.

Additive and Interactive Effects The final class of threats to internal validity considered by Shadish et al. (2002) consists of the additive and/or interactive effects of one or more of the previously listed internal validity effects. Shadish et al. focus their discussion on the interaction of selection with other threats, such as history, maturation, and instrumentation. Because selection with regard to participants is not relevant to single-case experimentation, neither are its interactions with other threats. However, it has sometimes been argued that because observation times rather than persons may be considered the "units" in a single-case experiment, random assignment of observation times to conditions serves the same function as random assignment of participants to conditions in a group design and is necessary to rule out the interaction of selection with threats such as history and maturation

(Edgington, 1980, 1996). Randomized single-case designs are, however, rare in applied behavior analysis, except for alternating-treatments designs. Attempting randomization with other single-case designs may present practical problems, along with generally compromising the goals and logic of single-case experimentation (Kazdin, 1980).

Apart from selection, the combined influence of any two or more internal validity threats could potentially produce an effect that might be mistakenly attributed to treatment in the absence of proper experimental design. This should be kept in mind when designing and interpreting single-case experiments, as well as other experiments.

Construct Validity

Whereas statistical conclusion and internal validity address the issue of experimental control, construct and external validity are concerned with the generality of results. Following the tradition of Cronbach (1982), Shadish et al. (2002) consider four aspects of a study about which one may wish to generalize: persons (or units), treatments, observations, and settings. Construct validity is described as “the validity of inferences about the higher order constructs that represent sampling particulars” (p. 38), and threats to this type of validity are said to “concern the match between study operations and the constructs used to describe those operations” (p. 72). In other words, construct validity concerns the extent to which the researchers’ conceptualization of their participants, settings, treatments, and dependent variables captures the critical features of the experiment that were responsible for an observed effect. If this is not the case, the results may be impossible to replicate and/or they may not generalize from laboratory to real-world situations. Shadish et al. (2002) list a total of 14 potential threats to construct validity. In the following sections, we have divided these threats into three categories. The first category contains threats related to the adequacy of operational definitions. The second concerns specific aspects of the experimental situation or measurement system that may interact with the treatment to produce an effect, and the third concerns the behavior of individuals exposed to control conditions that may either obscure or exaggerate a treatment effect.

Adequacy of Operational Definitions Shadish et al. (2002) list five threats to construct validity that may be considered to result from inadequate operationalization of the persons, treatments, observations, or settings about which inferences are to be drawn: (a) *inadequate explication of constructs*, (b) *construct confounding*, (c) *mono-operation bias*, (d) *mono-method bias*, and (e) *confounding constructs with levels of constructs*.

First, *inadequate explication of constructs* refers primarily to the researchers’ failure to conceptualize important variables

or communicate specific details of their method. The operational definition of the dependent variable (e.g., aggression) may not include all observed topographies of aggressive behavior, or it may not explicitly describe all topographies that were in fact counted as instances of aggression. Similarly, the researchers may fail to communicate or consider relevant details of their treatment that may be critical for its success or may fail to describe critical characteristics of participants or settings.

The second threat, *construct confounding*, is similar to the first, except that it refers specifically to situations in which the researchers identify a specific feature of the experiment as critical, when in fact a different but correlated feature is the critical one. With respect to treatment, construct confounding may occur anytime an experimental manipulation introduces more than one potential source of influence, or two manipulations are introduced at the same time. Other aspects of the experiment (i.e., observations, persons, and settings) may also be susceptible to construct confounding. In behavior-analytic research, confounding of observations may be most likely to occur when the target behavior is measured by means other than direct observation. For example, when measuring permanent products of behavior, construct confounding may occur if the product can be produced via means other than the behavior. Similar problems may occur with behavioral data that are self-recorded by participants. With respect to participants and settings, construct confounding is of concern primarily when the intent is to compare the effects of a treatment across different types of participants or different types of settings. Problems of interpretation then arise if the participants or settings differ from one another in ways other than those considered critical by the researchers. Although single-case experiments are typically not designed for such comparisons, this type of construct confounding may arise when failures to replicate are interpreted in terms of participant, setting, or measurement variables that are in fact not the ones responsible for the different outcomes.

Mono-operation bias refers to limitations that may arise when a variable of interest is operationalized only one way. According to Shadish et al. (2002), any single operational definition is imperfect in that it may contain irrelevancies and/or fail to capture all relevant instances. The use of multiple operationalizations may thus increase our confidence in the results if there is agreement between data obtained with each method. Shadish et al.’s discussion of mono-operation bias with respect to observations focuses primarily on indirect measures of hypothetical constructs, whereas in applied behavior analysis, it is often the case that the dependent variable of interest is directly observable behavior, which may be somewhat less susceptible to this threat (Barlow et al., 2009). Nevertheless, multiple measures of a single dependent variable may increase our confidence that the behavior change observed in the study actually represents the problem of

interest. For example, researchers who study the treatment of food refusal typically employ multiple measures of refusal (e.g., Piazza, Patel, Gulotta, Sevin, & Layer, 2003) because employing only a single measure (e.g., food expulsions) might lead to incorrect conclusions regarding food intake. As with the previous threats, mono-operation bias applies not only to dependent variables but also to treatments, participants, and settings. For example, anytime specific participant characteristics (e.g., diagnoses, verbal skills) are critical for the interpretation of results, multiple measures of those characteristics may be beneficial.

Mono-method bias refers to reliance on a single method of collecting data or delivering treatment. As an example of mono-method bias in data collection, a researcher studying disruptive behavior in a preschool classroom might define several topographies of disruptive behavior and then develop an observation code to measure all topographies via a discontinuous measurement method (e.g., partial-interval recording). However, exclusive reliance on the discontinuous measurement method might result in an overestimation or underestimation of disruptive behavior relative to that which would have been obtained by continuous measurement methods. This might lead to flawed conclusions regarding the effect of an intervention. With respect to treatment delivery, mono-method bias may be a threat to the construct validity of treatment comparisons when all treatments are delivered the same way, resulting in the observed effect being specific to that method of delivery. For example, both treatments may be delivered by the same therapist, who may be more experienced or more competent with one of the treatments than the other (Kazdin, 2003).

Finally, the *confounding of constructs with levels of constructs* is a threat that occurs whenever there is a failure to recognize that the outcome of an experiment may be a function of the specific levels of the variables (persons, settings, treatments, dependent variables) that were included in the study. For example, a failure to find an effect of a reinforcement-based intervention might be due to insufficient density or magnitude of reinforcement, or the outcome of a study on the effects of a reading program on reading acquisition by individuals with intellectual disability may be specific to the level of language ability of participants in the study.

Each of the aforementioned construct validity threats related to operationalization may lead to flawed conclusions and preclude successful replication. Shadish et al. (2002) recommend careful preexperimental planning and selection of measures as strategies for preventing those threats, as well as assessing their potential operation postexperimentally through critical analysis, discussion, and perhaps a resulting respecification of conclusions. These recommendations in general apply to single-case experiments in the same manner that they apply to other research. Single-case designs rely heavily on replication as a mechanism for establishing

generality, and in applied behavior analysis, a heavy emphasis has accordingly been placed on describing treatment ingredients and measurement systems in detail, such that this *technological* practice has been considered one of the defining features of the discipline (Baer, Wolf, & Risley, 1968). Research and conceptual traditions in the discipline may therefore contain elements that help defend against some of these threats.

Interaction of Treatment with Experimental Arrangements

Shadish et al. (2002) list five threats to construct validity that describe how participants' behavior may be affected by their interactions with features of the experiment other than the treatment itself, such as measurement, setting, instructions, and experimenter behavior: (a) *treatment-sensitive factorial structure*, (b) *reactive self-report changes*, (c) *reactivity to the experimental situation*, (d) *novelty and disruption effects*, and (e) *experimenter expectancies*.

Treatment-sensitive factorial structure and *reactive self-report changes* are both concerned with the reactivity of measures. The first is specifically associated with multiple-item subjective rating instruments and refers to situations in which a treatment effect is obscured because data analysis is not sensitive to treatment-related changes in the factorial structure of the instrument. The second is, according to Shadish et al. (2002), associated with self-report measures and refers to the effects of "participant motivation to be in a treatment condition" (p. 73) upon responding on a pretest. Because those threats are specifically associated with certain data-collection methods that are uncommonly used in conjunction with single-case designs, they may rarely be of concern to single-case experimenters. It should, however, be noted that the two threats appear to have replaced the *interaction of testing and treatment* discussed in earlier work (Campbell & Stanley, 1966; Cook & Campbell, 1979), also known as *pretest sensitization*. This threat was said to refer to the possibility that "a pretest might increase or decrease the respondent's sensitivity or responsiveness to the experimental variable" (Campbell & Stanley, 1966, pp. 5–6) and appeared applicable to any situation in which the presence of pretreatment observation could plausibly render the treatment more effective than it otherwise would be. To the extent that such effects may occur with the use of direct observation, pretest sensitization may be of concern in applied behavior-analytic research. This threat may be difficult to rule out without violating the requirements of most single-case designs for baseline observations. However, it may be worth noting that the *practice* of behavior analysis ideally also involves data collection before and during treatment. If interactions between data-collection practices and treatment ever do arise in research, those interactions should therefore become part of the treatment construct and not necessarily undermine the validity of the findings.

Shadish et al. (2002) describe *reactivity to the experimental situation* as encompassing a class of threats in which “participant responses reflect not just treatments and measures but also participants’ perceptions of the experimental situation” (p. 73). Under this heading, the authors discuss placebo effects resulting from knowledge that one is receiving treatment, effects of evaluation apprehension, and the effects of demand characteristics (responding in ways that may please the experimenter). Those threats may affect the outcome of single-case experiments in a similar manner as they affect other research. In applied behavior analysis, the most frequent concern may be reactivity to direct observation, such as when a problem behavior is suppressed in the presence of an observer, or the observation procedure involves repeated opportunities to perform the target behavior, which may result in practice effects (Cooper, Heron, & Heward, 2007). Shadish et al. (2002) do not specifically discuss reactivity to observation, except to the extent that it can be attributed to evaluation apprehension, but it appears to fall under this class of threats.

The extent to which reactivity threats are plausible may depend on certain characteristics of the individuals under study, such as their verbal repertoires (e.g., ability to tact experimental conditions). Shadish et al. (2002) list several precautions that may be taken ahead of time to reduce reactive effects if they are suspected to be a potential threat. Their recommendations include measuring the dependent variable under circumstances that are as natural as possible, keeping participants unaware of the experimental hypothesis in order to reduce demand characteristics, keeping participants unaware of whether they are receiving treatment in order to prevent placebo effects, and avoiding pretests that may provide information about expected outcomes to participants. In single-case design experiments, several strategies are available for preventing and detecting reactivity threats. When data are collected via direct observation, reactivity to observation may be reduced by allowing enough time for the participants to become accustomed to being observed, or by recording unobtrusively (Barlow et al., 2009). With respect to placebo effects, single-case researchers have sometimes controlled for this threat by including inactive treatment in control conditions (e.g., Dicesare, McAdam, Toner, & Varrell, 2005; Northup et al., 1999). Some of the other reactivity threats may be more complicated to address. In applied behavior analysis, as in much other applied research, it may not be feasible or possible to keep verbally competent participants unaware of the expected outcome of an experiment and, thus, prevent effects of demand characteristics. In many cases, the participants are seeking treatment for a particular problem and it is clear to them that the goal of treatment is to alleviate that problem. This may not necessarily detract from the generalizability of the results to nonexperimental situations, in which clients are also seeking treatment with a particular outcome in mind. It is possible, however, that knowing that a treatment is

experimental and the therapists are “researchers” may function as motivating variables for behavior change. Fortunately, single-case design methodology possesses as an advantage the ease of conducting experiments as part of the typical practice of clinicians or consultants. If an original experiment is conducted under distinctly experimental conditions, successful systematic replications that are conducted as part of normal practice may help rule out demand characteristics as a threat to construct validity.

Another class of threats, *novelty and disruption effects*, refers to a specific type of reactivity that results from the mere newness of an experimental treatment. A novelty effect occurs if a treatment’s superiority results from its being new and different from other treatments, either to the participants or the person(s) delivering it. Disruption, by contrast, occurs if the new treatment is less effective than it otherwise would be, either because it is less familiar to the experimenter or because it disrupts ongoing services for the participants. Both effects should be expected to dissipate with time (Bracht & Glass, 1968). In single-case designs, their presence is therefore likely to be detected when data are collected over an extended period of time. Both effects may be of concern mainly in studies in which treatment is evaluated in brief experimental sessions or phases. In those cases, confidence in construct validity may be increased if follow-up data are reported or if findings are successfully replicated when treatment is delivered over longer periods of time. Further, extensive training of therapists prior to the study should minimize novelty and disruption effects related to implementation.

Finally, Shadish et al. (2002) cite *experimenter expectancies* as a potential threat to construct validity. This class of threats refers to the possibility that even when participants are to be kept unaware of the expected outcome of an experiment, the behavior of the experimenter might nevertheless convey his or her hypothesis or influence participant behavior in other ways. The effects of participants being aware of expected outcomes have been discussed earlier in the context of reactivity to the experimental situation. However, experimenter behavior might also affect participants’ behavior without their awareness. For example, in a study comparing two reinforcement-based interventions, in which reinforcer delivery is accompanied by social praise, an experimenter expecting one intervention to be more effective than the other might unintentionally deliver more enthusiastic praise in that condition or deliver the reinforcer with greater immediacy. As noted by Kazdin (2003), “expectancies” may not always be the most parsimonious explanation of such effects, which may be better described as differential adherence to conditions or criteria. As such, monitoring of treatment integrity, as well as successful replication across different experimenters or therapists, may be of primary importance for ruling out this threat (Kazdin, 2003).

Behavior Change in Control Conditions The final four threats to construct validity considered by Shadish et al. (2002) concern the effects of behavior change in the control group that occurs as a result of treatment implementation for the experimental group: (a) *compensatory equalization*, (b) *compensatory rivalry*, (c) *resentful demoralization*, and (d) *treatment diffusion*. *Compensatory equalization* refers to members of a control group seeking effective treatment outside of the study because they are not receiving treatment from the researchers, whereas *compensatory rivalry* refers to a tendency for the behavior of individuals in the control group to change in the same direction as that produced by the experimental treatment because of perceived competition with the experimental group. Both threats may obscure the effects of treatment. *Resentful demoralization*, by contrast, exaggerates treatment effects, as assignment to a control group leads to countertherapeutic behavior change correlated with resentment over being denied treatment. All of these threats are described specifically in the context of between-group designs and as such do not apply directly to single-case designs in which the same individual is exposed to both treatment and control conditions. It is conceivable, however, that analogous effects may occur as a result of a participant's exposure to a control condition in a single-case design. Lengthy exposure to a no-treatment baseline condition, for example, might cause a participant to seek treatment outside of the experiment, and withdrawal of treatment following successful implementation might result in effects similar to resentful demoralization, preventing reversal of responding to baseline level. The operation of these threats, however, has not to our knowledge been experimentally demonstrated in the context of single-case designs. Further, it is likely that the proper use of experimental designs, in which conditions are conducted to stability, is frequently sufficient to detect problems of this nature.

A final threat concerned with behavior change in a control condition is *treatment diffusion*, which refers to the exposure of participants in a control condition to some or all of the treatment components delivered in treatment conditions (Shadish et al., 2002). This may happen because participants in a treatment group communicate treatment-relevant information to control-group participants or because both groups are exposed to the same treatment provider who inadvertently administers some or all of the experimental treatment to the control group. Framed in the context of group designs, this is another example of a threat that may be of less relevance in single-case experiments than in group experiments. However, single-case designs may also be susceptible to treatment diffusion under some circumstances. For example, in multiple-baseline designs across participants, a participant already receiving treatment may communicate critical information to participants still in baseline, resulting in behavior change by those participants. In other designs, even in the absence of between-subject comparisons, diffusion might obscure an

effect if treatment were administered inadvertently when the experimental design required it to be withheld (e.g., in a reversal design). Therefore, researchers should take precautions ahead of time to minimize the likelihood of diffusion. Such precautions include preventing participants from communicating with one another, monitoring the behavior of persons implementing treatment, and refraining from asking relatively untrained persons in natural settings to withhold treatment once they have been trained to implement it. At the design level, treatment diffusion should be ruled out anytime behavior change is repeatedly shown to coincide with implementation and/or withdrawal of the independent variable.

External Validity

Shadish et al. (2002) define external validity as the “validity of inferences about whether the cause-effect relationship holds over variation in persons, settings, treatment variables, and measurement variables” (p. 38). Four of the external validity threats are accordingly labeled (a) *interaction of the causal relationship with units*, (b) *interaction of the causal relationship over treatment variations*, (c) *interaction of the causal relationship with outcomes*, and (d) *interaction of the causal relationship with settings*. Shadish et al. consider two types of such interactions that may occur.

The first concerns whether a functional relation observed in an experiment holds for all persons who participated in that experiment, all of the settings in which the experiment was conducted, all treatment variations that were employed, and all dependent measures. External validity is threatened if, for example, the relation holds only for a subgroup of the participants, limiting the extent to which results obtained from a large group can be generalized to individuals. This type of external validity may not be of particular concern in single-case experimentation, as the results are typically evaluated and interpreted individually for each participant, each dependent measure, and so on.

The second type of interaction occurs when the causal relationship does not hold for persons, settings, treatment variations, and dependent measures other than those employed in the experiment. At the level of persons, it concerns generalization from the individuals who participated in the experiment to other types of individuals or larger groups that may differ in composition from the one that participated. In single-case design research, this type of interaction may threaten the external validity of the results of any given experiment, in the sense that a single participant cannot be conceived of as being representative of any population. This characteristic of single-case designs may *appear* to place them at a disadvantage compared to group-design research, in which representative groups of individuals may be obtained through random sampling strategies. As Shadish et al. (2002) point out, however,

random sampling of persons is rarely practical or feasible in experiments, and even less feasible is the random sampling of settings, treatment variations, or dependent measures. Applied researchers must therefore typically rely on other methods for establishing external validity of their results. Shadish et al. recommend several strategies for preventing interaction threats, such as purposive sampling of heterogeneous instances. In single-case designs, such purposive sampling strategies might translate into the selection of participants that are deemed typical of the population under study, or ones that have been resistant to previous interventions and thus provide a stringent test of the treatment under study (Kazdin, 1981).

Shadish et al. (2002) also emphasize the importance of conducting programs of research with systematic variations in participants, settings, treatment variations, and dependent measures across studies. Such systematic replication strategies are precisely the mechanism that has been recommended for establishing the generality of findings from single-case experiments (Sidman, 1960). Although the generality of findings from a single-case experiment is initially unknown, repeating the studies with planned variations enables researchers to discover and further examine the interactions of various characteristics with treatment (Johnston & Pennypacker, 2009). A series of accumulated studies eventually provides the research consumer with sufficient information to determine under which circumstances a procedure is likely to be effective (Birnbrauer, 1981). In this context, comprehensive and well-conducted literature reviews that explicitly focus on the variations studied in systematic replications and any known boundary conditions may be helpful.

A fifth external validity threat described by Shadish et al. (2002) is *context-dependent mediation*. This threat arises when a variable observed to mediate a causal relationship in one context or setting differs from a variable that mediates the relationship in a different context. This type of validity threat thus appears to be of concern mainly in studies designed to analyze treatment components or test potential mechanisms by which a treatment may exert its effects. As with the other external validity threats, the operation of this threat in single-case experiments may be assessed through systematic replication.

Concluding Remarks

Some of the validity threats described by Shadish et al. (2002) appear to apply mainly to group-design experiments and/or experiments in which inferential statistics are employed for the purpose of data analysis. However, a majority of the threats are variables that need to be taken into consideration when designing and interpreting a single-case experiment. The single-case methodology typically used in applied behavior analysis appears to be well equipped to address all of the

relevant threats. In order to effectively prevent them, however, it is important to adhere to the conventions of experimental single-case designs, direct observation, and visual inspection. For example, the effective ruling out of many internal validity threats requires each condition to be conducted to stability, and statistical conclusion validity may be affected if measures are unreliable or if variability within conditions obscures treatment effects. Clear and unambiguous operational definitions help defend against various threats to statistical conclusion, internal, and construct validity; and assessment of treatment integrity may similarly enhance all three types of validity. Further, the establishment of external validity, as well as the ruling out of certain construct validity threats, requires conducting and publishing well-planned systematic replications of original studies.

Shadish et al.'s (2002) revised taxonomy of validity threats presents a thorough consideration of most variables that may be of concern for the validity of findings from single-case experiments. However, the authors no longer include in their taxonomy *multiple-treatment interference*, which was discussed by Campbell and Stanley (1966) as a threat to external validity, and by Cook and Campbell (1979) as a threat to construct validity. From the perspective of single-case experimentation, we believe that this is an important threat to consider, as it may affect the validity of results from single-case experiments designed to compare the effects of two or more treatments. Multiple-treatment interference occurs when the outcome of one treatment is affected by another previously administered treatment, resulting in possible failure to obtain similar results when the second treatment is administered to individuals who do not have a history of receiving the first. This threat is of concern when variations of reversal designs (e.g., an A-B-A-C-A design) are used to compare the effects of treatments (Kazdin, 2011). When treatments are compared in alternating-treatments designs, the frequent alternation of conditions rules out effects of treatment order as long as the alternations are random or semirandom (Barlow et al., 2009). However, another type of multiple-treatment interference, carryover effects, may occur in an alternating-treatments design if one condition influences behavior in an adjacent condition. Carryover effects may be prevented or reduced by spacing treatment sessions adequately, enhancing the discriminability of conditions, or reducing the speed of alternations (Barlow & Hersen, 1984). It is unclear to us why discussion of multiple-treatment interference has been omitted by Shadish et al. (2002). However, it is possible that they consider it an instance of the external validity threat—that is, an interaction of the causal relationship over treatment variations.

On a final note, users of single-case designs (Birnbrauer, 1981; Dermer & Hoch, 1999) have sometimes lamented that single-case experiments have been assigned the status of *quasi-experiments* rather than *true experiments* in the mainstream methodology literature because of a presumed inability to

control for various internal and external validity threats. Campbell and Stanley (1966) originally defined a true experimental design as one capable of controlling for all threats to internal validity and placed in this category only between-subjects experimental designs that employed random assignment of participants to groups. Designs not employing this strategy, including interrupted time-series designs, were by contrast characterized as quasi-experimental and described as ones in which the researcher “introduce[s] something like experimental design into his scheduling of data collection procedures. . . , even though he lacks full control over the scheduling of experimental stimuli. . . which makes a true experiment possible” (p. 34). Shadish et al. (2002), however, prefer the term *randomized experiment* to *true experiment*. While maintaining the distinction between designs that do and do not employ random assignment, they argue that referring to designs as “true experimental” incorrectly implies that they are the only correct experimental method for producing valid results. Shadish et al. argue that valid causal inferences may also be drawn from well-conducted quasi-experiments in which there are few *plausible* alternative sources of influence. Thus, although single-case designs still fall into the category of quasi-experimental design according to this distinction (unless they employ random assignment of data-collection times to conditions [Edgington, 1980]), there is no longer an implication that the validity of the results from a well-conducted study is thereby diminished.

Acknowledgments We thank Molli Luke for her comments on an earlier version of the manuscript.

Compliance with Ethical Standards

Conflict of Interest The authors declare no conflict of interest.

Ethical Approval This article does not contain any studies with human participants or animals performed by the authors.

References

- Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1, 91–97. <https://doi.org/10.1901/jaba.1968.1-91>.
- Bailey, J. S., & Burch, M. R. (2002). *Research methods in applied behavior analysis*. Thousand Oaks, CA: Sage.
- 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. <https://doi.org/10.1901/jaba.1979.12-199>.
- Barlow, D. H., & Hersen, M. (1984). *Single-case experimental designs: strategies for studying behavior change* (2nd ed.). New York, NY: Pergamon Press.
- Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: strategies for studying behavior change* (3rd ed.). New York, NY: Pearson.
- Bateman, M. J., & Ludwig, T. D. (2003). Managing distribution quality through an adapted incentive program with tiered goals and feedback. *Journal of Organizational Behavior Management*, 23, 33–55. https://doi.org/10.1300/J075v23n01_03.
- Bimbrauer, J. S. (1981). External validity and experimental investigation of individual behavior. *Analysis and Intervention in Developmental Disabilities*, 1, 117–132. [https://doi.org/10.1016/0270-4684\(81\)90026-4](https://doi.org/10.1016/0270-4684(81)90026-4).
- Bracht, G. H., & Glass, G. V. (1968). The external validity of experiments. *American Educational Research Journal*, 5, 437–474. <https://doi.org/10.3102/00028312005004437>.
- Branch, M. N. (1999). Statistical inference in behavior analysis: some things significance testing does and does not do. *The Behavior Analyst*, 22, 87–92. <https://doi.org/10.1007/BF03391984>.
- Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: new directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Lawrence Erlbaum.
- Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.
- Charlop-Christy, M. H., & Daneshvar, S. (2003). Using video modeling to teach perspective taking to children with autism. *Journal of Positive Behavior Interventions*, 5, 12–21. <https://doi.org/10.1177/10983007030050010101>.
- Cochrane Consumer Network. (n.d.). *Levels of evidence*. Retrieved from <http://consumers.cochrane.org/levels-evidence>.
- Conyers, C., Miltenberger, R., Maki, A., Barenz, R., Jurgens, M., Sailer, A., ... Kopp, B. (2004). A comparison of response cost and differential reinforcement of other behavior to reduce disruptive behavior in a preschool classroom. *Journal of Applied Behavior Analysis*, 37, 411–415. <https://doi.org/10.1901/jaba.2004.37-411>.
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field settings*. Chicago, IL: Rand McNally.
- Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Pearson.
- Cronbach, L. J. (1982). *Designing evaluations of educational and social programs*. San Francisco, CA: Jossey-Bass.
- Crosbie, J. (1999). Statistical inference in behavior analysis: useful friend. *The Behavior Analyst*, 22, 105–108. <https://doi.org/10.1007/BF03391987>.
- Dermer, M. L., & Hoch, T. A. (1999). Improving descriptions of single-subject experiments in research texts written for undergraduates. *The Psychological Record*, 49, 49–66. <https://doi.org/10.1007/BF03395306>.
- Dicesare, A., McAdam, D. B., Toner, A., & Varrell, J. (2005). The effects of methylphenidate on a functional analysis of disruptive behavior: a replication and extension. *Journal of Applied Behavior Analysis*, 38, 125–128. <https://doi.org/10.1901/jaba.2005.155-03>.
- Dixon, M. R. (2002). Single-subject research designs: dissolving the myths and demonstrating the utility for rehabilitation research. *Rehabilitation Education*, 16, 331–344.
- Edgington, E. S. (1980). Validity of randomization tests for one-subject experiments. *Journal of Educational Statistics*, 5, 235–251. <https://doi.org/10.2307/1164966>.
- Edgington, E. S. (1996). Randomized single-subject experimental designs. *Behaviour Research and Therapy*, 34, 567–574. [https://doi.org/10.1016/0005-7967\(96\)00012-5](https://doi.org/10.1016/0005-7967(96)00012-5).
- Gorman, B. S., & Allison, D. B. (1997). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 159–214). Mahwah, NJ: Lawrence Erlbaum.
- Harris, F. N., & Jenson, W. R. (1985). Comparisons of multiple-baseline across persons designs and AB designs with replication: issues and confusions. *Behavioral Assessment*, 7, 121–127.

- 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.
- Hopkins, B. L., Cole, B. L., & Mason, T. L. (1998). A critique of the usefulness of inferential statistics in applied behavior analysis. *The Behavior Analyst*, 21, 125–137. <https://doi.org/10.1007/BF03392787>.
- Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Kazdin, A. E. (1980). Obstacles in using randomization tests in single-case experimentation. *Journal of Educational Statistics*, 5, 253–260. <https://doi.org/10.2307/1164967>.
- Kazdin, A. E. (1981). External validity and single-case experimentation: Issues and limitations (a response to J. S. Bimbrauer). *Analysis and Intervention in Developmental Disabilities*, 1, 133–143. [https://doi.org/10.1016/0270-4684\(81\)90027-6](https://doi.org/10.1016/0270-4684(81)90027-6).
- Kazdin, A. E. (1982). *Single-case research designs: Methods for clinical and applied settings*. New York, NY: Oxford University Press.
- Kazdin, A. E. (2003). *Research design in clinical psychology* (4th ed.). Boston, MA: Allyn & Bacon.
- Kazdin, A. E. (2011). *Single-case research designs: methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.
- Lechago, S. A., Carr, J. E., Grow, L. L., Love, J. R., & Almason, S. M. (2010). Mands for information generalize across establishing operations. *Journal of Applied Behavior Analysis*, 43, 381–395. <https://doi.org/10.1901/jaba.2010.43-381>.
- Michael, J. (1974). Statistical inference for individual organism research: mixed blessing or curse? *Journal of Applied Behavior Analysis*, 7, 647–653. <https://doi.org/10.1901/jaba.1974.7-647>.
- Morgan, D. L., & Morgan, R. K. (2001). Single-participant research design: bringing science to managed care. *American Psychologist*, 56, 119–127. <https://doi.org/10.1037/0003-066X.56.2.119>.
- Northup, J., Fusilier, I., Swanson, V., Huete, J., Bruce, T., & Freeland, J. (1999). Further analysis of the separate and interactive effects of methylphenidate and common classroom contingencies. *Journal of Applied Behavior Analysis*, 32, 35–50. <https://doi.org/10.1901/jaba.1999.32-35>.
- O’Leary, K. D., Kent, R. N., & Kanowitz, J. (1975). Shaping data collection congruent with experimental hypotheses. *Journal of Applied Behavior Analysis*, 8, 43–51. <https://doi.org/10.1901/jaba.1975.8-43>.
- Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: new directions for psychology and education* (pp. 15–40). Hillsdale, NJ: Lawrence Erlbaum.
- Perone, M. (1999). Statistical inference in behavior analysis: experimental control is better. *The Behavior Analyst*, 22, 109–116. <https://doi.org/10.1007/BF03391988>.
- Peterson, L., Homer, A. L., & Wonderlich, S. A. (1982). The integrity of independent variables in behavior analysis. *Journal of Applied Behavior Analysis*, 15, 477–492. <https://doi.org/10.1901/jaba.1982.15-477>.
- Piazza, C. C., Patel, M. R., Gulotta, C. S., Sevin, B. M., & Layer, S. A. (2003). On the relative contributions of positive reinforcement and escape extinction in the treatment of food refusal. *Journal of Applied Behavior Analysis*, 36, 309–324. <https://doi.org/10.1901/jaba.2003.36-309>.
- Ryan, C. S., & Hemmes, N. S. (2005). Effects of the contingency for homework submission on homework submission and quiz performance in a college course. *Journal of Applied Behavior Analysis*, 38, 79–88. <https://doi.org/10.1901/jaba.2005.123-03>.
- Ryan, S., Ormond, T., Imwold, C., & Rotunda, R. J. (2002). The effects of a public address system on the off-task behavior of elementary physical education students. *Journal of Applied Behavior Analysis*, 35, 305–308. <https://doi.org/10.1901/jaba.2002.35-305>.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Houghton Mifflin.
- Sidman, M. (1960). *Tactics of scientific research*. Boston, MA: Authors Cooperative.
- Tate, R. L., Perdices, M., Rosenkoetter, U., McDonald, S., Togher, L., Shadish, W., ... Vohra, S. (2016). The single-case reporting guideline in behavioral interventions (SCRIBE) 2016: Explanation and elaboration. *Archives of Scientific Psychology*, 4, 10–31. <https://doi.org/10.1037/arc0000027>.
- Van Houten, R., & Malenfant, J. E. L. (2004). Effects of a driver enforcement program on yielding to pedestrians. *Journal of Applied Behavior Analysis*, 37, 351–363. <https://doi.org/10.1901/jaba.2004.37-351>.
- Vollmer, T. R., Iwata, B. B. A., Zarcone, J. R., Smith, R. G., & Mazaleski, J. L. (1993). Within-session patterns of self-injury as indicators of behavioral function. *Research in Developmental Disabilities*, 14, 479–492. [https://doi.org/10.1016/0891-4222\(93\)90039-M](https://doi.org/10.1016/0891-4222(93)90039-M).
- Vollmer, T. R., Sloman, K. N., & St. Peter Pipkin, C. (2008). Practical implications of data reliability and treatment integrity monitoring. *Behavior Analysis in Practice*, 1(2), 4–11. <https://doi.org/10.1007/BF03391722>.