

# Single-Case-Design Research in Special Education: Next-Generation Guidelines and Considerations

Exceptional Children  
2023, Vol. 89(4) 379-396  
© The Author(s) 2022  
Article reuse guidelines:  
sagepub.com/journals-permissions  
DOI: 10.1177/00144029221137656  
journals.sagepub.com/home/ecx



Jennifer R. Ledford <sup>1</sup>, Joseph M. Lambert<sup>1</sup>,  
James E. Pustejovsky<sup>2</sup>, Kathleen N. Zimmerman <sup>3</sup>,  
Nicole Hollins<sup>4</sup>, and Erin E. Barton<sup>1</sup>

## Abstract

Single-case design has a long history of use for assessing intervention effectiveness for children with disabilities. Although these designs have been widely employed for more than 50 years, recent years have been especially dynamic in terms of growth in the use of single-case design and application of standards designed to improve the validity and applicability of findings. This growth expanded possibilities and inspired new questions about the contributions this methodology can make to generalizable knowledge about intervention in special education. In this article, we discuss and extend previous standards for studies using single-case designs. We identify new suggestions for internal validity, generality and acceptability, and reporting. We also provide considerations for single-case synthesis and discuss the complexities of assessing accumulating evidence for a given practice.

Single-case designs (SCDs) allow for causal (functional) relations to be established between environmental conditions and participant behaviors (Horner et al., 2005; Kazdin, 2016; Riley-Tillman et al., 2020). Several features characterize SCDs: (a) focus on within-participant or within-group changes in behavior rather than between-participant or between-group changes (i.e., the individual or case serves as their own control), (b) repeated measurement over time in at least two conditions, and (c) following prespecified rules for introduction (and/or withdrawal) of conditions. These rules entail one of three paradigms: sequential introduction and withdrawal, rapid iterative alternation, or time-lagged implementation. These rules and specific design variations (e.g., multiple baseline, withdrawal) are described in numerous other sources. Readers unfamiliar with basic features of SCD may benefit from reviewing those (Johnston et al., 2010; Kazdin, 2016; Ledford

et al., 2018; Ledford & Gast, 2018), as this discussion of contemporary issues with SCD assumes fluency with the logic and rationale behind these methods.

SCDs have been important for special education research since their inception, and they remain both frequently used and necessary for accumulating evidence in the field (e.g., 83% of studies assessing interventions for individuals with autism use SCDs; Steinbrenner et al., 2020). Moreover, the individual focus

---

<sup>1</sup>Vanderbilt University

<sup>2</sup>University of Wisconsin

<sup>3</sup>University of Kansas

<sup>4</sup>EdBeeConsultations

## Corresponding Author:

Jennifer R. Ledford, Vanderbilt University, Peabody  
College Box 228, Nashville, TN, 37203, USA.  
Email: Jennifer.ledford@vanderbilt.edu

of SCDs is aligned well with the goals of education in general and special education in particular (Repp & Lloyd, 1980). Despite wide use, agencies and researchers have often excluded SCD research evidence from systematic reviews due to difficulty quantifying and synthesizing outcomes (Shadish et al., 2015), related to the use of visual analysis as the historically preferred method for data analysis. SCDs were historically derived using an inductive model of research, with the purpose of isolating conditions that control specific behaviors for a given participant (Johnson & Cook, 2019; Repp & Lloyd, 1980). More recently, scholars have used SCD in alignment with a deductive approach, emphasizing evaluation of an independent variable for improving a dependent variable, given hypotheses based on theory and previous work. The synthesis of data across studies (e.g., systematic review and meta-analysis) for the purpose of determining the evidence base of an intervention is necessarily deductive in nature.

Outside of textbooks, the first notable attempt to set standards for SCD came when the Council for Exceptional Children (CEC) Division of Research (DR) commissioned a task force that developed quality indicators for SCD, published in *Exceptional Children* in 2005. These authors identified quality indicators for SCD research and proposed how SCD studies could contribute to evidence for effective intervention practices for disabled people. (Throughout this manuscript, we use a mix of identity-first and person-first language, given many self-advocates prefer identity-first but while acknowledging that some individuals prefer person-first language; e.g., Bury et al., 2020; Kenny et al., 2016.) The authors formalized criteria for rigorous SCD research, which we will refer to as the DR-SCD Standards (Horner et al., 2005). This group argued its proposed standards would allow readers to determine whether an individual study was a “credible example” of SCD research and whether a practice was validated across studies as “evidence-based” (Horner et al., 2005, p. 165). The measurable criteria proposed by the authors allowed for more consistent evaluation of previously conducted studies and provided benchmarks for

researchers as they planned and conducted new studies. This article solidified historically accepted but inconsistently applied *conventions* described by textbooks and established them instead as *standards*. This was intended to be helpful for the field, with clear expectations leading to more consistent production and assessment of SCD research.

Almost a decade after the DR-SCD Standards appeared, a different group proposed new standards, as part of a different task force commissioned by the CEC (2014) and called the “Standards for Evidence-Based Practices in Special Education” (referred to hereafter as the CEC-EBP Standards). These standards combined expectations for group designs and SCDs, with some standards applying to both and others to only SCD or only group designs. Generally, the CEC-EBP Standards were similar to those proposed by the DR in 2005 (see Supplemental File 1 for a crosswalk).

Alongside standards developed via CEC, the What Works Clearinghouse (WWC) of the Institute for Education Sciences (IES) designed a more limited set of standards focused on inclusion of single-case studies in WWC reviews (WWC, 2010). IES also explicitly identified SCDs as appropriate for funded research, especially related to intervention development and iterative refinement (Kaiser, 2014). The WWC standards were much more limited than DR-SCD and CEC-EBP standards, including only a few items directly related to internal validity. The WWC has also continued to iterate on their proposed standards, which were recently graduated from “pilot” status to become a regular component of their standards (WWC, 2020). The DR-SCD, CEC-EBP, and WWC standards have been widely used and cited in relation to judging study rigor for inclusion in systematic reviews and for assessing quality of included studies.

The DR-SCD, CEC-EBP, and WWC standards have all contributed to developments in the field, including increased acceptance of SCD research, greater attention to rigor of SCD research, and increased consideration for how to consider outcomes for large bodies of SCD research. Benefits of these advancements cannot be overstated. Given this context, we propose to qualify and

extend guidance for the field by offering insights related to the strengths and weaknesses of these commonly used standards. Our focus will be on the importance of accounting for contextual factors (e.g., measurement of generalized behaviors in typical settings, behavior complexity, intervention development stage) that impact attainment of conventional standards and on the nuance required to understand evidence accumulation in special education research. Our recommendations are designed to assist researchers in planning and conducting highly rigorous SCD studies and to help them identify areas of strength and weaknesses in other studies. They may be helpful, for example, for conducting assessments of whether certain study characteristics are associated with outcomes (e.g., if studies without interobserver agreement (IOA) data have larger effects, we might surmise that observer bias accounts for differences). We propose that these guidelines serve not as a minimum standard for inclusion in reviews (such as the WWC standards) or for publication but rather as a set of recommended practices for the field. We intend for the recommendations to be considered preliminary and as a starting point for future work, especially related to developing consensus among SCD researchers.

In the first section of the manuscript, we explain guiding principles used to develop the recommended practices and then provide the recommendations, which are listed in Tables 1 through 3. In Section 2, we discuss SCD meta-analysis and synthesis. In Section 3, we examine the complexities associated with identifying “evidence-based practices.” In all sections, we use the guiding principles and draw heavily from previous reviews (e.g., DR-SCD, 2005; WWC, 2020).

## **Section I: Recommended Practices for Conducting SCD Research**

### *Guiding Principles*

1. **Strengths of studies should be considered separately by domain (i.e.,**

**internal validity vs. external validity).** The DR-SCD and CEC-EBP standards are divided into seven or eight domains (e.g., internal validity, dependent variables) with many standards related to reporting, several related to internal validity, and some related to external validity; the WWC standards solely focus on internal validity. We separate recommendations for internal validity, external validity, and reporting, with the intention of setting clear expectations for the production of new evidence and for better characterization of accumulated evidence in systematic reviews. We use the term *external validity* to refer to generality outside of the original study contexts or populations and *social validity* to refer to the extent to which stakeholders and participants find interventions, goals, and procedures to be acceptable.

2. **Emphasis on internal or external validity should be based on the stage of research and its purposes.** DR-SCD standards suggest researchers should attend to socially important dependent variables, typical intervention agents, and practical and cost-effective intervention implementation. CEC-EBP retained a few of these considerations (e.g., dependent variables should be “socially important”). However, guidelines related to external and social validity may not be similarly applicable for all studies—instead, they are relevant when the original researcher or systematic reviewer has an associated research question. Relatedly, when the research question calls for it, social validity should be sacrificed for internal validity. For example, when comparing instructional procedures, internally valid comparisons require the use of stimuli that are unlikely to be learned outside the study context (i.e., to protect against history threats). In this case, having

socially *unimportant* dependent variables is required for adequate internal validity (i.e., participants cannot contact the information in typical activities outside of the study). The ability to answer the research question (i.e., Which procedure leads to faster acquisition?) is facilitated by failing to meet the standard of social importance. We thus emphasize that the types and qualities of evidence provided by a study should be scaled in accordance with the purpose of that study (e.g., increased emphasis on internal and less on external validity when trying to establish the existence of a phenomenon under controlled conditions and the converse when trying to demonstrate social validity).

3. **Guidelines for the field should be flexible.** Expectations to adhere to inflexible rules may be a context that discourages important work and encourages problematic research practices. For example, researchers might select a behavior that is easy to measure rather than a more complex one that better represents the construct of interest (e.g., contrived interactions during specific trials rather than interactions occurring during typical activities) due to ease of reaching specific interobserver agreement standards. Or researchers might collect data that increase the likelihood of high agreement (e.g., gross agreement rather than point by point) but result in decreased ability to have discrepancy discussions (i.e., exact disagreements are not apparent), thus impeding accurate data collection. Whereas these decisions allow researchers to meet some standards as conceived by the field, they actually decrease the extent to which threats to internal validity are controlled. This consequence might be avoided by allowing for flexibility while requiring authors to understand and provide rationales for their

decisions. Of course, such flexibility creates a different but potentially impactful problem: When there is no official “bar” to meet, who decides whether an author’s rationale is sufficient? One important consideration when determining to what extent guidelines should be applied is whether researchers are using inductive or deductive logic. For example, the extent to which preregistration of procedures and analyses is appropriate is dependent on whether researchers take an inductive approach and use frequent, dynamic decision-making or whether they take a more deductive approach that is relatively static.

4. **Guidelines for the field should be context specific.** This principle is related to the previous two, in that we argue that nuance is necessary. For example, it may be considerably easier to collect reliability data (i.e., IOA and fidelity) during researcher-implemented, trial-based, video-recorded research in the context of short sessions rather than when measuring free-operant behaviors during sessions conducted by endogenous implementers in typical contexts. If we hold all research to the same rules (e.g., collection of sufficient reliability data with high agreement and adherence), we may unintentionally reduce meaningful research on socially valid outcomes that are representative of typical contexts because it is harder to meet certain benchmarks in this type of work. As an example, consider reliable data collected in the context of an SCD related to a clinic-based intervention for challenging behavior, with a clearly demonstrated functional relation. These data “meet standards,” which has implications for both acceptance for publication and inclusion in systematic reviews. Now consider a context in which, still as part of a SCD, a caregiver collects data on challenging behavior that occurs *outside* of sessions. It would be unreasonable to

assess parent fidelity to procedures throughout the day or to assess observer reliability when the variable of interest (e.g., generality) occurs in isolated contexts in which a single adult is providing care (e.g., at home, across a typical day; Lambert et al., 2022). Disregarding these data for publication and synthesis, however, would be a serious disservice to the field. Again, we run into a different, potentially impactful problem: Who decides when collecting fewer reliability data or none at all is warranted? To answer these questions, we argue that the peer review process, when conducted by knowledgeable agents, should serve this purpose. This would be facilitated by widespread acceptance of flexible standards. Expert review and consensus building across many groups of researchers is also needed (e.g., via the use of a series of Delphi studies).

5. **Diversity, equity, and inclusion should be a priority.** It is apparent we should engage in equitable, diverse, and inclusive behaviors when conducting research and providing clinical services in special education settings (Kratochwill et al., 2021; Morris et al., 2021; Pritchett et al., 2021). Unfortunately, systemic equity issues, like structural racism and lack of cultural awareness, likely impact all stages of research. This includes steps such as determining dependent variables of interest, identification of participants who receive interventions and the settings in which they receive them, the extent to which developed interventions and materials are inclusive and culturally appropriate for a diverse range of participants, and retention of and outcomes for participants. To ensure goals, methods, and outcomes remain socially valid, it is crucial for researchers to ensure that culturally relevant and inclusive practices are embedded into research protocols (cf. Hayes & Toarmino, 1995; Shogren et al., 2021). Because SCD research is flexible,

it offers opportunities for researchers to recruit participants from underserved populations and use implementation science and participatory research models to design effective and valued interventions (Pritchett et al., 2021; Sutherland et al., 2022; Yee, 2016). Thus, researchers should consider individual and cultural differences that shape value systems, experiences, and needs of participants. This has the potential to improve social validity of given interventions across diverse groups. When appropriate, the use of mixed-methods research may help answer social validity questions associated with cultural relevance in tandem with experimental questions (Corr et al., 2020).

6. **Science should be conducted transparently.** Replicability is a cornerstone of single-case research and science more generally. Replicability and transparency of SCD research studies are crucial for maintaining the self-correction processes on which science relies (Cook et al., 2021a, 2021b). Promotion of transparency has become increasingly attainable through development of preregistration practices and facilities for providing online supplements to published articles, which reduces the constraints of page limits on the ability of researchers to fully explain decisions related to the design, production, and analysis of SCD research.

### *Conduct and Assessment of SCD Studies*

Given recent advancements and our guiding assumptions, we provide recommendations for SCD analysis in Tables 1 through 3. We identify recommended practices related to internal validity, external validity, and reporting for (a) design and research questions, (b) dependent variables and measurement, (c) data collection and analysis, (d) participants and contexts, (e) independent variables and implementers, and (f) transparency, ethics, and cost. Citations are intended to direct

readers to additional resources; Supplemental File 2 includes these references. Throughout, we avoid specific criteria (e.g., 80% agreement) and instead encourage researchers to make a priori decisions, justify them based on context, and report them transparently. Next, we describe general considerations for each domain and some major differences between our recommendations and common standards in the field (e.g., DR-SCD, WWC). Supplemental File 3 includes rationales for each guideline, and Supplemental File 4 lists guidelines in a matrix format with internal validity, external validity, and reporting listed together for each domain.

**Internal Validity.** In Table 1, we identify internal validity guidelines, including maximizing ability to detect threats via design selection, measurement practices, and transparent planning. We specify that designs must include three potential demonstrations of effect (as required by WWC) rather than three demonstrations of effect (as required by DR-SCD, CEC-EBP). This clarification prevents studies with noneffects from being disqualified as sources of evidence, given other criteria are met (Tincani & Travers, 2018). Notably, we do not specify that a pattern of experimental control is necessary, as doing so could result in exclusion of noneffects from systematic reviews. Although consistent noneffects do not always allow us to draw confident conclusions (i.e., plausible alternative explanations might exist), bodies of work including noneffects are important for identifying contexts in which a given intervention is unlikely to produce positive outcomes. Consistent with historical precedent, we assert that expected changes that occur at least three times and are consistent within the design provide evidence for a functional relation.

In our guidelines, we emphasize importance of stable data patterns rather than a fixed number of data points (e.g., WWC, 2020) and specify that randomization and use of naive observers (blind or masked assessors) can reduce expectancy bias in some circumstances. We recommend use of visual analysis as the analysis method appropriate for identifying functional relations—because

it allows determination of both whether outcomes were positive (i.e., changed when and only when conditions changed) and whether common threats to internal validity are likely (e.g., history, maturation). Visual analysis guidelines are available in Supplemental File 5. We clarify that effect sizes can be reported as an estimate of magnitude of change between conditions and that some recommendations are relevant only when a deductive, rather than an inductive, approach is used to guide the research (Cook et al., 2021b; Johnson & Cook, 2019).

**External and Social Validity.** In Table 2, we specify that social validity, generalization, maintenance, and cost outcomes may be relevant given a researcher's questions. We assert that study parameters (i.e., participants, contexts, implementers) should be representative of situations to which authors aim to generalize or that differences should be explicitly identified. We specify that social validity procedures less subject to social desirability bias are preferred and that authors should explicitly address social desirability if more subjective measures are used (e.g., when a researcher associated with implementation administers a questionnaire). Notably, we do not specify that the social importance of the dependent variable, social value of behavior change, cost-effectiveness, or use of endogenous implementers or contexts is essential for every study. The importance of each should be determined based on the contribution of a study in the context of the broader literature. For example, if a single study contributes initial exploration of a practice, then social validity may not be a focus. However, an assessment of an intervention contributing to a well-established literature base may promote translation of the intervention to typical contexts by considering cost-effectiveness and employing endogenous implementers in typical contexts.

**Reporting.** In Table 3, we specify specific reporting guidelines crucial for interpreting SCD research. We acknowledge that researchers may not be able to report all recommended information in a typical manuscript but argue that inclusion of online supplemental materials (via journals or third-party repositories) is generally possible. Notable differences from

**Table 1.** Recommendations for Improving Internal Validity.

## Internal Validity

**Design and research questions: Select and use designs that allow for detection of threats to internal validity so that research questions can be answered.**

- Use design appropriate for research questions and dependent variables (Ledford et al., 2019a).
- Use a design with at least three potential demonstrations of effect between adjacent conditions (e.g., three A-to-B comparisons). Designs with fewer than three demonstrations due to participant withdrawal or noneffects can be considered as potentially important evidence of intervention limits.
- Determine a priori a strategy for assigning participants to intervention start times (time-lagged designs) or condition order (rapid iterative alternation designs). Randomization, including blocked randomization, is acceptable, as is purposeful ordering (e.g., planning functional analysis conditions in an order designed to minimize carryover effects) (Kratochwill & Levin, 2014; Ledford, 2018; Levin et al., 2019).
- Use multiple-baseline designs only when baseline lengths are sufficiently different in terms of time, number of sessions, and date (Slocum et al., 2022).

**Dependent variables and measurement: Measure behaviors in ways that increase believability.**

- When possible, have data collectors who are naive to study condition. In some studies, condition changes are apparent (e.g., different materials across conditions). In other situations, resources or practical constraints prohibit the use of naive raters. When use of naive coders is not feasible or possible, report explicitly that increased risk of bias may exist (Yoder et al., 2018b; Yoder & Crandall, 2019).
- Use a measurement system representative of the dimension of interest (e.g., duration, number).
- Collect reliability data frequently and obtain acceptable agreement for each condition and participant. Determine criteria for the frequency and level of agreement data appropriate for the specific study context, and rationalize decisions (e.g., trial-based data in a contrived context may require high agreement, whereas free-operant behaviors measured in a classroom may require lower agreement). If collecting reliability data is not possible, criteria are low, or criteria are not met, explain the extent to which this impacts data interpretation (Repp et al., 1976).
- Regularly assess agreement between primary and secondary coders throughout the study (e.g., by graphing data from both observers and assessing for bias or drift) and take steps to resolve discrepancies as needed (e.g., retraining) (Ledford & Wolery, 2013; Yoder et al., 2018b).

**Data collection and analysis: Collect enough data to draw confident conclusions about effects and analyze according to predetermined protocols.**

- When using a deductive approach, identify the primary variable a priori. This variable should be used in response-guided decision-making and to draw primary conclusions.
- When using a deductive approach, identify secondary variables (e.g., corollary, generalization) and explicitly identify whether these variables represent exploratory or confirmatory relations.
- When using a deductive approach, identify a priori expected data patterns within and between conditions (e.g., stable baseline data with increasing trend in intervention for skill acquisition data, variability in both conditions, with a change in level between conditions for challenging behavior data collected in a classroom).
- Collect data on a sufficient number of measurement occasions to establish level, trend, and variability in each condition. Behaviors with more stable data patterns or those that are subject to testing threats require fewer data points to establish internal validity. Exceptions should be made for dangerous or otherwise highly problematic behaviors (Kazdin, 1978). When using response-guided decision-making, stable data patterns are more important than specific numbers of data points.
- If it is aligned with a research question, calculate an appropriate effect size to describe the magnitude of differences between conditions.

**Independent variables and measurement: Measure implementation to ensure differences between conditions occurred as planned.**

- Identify components of intervention intended to result in desired outcomes (i.e., presumed active ingredients), including materials and implementer behavior. Consider how ingredients are used (e.g., rules), by whom, for how long (dosage, including time and the number of trials, if applicable), and how frequently (e.g., over what time frame). Identify what components are present in intervention only (i.e., independent variables) and those that are present across conditions (i.e., control variables; Abry et al., 2014; Cook et al., 2017; Frost et al., 2020; Yoder et al., 2012).
- Collect data providing evidence that all conditions are implemented as described using direct

(continued)

**Table 1.** (continued)**Internal Validity**

measurement of implementation, including systematic direct observation by a nonimplementing observer for all participants in all conditions (Collier-Meek et al., 2020, 2021). When direct observation is not possible or feasible, note that data may not be reliable and use other safeguards (e.g., regular reminders sent to parents who collected data at home).

- When the likelihood of expectancy bias is high (e.g., data collection involves subjective assessments or rating scales), collect agreement data on fidelity measurement.

**Transparency, ethics, and cost: Improve replicability and generality by transparently reporting plans, deviations, and relevant data.**

- Preregister studies or explicitly plan and document via a written protocol which components of the procedures are determined a priori and which components are changed using response-guided decision-making (Cook et al., 2021a, 2021b). If response-dependent decisions are made (e.g., condition change decisions, intervention modifications), describe the processes for determining these decisions (e.g., operational definitions of data stability or nonresponse).
- Share data and additional study information (e.g., protocols, instructional materials) using open-science resources (Cook et al., 2021a, 2021b).

the previous standards include recommendations for increased transparency and reporting information necessary to judge ethical administration of SCD research, such as conflicts of interest and adverse events. We also encourage authors to share preprints or postprints to increase access to research; most journals allow for posting of unformatted postprints even after the article has been published.

## Section 2: Synthesis of SCD Research

As the availability of findings from SCD studies has grown, researchers have sought tools to draw broader conclusions about intervention practices using evidence from multiple studies. Some tools for doing so are based on rules of thumb or professional conventions, such as the approach described in the DR-SCD Standards (described further in the next section). Other approaches are based on systematic review and meta-analysis methods, which involve quantitative analysis of results from multiple SCD studies.

Over the past two decades, production of systematic reviews and meta-analyses has burgeoned, although methodologists have noted important shortcomings in conduct and reporting (Jamshidi et al., 2018). Along with increased production of reviews, the field has also seen a methodological renaissance, with

many new—and increasingly sophisticated—proposals for how to quantitatively analyze and meta-analyze SCD data (Manolov et al., 2021; Shadish, 2014). Many of the processes and considerations involved in conducting a systematic review and synthesis of SCD research are closely analogous to those involved in synthesizing findings from between-group experimental studies or other types of research designs (Pustejovsky & Ferron, 2017). Readers interested in the how-to of synthesis can look elsewhere for overviews (Pustejovsky & Ferron, 2017) and comprehensive, book-length guides (Cooper, 2010). However, several considerations are more unique to synthesis of SCD research. In this section, we review several of the central issues related to (a) selecting an effect size metric and an approach to quantitative synthesis, (b) unique features of SCDs, and (c) selective reporting.

### *Effect Size Metrics*

Many of the methodological developments in quantitative analysis of SCD studies have focused on effect size measures, which are quantitative indices that summarize the direction and strength of an intervention's relation with a dependent variable (Pustejovsky & Ferron, 2017). A broad array of effect size measures have been explored specifically for use with SCD studies, including both measures based on parametric statistical models



**Table 2.** Recommendations for Improving External and Social Validity.

## External and social validity

**Dependent variables and measurement: Define and measure behaviors that are representative of those expected to change, given theory of change that drives intervention selection or development.**

- Select and define a behavior that is representative of the construct of interest.
- When applicable to research questions, use social validity measures with guards against potential social desirability bias, including naive raters of conditions, normative comparison data of peers who do not meet inclusion criteria, continued intervention use by endogenous implementers, participant preference for intervention, stakeholder choices regarding intervention or procedures, or the use of interviewers who are not apparently affiliated with study implementation (Kennedy, 2002; Ledford et al., 2016; Snodgrass et al., 2021).
- Ensure that intervention and intervention targets are socially valid for all participants throughout the research study (e.g., initial interviews, intervention check-ins, poststudy questionnaires) and that both are developed to be inclusive of participants' cultural context (Colic et al., 2021; Lovelace et al., 2018).
- When applicable to research questions and in alignment with the theory of change, assess outcomes under generalization conditions that permit inferences about participants' ability to use target behaviors in contexts of interest (e.g., at home with caregivers, in classroom with teachers; e.g., Ramirez et al., 2019). Generalization data provided in the context of single-case design (e.g., data in generalization context collected concurrently with data in training context) provides the most compelling evidence of the presence or absence of generalized behavior change, whereas more intermittent data collection, pre- and postintervention data collection, or postintervention-only data collection provides nonexperimental but still potentially important evidence (Ledford & Gast, 2018).
- When applicable to research questions and in alignment with the theory of change, assess outcomes not directly targeted by the intervention (e.g., corollary or generalization outcomes).
- When applicable to research questions and in alignment with the theory of change, assess behavioral changes in a postintervention condition in which no intervention components are present. Contextually bound or readily reversible behaviors may not be expected to maintain in the absence of intervention elements; authors should articulate why maintained outcomes are or are not expected to support understanding of why maintenance measurement did or did not occur.

**Independent variables and implementers: Select contexts that are representative of those to which authors would like to generalize findings.**

- Specify inclusion criteria and descriptive information for implementers likely to impact fidelity of implementation (e.g., teaching certification, at least one year of work in a classroom, prior experience with intervention). When researchers are intervention agents, specify researcher characteristics that might impact real-world application (e.g., resources, training, certification, experience).

**Transparency, ethics, and cost: When cost is of interest, identify whether interventions are feasible for adoption.**

- Calculate cost of an intervention when relevant, especially when considering conditions under which an intervention is likely to result in optimal outcomes. Cost is most important when authors specify that an intervention should be scaled up for use in typical settings (Bradshaw et al., 2020; Schiebel et al., 2022).

**Participants and contexts: Select participants and contexts to which generalizations are desirable.**

- Devise inclusion criteria that are necessary for participants to benefit from the intervention, and use those criteria (and only those criteria) to select participants.
- Ensure that recruitment methods are equitable and that minoritized or disadvantaged groups are not excluded based on recruitment, procedures, language spoken, or accessibility (Lovelace et al., 2018; Oh-Young, 2022; Pritchett et al., 2021) or that such exclusions are explicit to allow other researchers to draw appropriate generalizations (e.g., only English-speaking participants were recruited due to resource constraints).
- Describe recruitment methods (e.g., use of a particular school district) and population from which study participants were drawn (e.g., children with autism) separately from inclusion criteria, unless these characteristics are associated with likelihood of intervention benefit.
- Select contexts representative of settings and activities to which generalization is desirable. Researchers may elect to use more highly controlled environments, even when researching phenomena that are applicable in applied settings. This is acceptable if they explicitly acknowledge contextual differences between research and applied settings and that generalization to applied settings is unknown.

**Table 3.** Recommendations for Improving Reporting.

## Reporting

**Design and research questions: Report questions guiding the study and enough information about design to allow readers to draw accurate conclusions.**

- When using a deductive approach, report at least one directional and falsifiable research question (primary question). Report other questions as applicable (e.g., exploratory questions, corollary outcomes, generalization, maintenance, social validity). When questions are inductive in nature, specify that such is the case.
- Report design type, including concurrence (for multiple-baseline and multiple-probe designs).
- Report specific operational rules used to determine condition changes (response guided) or how condition lengths were selected and assigned (predetermined).
- Report rationale for selecting design and how potential threats to validity may be detected.
- Report randomization decisions and condition ordering rules transparently.

**Dependent variables and measurement: Report relevant information about how behaviors are defined and measured.**

- Explicitly report which outcome is primary (on which condition change decisions are made, in alignment with the theory of change) or secondary (corollary) and whether all measured dependent variables were planned a priori.
- Describe operational definitions and examples and nonexamples of target behaviors.
- Describe how coders were trained and whether they had to reach a specific criterion before beginning data collection.
- Describe measurement system with replicable precision, including system and features (e.g., partial interval recording with 15-s intervals and 5-s breaks), assessors (e.g., graduate students naive to condition), and timing (e.g., collected during sessions or via video recording).
- When discontinuous measurement is used, explicitly report dimension of interest (e.g., momentary time sampling to estimate duration) and evidence that the discontinuous method used matches the dimension of interest (Prykanowski et al., 2018; Pustejovsky & Swan, 2015; Wood et al., 2016).
- Report explicitly the extent to which behaviors represent context-bound versus generalized and proximal versus distal outcomes. Avoid reporting conclusions about distal or generalized behavior change based on proximal or context-bound measurement (Sandbank et al., 2021).
- Report whether agreement data are collected. If so, report the percentage of sessions during which data were collected, percentage of agreement for each participant in each condition, and method of calculation used.

**Participants and contexts: Report relevant characteristics to allow for generalization beyond the study.**

- Report recruitment information, inclusion criteria, relevant descriptive data, and demographic characteristics (race-ethnicity, age, languages spoken, gender [including nonbinary choices], relevant individual or context-specific [e.g., school] socioeconomic status information). Report relationship between participant and implementer. Content experts are required to determine what descriptive data are relevant given study goals, but demographic data should always be reported.
- Refer to participants using terms indicated as preferred by them (e.g., “Hispanic” or “Latino”; “autistic adult” or “adult with autism”).
- Report data for all recruited participants, including those who withdraw before completion.
- Report whether inclusion was limited to English-speaking participants.
- Report characteristics of settings and activities likely to aid in replication attempts (e.g., presence of nonparticipating children or adults, whether the activity was typically occurring or contrived).

**Data collection and analysis: Report information about data analysis to allow readers to understand decision-making.**

- Present data for all measured variables using typical graphing conventions and avoiding misleading display (Dart & Radley, 2017, 2018; Ledford et al., 2019b). (See Supplemental Files 7–10.)
- Report information about the role of the data collectors in relation to the study (e.g., the principal investigator and implementer, undergraduate research assistant) and whether data collectors were naive to study conditions (with rationale for naive or informed coders).
- When social validity measures are more subject to bias (e.g., interviews and questionnaires conducted by the same person who implemented intervention procedures or who otherwise appears to have a

(continued)

**Table 3.** (continued)**Reporting**

stake in positive outcomes of a study), consider and explicitly acknowledge social desirability bias as a potential cause for concern.

- Report systematic visual analysis procedures (Ledford et al., 2018; Wolfe et al., 2019).
- When using a deductive approach, describe expected changes between conditions, established a priori, and extent to which data patterns corresponded to these predictions.
- Explicitly identify whether a functional relation exists for each opportunity (e.g., if there are two participants with two measured dependent variables, measured in the context of A-B-A-B designs, there are four potential functional relations).

**Independent variables and measurement: Report information about conditions and fidelity measurement to enhance data believability and replicability.**

- Describe all components of all conditions and materials used in each. Identify the extent to which procedures and materials were different across conditions.
- Describe dosage (how often condition was implemented and for how long).
- Describe the percentage of sessions during which fidelity data were collected and the percentage of agreement in each condition.
- Describe behaviors measured for fidelity purposes and measurement procedures. Report the relation between fidelity data collectors and implementers.
- Describe all modifications made to intervention procedures and whether these modifications were determined a priori or based on individual participant response. Explicitly describe the impact of modifications on confidence in functional relation determination.
- Describe intervention agents in terms of demographic characteristics, experience with participants (i.e., endogenous, outside researcher) and intervention elements, and professional qualifications (e.g., certified teacher).
- If implementers are endogenous, explain how the intervention agent was trained by researchers and evidence of fidelity of training (i.e., implementation fidelity). If a researcher was the intervention agent, describe training and expertise.
- Describe similarities and differences between generalization sessions and baseline and intervention sessions, including whether any intervention components or materials are present during generalization, and describe when and how often generalization sessions occurred.
- Describe similarities and differences between maintenance sessions and baseline and intervention sessions, including whether any intervention components or materials are present, and report latency between intervention stoppage and maintenance measurement.

**Transparency, ethics, and cost: Report transparently all procedures and information related to ethical administration of single-case-design studies.**

- When allowable, share preprints on a preprint server. Preprints do not necessarily need to be, but could be, shared prior to publication.
- Share raw data, including data from all measured dependent variables, either via supplemental materials or via online repositories, such as the Open Science Framework (<https://osf.io>).
- Systematically record adverse events throughout the study, and describe any adverse events associated with participation for any participant (or report no adverse events were recorded) (Bottema-Beutel et al., 2021a).
- Describe conflicts (and potential conflicts) of interest for each author (e.g., assessment of an intervention on which author has been paid to present; Bottema-Beutel et al., 2021b).
- Report that institutional review board permissions were obtained, and describe both consent and assent procedures.
- Report cost of intervention when relevant, especially when considering conditions under which it is likely to result in optimal outcomes demonstrated in tightly controlled clinical settings.

and measures based on the nonparametric concept of nonoverlap (Manolov et al., 2021). However, many of the indices that were once most widely used have severe limitations. In particular, some nonoverlap

indices (e.g., percentage of nonoverlapping data, percentage of all nonoverlapping data, robust improvement rate difference) are strongly influenced by incidental features, such as the number of baseline observations

or number of treatment phase observations (Allison & Gorman, 1994; Pustejovsky, 2019; White, 1987), which are irrelevant to the strength of an intervention's effect.

The most fundamental distinction between available effect sizes pertains to the metric—or scale—on which they quantify an intervention's effect. An effect size metric needs to be meaningful and interpretable for the interventions and dependent variables examined in the set of studies to be synthesized. For example, the log response ratio (Pustejovsky, 2018) describes change in a dependent variable in terms of the ratio (or percentage change) in mean level from a baseline phase to an intervention phase. It is therefore interpretable when dependent variables are measured on a ratio scale and where it is sensible to characterize change in percentage terms (e.g., a 50% reduction in aggressive behavior); it is inappropriate for behavior acquisition studies in which the behavior is absent (or nearly so) during baseline, such that percentage change is not meaningful. Other effect sizes are based on other conceptual metrics, such as distributional nonoverlap (e.g., nonoverlap of all pairs; Parker & Vannest, 2009), progress toward a prespecified goal (percentage of goal obtained; Ferron et al., 2020), mean differences relative to within-participant variability in an outcome (within-case standardized mean difference; Busk & Serlin, 1992), or mean differences relative to the total variability in an outcome (between-case standardized mean difference; Shadish et al., 2014). In selecting an effect size, researchers need to consider the properties of the dependent variables examined in included studies to identify a metric that is interpretable and can be meaningfully compared from one study to another.

In addition to effect size considerations, researchers also have several options for how to synthesize findings across studies. Three unique approaches to synthesis of SCDs have been developed, including approaches based on (a) integrative modeling of raw data from all included studies; (b) participant-specific effect size estimates, summarized in a multilevel meta-analysis; and (c) study-level summary effect size estimates, summarized using conventional meta-analytic methods. These options are useful in different contexts, depending on the

characteristics of the set of studies identified for synthesis. See Supplemental File 6 for more information.

### *Unique Features of SCD Studies*

SCD studies have unique features that bear consideration when conducting syntheses. One such feature is that researchers may make response-guided decisions about when to start intervention or change from one condition to another (Edgington, 1983; Ledford & Gast, 2018). Effect size estimation and synthesis methods do not generally account for use of response-guided practices, and the limited available methodological research comes to mixed conclusions about whether effect size estimates could be biased by use of these practices (Joo et al., 2018; Swan et al., 2020). As methodological investigation continues, we recommend researchers conducting syntheses of SCD studies attend to whether included studies use response-guided practices, potentially examining this factor as a moderator of effect size. Increased transparency in reporting would make such endeavors more feasible (see Table 1, Reporting).

Another feature of SCDs is that they are dynamic, allowing for modification of procedures or addition of intervention components. Because of this, SCD studies may contain results for multiple interventions or variations. Researchers conducting a synthesis need to determine not only whether a study should be included or excluded but which conditions within a study are relevant. For instance, in a study of a behavioral intervention, researchers might determine that the initial baseline phase and initial intervention phase are relevant for inclusion but that a subsequent phase involving a modified intervention is not relevant. The alternative of giving preference to phases with adaptations of the intervention could lead to a distorted picture because it would tend to exclude data where the initial form of intervention was ineffective. Increased transparency in reporting intervention components and modification plans would aid synthesists in making judgments about which phases to include (see Table 1, Reporting).

### Selective Reporting and Publication Bias

*Selective reporting* and *publication bias* refer to the phenomenon in which affirmative study findings are more easily published compared with those that are inconclusive, ambiguous, or counter to expectations. Selective reporting poses a major validity threat for systematic review and synthesis efforts because it affects what evidence is available for inclusion, leading to overrepresentation of affirmative findings (Rothstein et al., 2005). In synthesis of between-group designs, selective reporting is broadly understood to be driven by the statistical significance of study findings, and a wide array of statistical tools have been developed for assessing and correcting for biases created by selective reporting (Marks-Anglin & Chen, 2020; McClain et al., 2021). Evidence indicates that selective reporting and publication bias also present concerns for SCD research (Dowdy et al., 2020; Gage et al., 2017; Shadish et al., 2016; Sham & Smith, 2014). However, because statistical analysis is rarely the primary means of drawing conclusions in primary SCD studies, it is unlikely that statistical significance is a primary driver of selective reporting. Tools used for probing selective reporting in syntheses of group designs may therefore be unsuitable for investigating selective reporting in syntheses of SCD studies. Lacking tools, it becomes more important to *prevent* selective reporting and mitigate its biasing effects through practices such as study preregistration, publication of studies indicating noneffects, and identification of unpublished or “gray” literature for potential inclusion in systematic reviews (Johnson & Cook, 2019; Pustejovsky & Ferron, 2017; Tincani & Travers, 2018, 2019).

### Section 3: Identification of Evidence-Based Practices

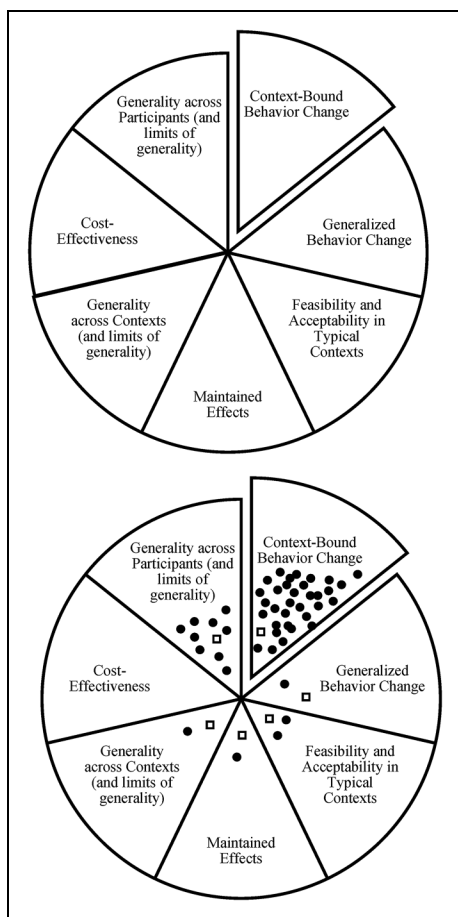
As noted already, the DR-SCD and CEC-EBP standards included guidelines for evaluating whether sufficient evidence existed across studies to identify a practice as evidence based, likely inspired by federal regulations requiring evidence-based educational interventions. The DR-SCD standards required positive outcomes demonstrated by at least three research groups

with five or more SCD studies including at least 20 participants (referred to as the 3-5-20 rule). The CEC-EBP standards use this rule as well but also allow some neutral or mixed effects if the ratio of positive to mixed or neutral is at least 3:1. The WWC (2020) requires 350 participants for moderate or strong evidence—a bar unlikely to be met by any intervention assessed exclusively via SCDs. No current standards explicitly assert that evidence-based practice entails considerations that extend beyond research evidence (e.g., clinical judgment and client characteristics, cultures, and preferences; American Psychological Association, 2006; Contreras et al., 2021).

All previous standards are at least partially a reaction to production of SCD studies increasing in rate and amassing over time, resulting in the need for consumable syntheses of what is known about specific practices for improving outcomes for students with disabilities. The 3-5-20 rule has been widely used to identify practices as evidence based in special education and related fields (e.g., Steinbrenner et al., 2020). However, especially when evidence of noneffects also exists, the 3-5-20 rule may be insufficient or erroneously applied (Ledford et al., 2021a), and even more nuanced rules (CEC-EBP) may not provide information about for whom and under what conditions interventions are likely to be effective or ineffective. No single criterion has been endorsed by the field, although important advances in effect size development and meta-analysis have occurred since previous standards were published, as described earlier.

Science is ever evolving, and evidence accumulates over time. Thus, a determination that a particular practice has a certain amount of research support may be relevant for a short period of time. Nonetheless, it is helpful for the field to address the extent to which certain practices have accumulated support for solving specific problems, for specified participants, in particular contexts, using a deductive approach. However, making a dichotomous decision of “evidence based” or “not evidence based” may be insufficient or misleading.

As shown in Figure 1, several conclusions can be drawn about a practice that meets a



**Figure 1.** Different types of evidence that can be accumulated for a practice. The data points included in the bottom frame are intended to represent the amount of evidence accumulated in each area for a hypothetical intervention. Filled circle data points represent rigorous evidence of positive effects. Unfilled square data points represent rigorous evidence of negative or noneffects.

minimum bar of evidence, all of which provide different information relevant to an *evidence based* determination. For example, one component of evidence accumulated for a practice is whether we have rigorous evidence for context-bound behavior change (i.e., change in the intervention context; Sandbank et al., 2021). Given this, additional work may be needed to establish external and social validity. The lower panel of Figure 1 shows hypothetical evidence for an unnamed practice, with each filled circular data point representing a positive outcome

from a rigorous study and each unfilled square data point representing a negative outcome from a rigorous study. For this practice, there is sufficient evidence of context-bound behavior change for a variety of participants but sparse and mixed evidence across other categories. This provides us with important next steps, for example, determining whether generalized behavior change is likely and whether the practice is feasible and acceptable in typical contexts.

As a further example, we highlight a review of interruption and redirection procedures, (Ledford et al., 2022). For this practice, evidence for decreases in vocal stereotypy for autistic individuals who are served in separate educational programs exists when the practice is being actively implemented. There is limited evidence for social validity and clear evidence for noneffects for generalized and maintained outcomes. Thus, although the practice meets clear criteria as “evidence based” using previous standards, limitations are considerable, boundaries are important, and questions about social and external validity remain unanswered.

Given the complexities associated with assessing practices (Ledford et al., 2021b), we propose that systematic reviews of interventions result not in a determination of “evidence based” or “not evidence based” but rather in a determination of the contexts in which the practice has more or less support and in what areas additional research is needed. Researchers are eager to be influenced by rules, but their behavior may be insensitive to contingencies that operate in opposition to expectation (cf. Hayes et al., 1989). The most straightforward solution may be the development and presentation of sufficiently dynamic rules and a willingness to modify such rules as evidence of their impact emerges. Specifically, we suggest authors describe (a) the contexts under which the practice has been evaluated; (b) the outcomes of the practice, including differences across dependent variable types, context characteristics, and context-bound versus generalized behaviors; and (c) the limits of current knowledge about the practice, including unanswered questions about maintenance, generalization, long-term effects, costs and benefits, and social validity. Given these data and the theory of change for the

practice, authors should be better able to describe situations in which the use of the practice aligns with evidence, and situations in which use of the practice is contraindicated or without evidence.

In reconsidering standards for SCD, our primary purposes were to provide contemporary updates and to acknowledge the complexity of assessment of SCD that can be lost when distilling syntheses to a single decision about a practice. These complexities make assessment of SCD studies more difficult and conclusions less likely to be straightforward. However, this complex engagement is needed to reflect the variable and nuanced human experiences SCD research is designed to investigate. We acknowledge several limitations in our approach, including lack of consensus among other leaders in the field. We propose that rigorous testing of these guidelines is warranted (e.g., via focus groups, Delphi study, application across various bodies of research) to establish consensus, evaluate validity, and improve uptake.

## References

- Allison, D. B., & Gorman, B. S. (1994). "Make things as simple as possible, but no simpler." A rejoinder to Scruggs and Mastropieri. *Behaviour Research and Therapy*, 32(8), 885–890. [https://doi.org/10.1016/0005-7967\(94\)90170-8](https://doi.org/10.1016/0005-7967(94)90170-8)
- American Psychological Association Presidential Task Force on Evidence-Based Practice. (2006). Evidence-based practice in psychology. *American Psychologist*, 61(4), 271–285. <https://doi.org/10.1037/0003-066X.61.4.271>
- Bury, S. M., Jellett, R., Spoor, J. R., & Hedley, D. (2020). "It defines who I am" or "It's something I have": What language do [autistic] Australian adults [on the autism spectrum] prefer? *Journal of Autism and Developmental Disorders*. Advance online publication. <https://doi.org/10.1007/s10803-020-04425-3>
- Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In Kratochwill, T. R., & Levin, J. R. (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Lawrence Erlbaum.
- Contreras, B., Hoffmann, A., & Slocum, T. (2021). Ethical behavior analysis: evidence-based practice as a framework for ethical decision making. *Behavior Analysis in Practice*, 15(3), 619–634. <https://doi.org/10.1007/s40617-021-00658-5>
- Cook, B. G., Fleming, J. I., Hart, S. A., Lane, K. L., Therrien, W. J., van Dijk, W., & Wilson, S. E. (2021a). A How-To Guide for Open-Science Practices in Special Education Research. *Remedial and Special Education*, 43(4), 270–280. <https://doi.org/10.1177/07419325211019100>
- Cook, B. G., Johnson, A. H., Maggin, D. M., Therrien, W. J., Barton, E. E., Lloyd, J. W., & Travers, J. C. (2021b). Open science and single-case design research. *Remedial and Special Education*, 43(5), 359–369. <https://doi.org/10.1177/0741932521996452>
- Cooper, H. M. (2010). *Research synthesis and meta-analysis* (4th ed.). Sage.
- Corr, C., Snodgrass, M. R., Greene, J. C., Meadan, H., & Santos, R. M. (2020). Mixed methods in early childhood special education research: Purposes, challenges, and guidance. *Journal of Early Intervention*, 42(1), 20–30. <https://doi.org/10.1177/1053815119873096>
- Dowdy, A., Tincani, M., & Schneider, W. J. (2020). Evaluation of publication bias in response interruption and redirection: A meta-analysis. *Journal of Applied Behavior Analysis*, 53(4), 2151–2171. <https://doi.org/10.1002/jaba.724>
- Edgington, E. S. (1983). Response-Guided experimentation. *Contemporary Psychology: A Journal of Reviews*, 28(1), 64–65. <https://doi.org/10.1037/021569>
- Ferron, J., Goldstein, H., Olszewski, A., & Rohrer, L. (2020). Indexing effects in single-case experimental designs by estimating the percent of goal obtained. *Evidence-Based Communication Assessment and Intervention*, 14(1–2), 6–27. <https://doi.org/10.1080/17489539.2020.1732024>
- Gage, N. A., Cook, B. G., & Reichow, B. (2017). Publication bias in special education meta-analyses. *Exceptional Children*, 83(4), 428–445. <https://doi.org/10.1177/0014402917691016>
- Hayes, S. C., Kohlenberg, B. S., & Melancon, S. M. (1989). Avoiding and altering rule-control as a strategy of clinical intervention. In *Rule-governed behavior* (pp. 359–385). Boston, MA: Springer.
- Hayes, S. C., & Toarmino, D. (1995). If behavioral principles are generally applicable, why is it necessary to understand cultural diversity? *The Behavior Therapist*, 18(1), 21–23.
- Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S., & Wolery, M. (2005). The use of single-subject research to identify evidence-based

- practice in special education. *Exceptional Children*, 71(2), 165–179. <https://doi.org/10.1177/001440290507100203>
- Jamshidi, L., Heyvaert, M., Declercq, L., Fernández-Castilla, B., Ferron, J., Moeyaert, M., & Van den Noortgate, W. (2018). Methodological quality of meta-analyses of single-case experimental studies. *Research in Developmental Disabilities*, 79, 97–115. <https://doi.org/10.1016/j.ridd.2017.12.016>
- Johnson, A. H., & Cook, B. G. (2019). Preregistration in single-case design research. *Exceptional Children*, 86(1), 95–112. <https://doi.org/10.1177/0014402919868529>
- Johnston, J. M., & Pennypacker, H. S. (2010). *Strategies and tactics of behavioral research*. Routledge.
- Joo, S.-H., Ferron, J. M., Beretvas, S. N., Moeyaert, M., & Van den Noortgate, W. (2018). The impact of response-guided baseline phase extensions on treatment effect estimates. *Research in Developmental Disabilities*, 79, 77–87. <https://doi.org/10.1016/j.ridd.2017.12.018>
- Kaiser, A. P. (2014). Using single-case research designs in programs of research. In Kratochwill, T. R., & Levin, J. R. (Eds.), *Single-case intervention research: Methodological and statistical advances* (pp. 309–323). American Psychological Association.
- Kazdin, A. E. (1978). Methodological and interpretive problems of single-case experimental designs. *Journal of Consulting and Clinical Psychology*, 46(4), 629.
- Kazdin, A. E. (2016). *Single-case research designs: Methods for clinical and applied settings*. Oxford University Press.
- Kenny, L., Hattersley, C., Molins, B., Buckley, C., Povey, C., & Pellicano, E. (2016). Which terms should be used to describe autism? Perspectives from the UK autism community. *Autism*, 20(4), 442–462. <https://doi.org/10.1177/1362361315588200>
- Kratochwill, T. R., Horner, R. H., Levin, J. R., Machalicek, W., Ferron, J., & Johnson, A. (2021). Single-case design standards: An update and proposed upgrades. *Journal of School Psychology*, 89, 91–105. <https://doi.org/10.1016/j.jsp.2021.10.006>
- Lambert, J. M., Sandstrom, A., Hodapp, R., Copeland, B. A., Paranczak, J. L., Macdonald, M. J., & Houchins-Juarez, N. J. (2022). Revisiting the social validity of services rendered through a university-based practicum addressing challenging behavior. *Journal of Applied Behavior Analysis*, 55(4), 1220–1238. <https://doi.org/10.1002/jaba.939>
- Ledford, J. R., & Gast, D. L. (Eds.). (2018). *Single case research methodology: Applications in special education and behavior sciences*. Routledge.
- Ledford, J. R., Lane, J. D., & Severini, K. E. (2018). Systematic use of visual analysis for assessing outcomes. *Brain Impairment*, 19(1), 4–17. <https://doi.org/10.1017/brimp.2017.16>
- Ledford, J. R., Lambert, J. M., Barton, E. E., & Ayres, K. M. (2021a). The evidence base for interventions for individuals with ASD: A call to improve practice conceptualization and synthesis. *Focus on Autism and Other Developmental Disabilities*, 36(3), 135–147. <https://doi.org/10.1177/10883576211023349>
- Ledford, J. R., Truemp, C., Chazin, K. T., Windsor, S. A., Eyler, P. B., & Wunderlich, K. (2021b). Systematic review of interruption and redirection procedures for autistic individuals. Advance online publication. *Behavioral Interventions*. <https://doi.org/10.1002/bin.1905>
- Manolov, R., Tanious, R., & Onghena, P. (2021). Quantitative techniques and graphical representations for interpreting results from alternating treatment design. *Perspectives on Behavior Science*, 45(1), 259–294. <https://doi.org/10.1007/s40614-021-00289-9>
- Marks-Anglin, A., & Chen, Y. (2020). A historical review of publication bias. *Research Synthesis Methods*, 11(6), 725–742. <https://doi.org/10.1002/jrsm.1452>
- McClain, M. B., Callan, G. L., Harris, B., Floyd, R. G., Haverkamp, C. R., Golson, M. E., Longhurst, D. N., & Benallie, K. J. (2021). Methods for addressing publication bias in school psychology journals. *Journal of School Psychology*, 84, 74–94. <https://doi.org/10.1016/j.jsp.2020.11.002>
- Morris, C., Detrick, J. J., & Peterson, S. (2021). Participant assent in behavior analytic research: Considerations for participants with autism and developmental disabilities. *Journal of Applied Behavior Analysis*, 54(4), 1300–1316. <https://doi.org/10.1002/jaba.859>
- Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, 40(4), 357–367. <https://doi.org/10.1016/j.beth.2008.10.006>
- Pritchett, M., Ala'i-Rosales, S., Cruz, A. R., & Cihon, T. M. (2021). Social justice is the spirit and aim of an applied science of human behavior: Moving from colonial to participatory research practices. *Behavior Analysis in*



- Practice*. <https://doi.org/10.1007/s40617-021-00591-7>
- Pustejovsky, J. E. (2018). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. *Journal of School Psychology, 68*, 99–112. <https://doi.org/10.1016/j.jsp.2018.02.003>
- Pustejovsky, J. E. (2019). Procedural sensitivities of effect sizes for single-case designs with directly observed behavioral outcome measures. *Psychological Methods, 24*(2), 217–235. <https://doi.org/10.1037/met0000179>
- Pustejovsky, J. E., & Ferron, J. (2017). Research synthesis and meta-analysis of single-case designs. In *Handbook of special education* (2nd ed., p. 63). Routledge.
- Repp, A. C., Deitz, D. E., Boles, S. M., Deitz, S. M., & Repp, C. F. (1976). Differences among common methods for calculating interobserver agreement. *Journal of Applied Behavior Analysis, 9*(1), 109. <https://doi.org/10.1901/jaba.1976.9-109>
- Repp, A. C., & Lloyd, J. (1980). Evaluating educational changes with single-subject design. In Gottlieb, J. (Ed.), *Educating mentally retarded persons in the mainstream*. University Park Press.
- Riley-Tillman, T. C., Burns, M. K., & Kilgus, S. P. (2020). *Evaluating educational interventions: Single-case design for measuring response to intervention*. Guilford.
- Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.), *Publication bias in meta-analysis: Prevention, assessment, and adjustments* (pp. 1–7). Wiley.
- Sandbank, M., Chow, J., Bottema-Beutel, K., & Woynarowski, T. (2021). Evaluating evidence-based practice in light of the boundedness and proximity of outcomes: Capturing the scope of change. *Autism Research, 14*(8), 1536–1542. <https://doi.org/10.1002/aur.2527>
- Shadish, W. R. (2014). Statistical analyses of single-case designs: The shape of things to come. *Current Directions in Psychological Science, 23*(2), 139–146. <https://doi.org/10.1177/0963721414524773>
- Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). *The role of between-case effect size in conducting, interpreting, and summarizing single-case research* (NCER 2015–002). National Center for Education Research.
- Shadish, W. R., Hedges, L. V., & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology, 52*(2), 123–147. <https://doi.org/10.1016/j.jsp.2013.11.005>
- Shadish, W. R., Zelinsky, N. A. M., Vevea, J. L., & Kratochwill, T. R. (2016). A survey of publication practices of single-case design researchers when treatments have small or large effects. *Journal of Applied Behavior Analysis, 49*(3), 656–673. <https://doi.org/10.1002/jaba.308>
- Sham, E., & Smith, T. (2014). Publication bias in studies of an applied behavior-analytic intervention: An initial analysis: Publication bias in applied behavior analysis. *Journal of Applied Behavior Analysis, 47*(3), 663–678. <https://doi.org/10.1002/jaba.146>
- Shogren, K. A., Scott, L., Hicks, T., Raley, S., Hagiwara, M., Pace, J., & Kiblen, J. (2021). Exploring self-determination outcomes of racially and ethnically marginalized students with disabilities in inclusive general education classrooms. *Inclusion, 9*(3), 189–205. <https://doi.org/10.1352/2326-6988-9.3.189>
- Steinbrenner, J. R., Hume, K., Odom, S. L., Morin, K. L., Nowell, S. W., Tomaszewski, B., & Savage, M. N. (2020). *Evidence-based practices for children, youth, and young adults with autism*. FPG Child Development Institute.
- Swan, D. M., Pustejovsky, J. E., & Beretvas, S. (2020). The impact of response-guided designs on count outcomes in single-case experimental design baselines. *Evidence-Based Communication Assessment & Intervention, 14*(1–2), 82–107. <https://doi.org/10.1080/17489539.2020.1739048>
- Tincani, M., & Travers, J. (2018). Publishing single-case research design studies that do not demonstrate experimental control. *Remedial and Special Education, 39*(2), 118–128. <https://doi.org/10.1177/0741932517697447>
- Tincani, M., & Travers, J. (2019). Replication research, publication bias, and applied behavior analysis. *Perspectives on Behavior Science, 42*(1), 59–75. <https://doi.org/10.1007/s40614-019-00191-5>
- What Works Clearinghouse. (2010). *Procedures and standards handbook (Version 1.0.)*. <https://files.eric.ed.gov/fulltext/ED510743.pdf>
- What Works Clearinghouse. (2020). *Procedures and standards handbook (Version 4.1)*. <https://ies.ed.gov/ncee/wwc/Protocols#>
- White, O. R. (1987). Some comments concerning “The Quantitative Synthesis of Single-Subject Research.” *Remedial and Special Education, 8*(2), 34–39. <https://doi.org/10.1177/074193258700800207>
- Wolfe, K., Barton, E. E., & Meadan, H. (2019). Systematic protocols for the visual analysis of single-case research data. *Behavior Analysis in Practice, 12*(2), 491–502. <https://doi.org/10.1007/s40617-019-00336-7>

Yee, A. (2016). Autism research's overlooked racial bias. The Atlantic. Available at: <https://med.fsu.edu/sites/default/files/news-publications/print/Autism%20%20Research's%20Racial%20Bias%20-%20The%20Atlantic.pdf>

### ORCID iDs

Jennifer R. Ledford  <https://orcid.org/0000-0002-2392-7103>

Kathleen N. Zimmerman  <https://orcid.org/0000-0003-0271-0965>

### Supplemental Material

The supplemental material is available in the online version of the article.

Manuscript received March 2022; accepted August 2022.