Guide for the Use of Single Case Design Research Evidence

from the Division for Research of the Council for Exceptional Children

Jennifer R. Ledford and Kathleen N. Zimmerman, Vanderbilt University
Ilene Schwartz, University of Washington
Samuel Odom, University of North Carolina at Chapel Hill

As outlined in the Policy and Position Statement from the Division for Research of the Council for Exceptional Children (CEC), CEC recognizes the legitimacy and importance of single case research (SCR) for identifying evidence-based practices in special education (Rodgers, Lewis, O'Neill, & Vannest, 2017). CEC has also identified quality indicators for determining quality of studies using SCR methodology (CEC, 2014).

The purpose of this document is to provide guidelines for practitioners to assess evidence from existing SCR studies to identify potential evidence-based interventions for improving specific behaviors for individuals with and at-risk for a variety of disabilities.

Experimental Designs

Randomized controlled trials (RCTs) are often considered the "gold standard" for assessing evidence for the effectiveness of an intervention (Shadish, Cook, & Campbell, 2002). RCTs are a group of *experimental* designs that allow for causal attributions to be made—any differences between groups can be credited to the intervention due to the pre-intervention randomization (i.e., any pre-existing differences on important attributes should not impact results, given large enough groups). Other designs, such as quasi-experimental (non-randomized) group comparison designs, provide less compelling evidence (What Works Clearinghouse [WWC], 2017). Single case designs, when implemented according to certain standards, are also considered *experimental* in nature (Horner, Carr, Halle, McGee, Odom, & Wolery, 2005; Ledford, 2017; Shadish, Hedges, Horner, & Odom, 2015; WWC, 2017). That is, SCR studies can be used to establish causal links between interventions and changes in behavior. In the remaining sections of the paper, we will explain: (a) under what conditions SCR is generally employed, (b) how experimental attributions are made without randomly assigning participants to groups, (c) three major types of single case designs, (d) assessing rigor and outcomes of SCR studies across the three major types, and (e) summarizing evidence across multiple SCR studies.

Introduction to Single Case Research

SCR has been widely used to establish evidence in a range of fields (general education, special education, related services, psychology) for more than 50 years. It was originally associated with teaching individuals with intellectual disabilities in clinical and institutional settings (see

Schumaker & Sherman, 1970). It remains especially prevalent in use with low-incidence populations and for young children with disabilities. However, rigorous SCR has recently been conducted with children with high-incidence disabilities, individuals without disabilities, and in non-educational contexts. Moreover, it has been recognized by a wide range of agencies as legitimate and useful (e.g., American Speech-Language-Hearing Association; Council for Exceptional Children; Institute of Education Sciences). Thus, SCR can be used when assessing intervention effectiveness across a broad range of participants and in a variety of contexts.

One of the primary advantages of SCR is that outcomes can be determined individually. For example, a teacher, clinician, or researcher can identify a problem for a specific student or client, an intervention that may be effective for improving the problem, and measurement procedures likely to be sensitive to expected behavior changes. Although termed "single" case research, this can also refer to establishing change in intact groups (e.g., class-wide problem behavior). Moreover, SCR studies are *dynamic* in nature—that is, if behavior does not improve and expected changes do not occur, changes can be made to the intervention (or an entirely new intervention can be assessed) while maintaining experimental control (see Barton et al., 2016).

Components of Single Case Design Research

Single case research requires repeated measurement of behavior over time. This repeated measurement may include, for example, assessing the extent to which a behavior occurs for multiple consecutive days. Most often, the types of behaviors appropriate for measurement in SCR are proximal and context-bound. That is, the behaviors measured are specifically targeted in the intervention (e.g., performance on single digit addition, rather than performance on a norm-referenced mathematics achievement test) and expected to change under specific conditions. Proximal and context-bound behaviors are often of interest to teachers and other practitioners who serve individuals with disabilities—thus, use of SCR to inform decisions regarding interventions to target these behaviors is recommended. This may be especially true for students for whom typical norm-referenced assessments may be insensitive to important changes in student behavior. For example, teaching a preschool student with autism and significant cognitive impairments to request items and activities may significantly impact his or her quality of life and participation in school environments without resulting in a substantial change in overall language performance as measured via a formal assessment. Table 1 shows several examples of pairs of studies intended to improve skills in the same general domain in the context of RCTs versus single case designs. Note that SCR typically employs researcher-developed, observational measures while RCTs typically employ standardized, pre/post measures.

Single Case Research Logic

To rule out alternative explanations for behavior change, systematic condition ordering (rather than randomization to a group; Ledford, 2017; Shadish et al., 2015) is used in SCR. This is accomplished by comparing data for each participant across adjacent conditions (generally, *baseline* and *intervention* conditions); this is referred to as "baseline logic." In SCR, data are collected repeatedly over time, generally via direct observation of observable behaviors. Condition changes are made dependent on data stability within conditions and changes between conditions. If measured behavior (the dependent variable) changes *when and only when* conditions change, in the expected direction, causal attributions can be made if a rigorous and appropriate single case design is used. Decisions regarding whether a causal attribution can be made are dependent on determining whether (a) the study is sufficiently rigorous, and (b) visual analysis indicates a functional relation is present.

Types of Single Case Research

There are three primary condition-ordering variations: (a) time-lagged introduction (including multiple baseline designs), (b) sequential introduction and withdrawal (including A-B-A-B designs), and (c) rapid iterative alternation (including alternating treatments designs).

Time-lagged introduction. The most commonly used single case design is the multiple baseline design (Hammond & Gast, 2010). Other time-lagged designs include changing criterion and multiple probe designs. When multiple baseline or multiple probe designs are used, baseline data are collected concurrently for three or more participants, contexts, or behaviors, and the intervention is introduced to each at a different point in time (see Figure 1 for examples of multiple baseline across participants and multiple probe across behaviors designs). When changing criterion designs are used, the intervention is time-lagged across several criterion levels following an initial baseline condition.

Sequential introduction and withdrawal. Withdrawal designs, also referred to as A-B-A-B designs, are commonly used SCR designs. These designs consist of initial baseline data collection (A1), intervention implementation (B1), return to baseline condition (A2), and a second intervention condition (B2). A common variation is the multitreatment design (A-B-C-B-C), in which two intervention conditions, rather than a baseline and intervention condition, are compared to each other. These designs must only be used to measure *reversible* behaviors; that is, behaviors that are likely to change readily based on the context (e.g., problem behavior or social interactions) rather than on learning history (e.g., academic or self-help skills). An A-B-A-B withdrawal design is shown in Figure 2.

Rapid iterative alternation. Rapid iterative alternation designs include the commonly-used alternating treatments design (ATD) and less commonly-used adapted alternating treatments design (AATD) and repeated acquisition design (RAD). The ATD includes at least two conditions that can be rapidly alternated; conditions are often ordered using restricted randomization (Ledford, 2017). An ATD is shown in Figure 3; ATDs are used when measuring reversible behaviors. AATDs and RADs are used when measuring non-reversible behaviors, such as correct academic responding.

Assessing Rigor

Some rigor expectations are applicable for all design types. Quality indicators and proposed standards (e.g., CEC, 2014; WWC, 2017) vary somewhat, but generally include: (a) three or more potential demonstrations of effect, (b) adequate data collection in each condition, and (c) adequate evidence of reliability of dependent and independent variables. If a study meets these requirements, a functional relation can be assessed. Specific requirements are detailed below, and additional information, examples, and non-examples are presented in Table 2.

Replication. All single case studies should include three potential demonstrations of effect. That means that all SCR studies should include *replicated* behavior changes. In sequential introduction and withdrawal (e.g., A-B-A-B) and rapid iterative alternation designs (e.g., ATD), these changes occur as two or more conditions are alternated—for example, in an A-B-A-B design, there are three potential points in time in which behavior change occurs—when the initial intervention condition is implemented ($A_1 \rightarrow B_1$), when the intervention is withdrawn ($B_1 \rightarrow A_2$), and when it is implemented again ($A_2 \rightarrow B_2$). In time-lagged designs, the behavior change is implemented with different participants, behaviors, or contexts. Regardless of the design, *replication* of behavior change is at the heart of SCR research. In Figures 1-3, each potential demonstration is denoted with a number; note that all have three or more.

Data sufficiency. Within each condition, there must be a sufficient number of data points such that evaluation of the data is possible. What is considered "sufficient" has varied over time and there is not agreement on what should be considered acceptable (see Ledford, Barton, Severini, & Zimmerman, 2018; WWC, 2017). Generally, there should be at least three data points per condition because that is the minimum number of data points needed to determine whether a trend is present. More data may be needed if data values are not consistent over time, while three total measurement occasions may be enough, particularly if data are at ceiling or floor values (e.g., 0% correct responding for three consecutive sessions may provide convincing evidence that a child's correct responding is unlikely to improve in the absence of intervention).

Reliability. Assessments of reliability are intended to provide information about the believability of data. Generally, reliability is assessed via interobserver agreement—that is, two observers record the same behaviors at the same time and determine the similarity of their measurements. Agreement is typically reported as a percentage, and higher percentages indicate higher believability. If agreement is low (e.g., below 80%), we have less confidence that the observer's data reflect the true occurrence of behavior. For example, low agreement may indicate observer bias, such that the observer's expectation of behavior change during intervention conditions affected the data he or she collected, rather than the actual behavior occurrence.

Fidelity. In addition to assessing reliability of the measurement of target behaviors (dependent variables), researchers should also assess the accuracy with which procedures were executed as planned. This measurement of the extent to which researchers (or other intervention providers) implemented procedures as intended increases confidence that the intervention *and only the intervention* is responsible for any changes in behavior.

Other characteristics. In addition to the characteristics of rigorous studies, the ability to assess generality is improved if authors adequately describe participants, settings, procedures, measurement, and training of implementers. In time-lagged designs, it is critical that baseline data are collected concurrently (i.e., baseline conditions start at the same time) and that sufficient baseline data are collected to enable analysts to assess whether behavior changes when and only when intervention is implemented (i.e., to assess stability of baseline data prior to intervention implementation).

Social validity, the extent to which the goals, procedures, and outcomes of a study are acceptable to relevant consumers, may also be a critical consideration. For example, if a teacher of students with visual impairments wants to evaluate whether repeated reading is evidence-based for improving reading rates, she would be interested in the extent to which researchers included individuals like her students. She also might want to know whether the goals, procedures, and outcomes of the study were meaningful, as assessed by the students, teachers, parents, or other relevant individuals.

Likewise, ecological validity, the extent to which the study is similar to intervention contexts of interest, may be important. For example, a teacher might be interested in whether the repeated reading interventions have been conducted by classroom teachers, in classroom contexts, and with typical classroom resources and supports (e.g., low-cost materials, without extensive researcher assistance). A meaningful and rigorous single case study can have low ecological validity and no evidence of social validity; however, the extent to which the results are applicable to a given context may depend on these two factors.

Assessing Outcomes

When SCR is used, outcomes assessment is conducted using visual analysis. Data are generally depicted via a line graph, with dependent variable values on the y-axis and time on the x-axis. A functional relation is identified if visual analysis indicates that behavior change occurs in the predicted direction *when and only when* condition changes occur. Decisions made from visual analysis are dichotomous although differences in magnitude of effect (e.g., small to large) or confidence (e.g., uncertain to very certain) can exist. Figure 1 shows evidence of a functional relation; data change when and only when the intervention is introduced across tiers. Figure 2 shows evidence of *no* functional relation; data do not systematically change when A→B condition changes occur.

Six data characteristics can be evaluated via visual analysis:

- Level: Level refers to the value of the dependent variable. When a functional relation is present, levels are similar within identical conditions and different across conditions.
- Trend: Trend refers to the slope of the data and can be increasing or decreasing. For some dependent variables, we expect a gradually increasing trend (e.g., learning academic content; see Figure 1), and for others, we expect an immediate change in level with no trend changes (e.g., engagement; see Figure 2).
- Variability: Variability refers to the session-to-session differences in data value; if there are no trends in the data, variability can be represented as the range of values (e.g., 15-25% of intervals). Interventions tend to target changes in level (with or without changes in trend), but variability is important because high variability within conditions precludes the ability to confidently identify a functional relation.
- Overlap: The degree to which data are similar in level across conditions can be evaluated by describing the degree to which data overlap between conditions. For example, in Figure 2, a single data point in the intervention condition overlaps with baseline values; this results in a high degree of non-overlapping data. Generally, more overlap suggests a functional relation does not exist or results in less confident conclusions.
- Consistency: The degree to which data are similar *within conditions* and the degree to which behavior change is similar *across condition changes*.
- Immediacy: Confidence in the presence of a functional relation is improved when behavior change occurs simultaneously with condition changes. However, if delayed changes are (a) consistent across conditions, and (b) expected a priori, they are less problematic.

Synthesizing Results across Studies

Evidence-based practices are identified by synthesizing results across multiple studies. SCR syntheses involve evaluating bodies of literature to (1) identify for whom and under what conditions an intervention is effective, (2) describe interventions designed to evaluate specific behaviors (e.g., problem behavior), or (3) examine if sufficient evidence exists to conclude an intervention is an evidence-based practice. SCR synthesis should include narrative descriptions of the participants, settings, implementers, and intervention characteristics as well as evaluations of the quality and rigor *and* outcomes (Ledford & Gast, 2018; Pustejovsky & Ferron, 2017).

Narrative descriptions. Narrative descriptions of SCR should be summarized across studies to describe for whom and under what conditions the interventions were evaluated. Descriptive information, including descriptions of intervention implementers' background and required training, may assist readers in understanding the necessary qualifications for applying interventions in practical settings (e.g., schools and community settings). This information may be easily displayed in tables to allow the reader to synthesize common themes across studies (see Lloyd, Weaver, & Staubitz, 2016).

Quality and rigor. After describing for whom and under what conditions an intervention has been conducted in the existing literature, syntheses should evaluate the quality and rigor of SCR to determine if designs were *experimental* in nature and to what extent studies were rigorous in their evaluations of the intervention. Evaluations of the quality and rigor of SCR are often guided by existing quality evaluation tools such as quality indicators (CEC, 2014), standards (WWC, 2017), and frameworks (Goldstein, Lackey, & Schneider, 2014). Selection of an appropriate tool can be aided by reading reviews comparing available tools (see Maggin, Briesch, Chafouleas, Ferguson, & Clark, 2014). Quality and rigor evaluations should at minimum include evaluations of (1) sufficient potential demonstrations of effect (at least three at three different points in time), (2) reliability of dependent variables, and (3) measurement of fidelity of independent variable implementation *across all conditions in the design*.

Outcomes. Finally, outcomes should be evaluated in the context of *experimental* SCR designs that meet sufficient quality and rigor guidelines. Failure to limit outcome evaluation to quality SCR designs may reduce confidence that results are attributable to the intervention rather than other factors (e.g., changes in the participants' environments outside of the study context or failure of the authors to adequately control for threats to internal validity; Ledford & Gast, 2018). First, outcomes should be evaluated using systematic visual analysis procedures (see above and Ledford, Lane, & Severini, 2018. Quantitative metrics can also be used to describe the magnitude of behavior change. Visual analysis evaluates components such as consistency and immediacy of change across the entire design, whereas quantitative metrics evaluate each AB comparison separately within the design. Quantitative metrics used to evaluate outcomes include non-overlap measures, parametric effect sizes, and hierarchical linear models. A comprehensive summary of quantitative metrics that can be used to synthesize SCR is beyond the scope of this guide. We suggest readers consult Pustejovsky and Ferron (2017) for a comprehensive evaluation and comparison of quantitative metrics.

Summary
Single case research designs are often used to evaluate the effectiveness of interventions for improving behaviors of children with disabilities. Three primary groups of designs are characterized by the way conditions are ordered—time-lagged implementation, sequential introduction and withdrawal, and rapid iterative alternation. Regardless of design type, SCR studies should include (1) an adequate number of potential demonstrations of effect (usually three), (2) a sufficient number of data points in each experimental condition to establish level, trend, and variability; (3) reliability of dependent variable measurement, and (4) accurate implementation of procedures (i.e., fidelity). Outcomes are assessed via visual analysis, using several data characteristics—level, trend, variability, overlap, consistency, and immediacy. Like RCTs, high-quality SCR designs can be used to evaluate single interventions. In addition, evidence-based practices can be established given multiple high-quality SCR studies with positive outcomes.

References

Asmus, J. M., Carter, E. W., Moss, C. K., Biggs, E. E., Bolt, D. M., Born, T. L., ... & Weir, K.. (2017). Efficacy and social validity of peer network interventions for high school students with severe disabilities. *American Journal on Intellectual and Developmental Disabilities*, *122*, 118-137.

Barton, E. E., Ledford, J. R., Lane, J. D., Decker, J., Germansky, S. E., Hemmeter, M. L., & Kaiser, A. (2016). The iterative use of single case research designs to advance the science of EI/ECSE. *Topics in Early Childhood Special Education*, *36*, 4-14.

Council for Exceptional Children. (2014). Council for Exceptional Children standards for evidence-based practices in special education. Retrieved from: https://www.cec.sped.org/Standards/Evidence-Based-Practice-Resources-Original

Fuchs, L. S., Seethaler, P. M., Powell, S. R., Fuchs, D., Hamlett, C. L., & Fletcher, J. M. (2008). Effects of preventative tutoring on the mathematical problem solving of third-grade students with math and reading difficulties. *Exceptional Children*, *74*, 155-173.

Garfinkle, A. N., & Schwartz, I. S. (2002). Peer imitation: Increasing social interactions in children with autism and other developmental disabilities in inclusive preschool classrooms. *Topics in Early Childhood Special Education*, *22*, 26-38.

Goldstein, H., Lackey, K. C., & Schneider, N. J. B. (2014). A new framework for systematic reviews: Application to social skills interventions for preschoolers with autism. *Exceptional Children, 80*, 262-286.

Hammond, D., & Gast, D. L. (2010). Descriptive analysis of single subject research designs: 1983—2007. *Education and Training in Autism and Developmental Disabilities*, 187-202.

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

Kamps, D., Wills, H. P., Heitzman-Powell, L., Laylin, J., Szoke, C., Petrillo, T., & Culey, A. (2011). Class-Wide Function-Related Intervention Teams: Effects of group contingency programs in urban classrooms. *Journal of Positive Behavior Interventions, 13*, 154-167.

Ledford, J. R. (2017). No randomization? No problem. Experimental control and random assignment in single case research. *American Journal of Evaluation. doi:* 10.1177/1098214017723110

Ledford, J. R., Barton, E. E. Severini, K. E., & Zimmerman, K. N. (in press). A primer on single case research designs: Contemporary use and analysis. *American Journal of Intellectual and Developmental Disabilities.*

Ledford, J. R. & Gast, D. L. (2018). *Single case research methodology: Applications in special education and behavioral sciences*. New York, NY: Routledge.

Ledford, J. R., Lane, J. D., & Severini, K. E. (2017). Systematic Use of Visual Analysis for Assessing Outcomes in Single Case Design Studies. *Brain Impairment*. Advance online publication. https://doi.org/10.1017/BrImp.2017.16

Lemons, C. J., King, S. A., Davidson, K. A., Puranik, C. S., Al Otaiba, S., …Fidler, D. J. (2017). Developing an early reading intervention aligned with the Down syndrome behavioral phenotype. *Focus on Autism and other Developmental Disabilities.* doi: 10.1177/1088357615618941

Lloyd, B. P., Weaver, E. S., & Staubitz, J. L. (2016). A review of functional analysis methods conducted in public school classroom settings. *Journal of Behavioral Education*, *25*, 324-

356.
Maggin, D. M., Briesch, A. M., Chafouleas, S. M., Ferguson, T. D., & Clark, C. (2014). A comparison of rubrics for identifying empirically supported practices with single-case research. *Journal of Behavioral Education, 23*, 287-311.
Pustejovsky, J. E., & Ferron, J. M. (2017). Research synthesis and meta-analysis of single-case designs. In J. M. Kauffman, D. P. Hallahan, & P. C. Pullen (Eds.), *Handbook of Special Education*, 2nd Edition. New York, NY: Routledge.
Schumaker, J., & Sherman, J. A. (1970). Training generative verb usage by imitation and reinforcement procedures. *Journal of Applied Behavior Analysis, 3*, 273-287.
Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs*. Belmont, CA: Wadsworth Cengage Learning.
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). Washington, DC: National Center for Education Research, Institute of Education Sciences, U. S. Department of Education. Available from http://ies.ed.gov/.
Sreckovic, M. A., Hume, K., & Able, H. (2017). Examining the efficacy of peer network interventions on the social interactions of high school students with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 47.* doi: 10.1007/s10803-017-3171-8
Tankersley, M., Kamps, D., Mancina, C., & Weidinger, D. (1996). Social interventions for Head Start children with behavioral risks: Implementation and outcomes. *Journal of Emotional and Behavioral Disorders, 4*, 171-181.
What Works Clearinghouse, Institute of Education Sciences. (2017). *Procedures handbook, version 4.0.* Retrieved from https://ies.ed.gov/ncee/wwc/Docs/ReferenceResources/wwc_procedures_handbook_v4_draft.pdf

Table 1

Descriptions of Studies Using Randomized Controlled Trials and Single Case Designs to Answer Questions in the Same Domain

Domain	Randomized Controlled Trial	Single Case Design
Improving imitation skills for children with autism	Assessed the effects of imitation training for children with autism via two imitation assessments, compared to a no-intervention control group (Ingersoll, 2010)	Assessed the effects of a peer-mediated intervention for improving imitation for young children measured via direct observation of baseline and treatment sessions (Garfinkle & Schwartz, 2002)
Improving academic outcomes for children with learning and intellectual disabilities	Assessed the effects of peer tutoring on word problem performance, measured by standardized and researcher-developed tests, compared to a business-as-usual control group (Fuchs et al., 2008)	Assessed the effects of a researcher-devised reading intervention for improving reading behaviors of children with Down syndrome, measured via counts of correct responding during probe sessions during baseline and treatment conditions (Lemons et al., 2017)
Improving social interactions for individuals with intellectual disabilities	Assessed the effects of peer social networks on social interactions, measured via a standardized teacher report, compared to a business-as-usual control group (Asmus et al., 2017)	Assessed the effects of peer social networks on social interactions, measured via direct observation of baseline and treatment sessions (Sreckovic, Hume, & Able, 2017)
Decreasing problem behavior for children with behavior disorders	Assessed the effects of a school-based social skills and positive interactions prevention program, measured via direct observation, compared to a business-as-usual control group (Tankersley, Kamps, Mancina, & Weidinger 1996)	Assessed the effects of CW-FIT on disruptive behaviors via direct observation of baseline and treatment sessions (Kamps et al., 2011)

Table 2

Basic Rigor Requirements for Single Case Designs

Domain	**Requirement**	**Examples**	**Non-Examples**
Replication	Three potential demonstrations of effect between identical conditions (e.g., A to B)	• Multiple baseline design with three tiers and three separate start points • A-B-A-B design • B-C-B-C design	• Multiple baseline design with two tiers or three tiers but only two start points • A-B-A-C design
Data Sufficiency	Sufficient data to identify level, trend, and variability *and* to evaluate change between conditions	• Three data points in a baseline condition at floor levels (e.g., 0% correct) • Six data points in a moderately variable baseline condition with no apparent trend	• Three data points in a baseline condition with some variability (e.g., 0%, 20%, 10% correct) • Six data points in a baseline condition with a therapeutic or contratherapeutic trend (e.g., 0%, 10%, 20%...)
Reliability	Evidence that data were recorded as planned and align with behavior occurrence	• Agreement data collected frequently in baseline and intervention conditions • High interobserver agreement or Kappa coefficient	• Agreement data collected in only one condition • Agreement data collected infrequently (e.g., <20% of sessions) • Low interobserver agreement scores (e.g., <80 or 90%, depending on dependent variable complexity)
Fidelity	Evidence that experimental conditions were carried out as planned	• Fidelity data collected frequently in baseline and intervention conditions • High fidelity to planned procedures	• Fidelity data collected only during intervention conditions (sometimes referred to as treatment integrity) • Low fidelity to planned procedures

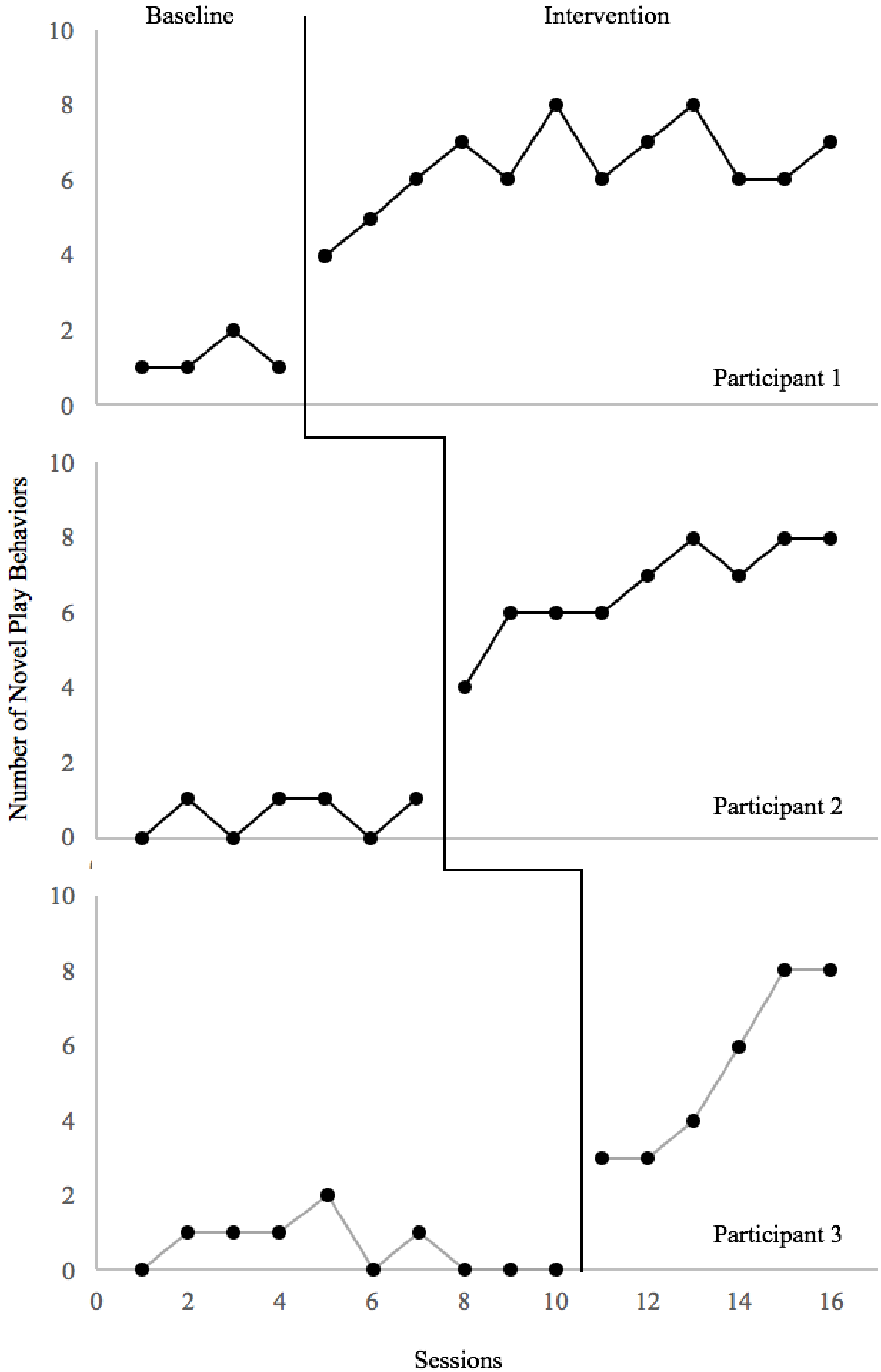

Figure 1. Multiple baseline across participants design.

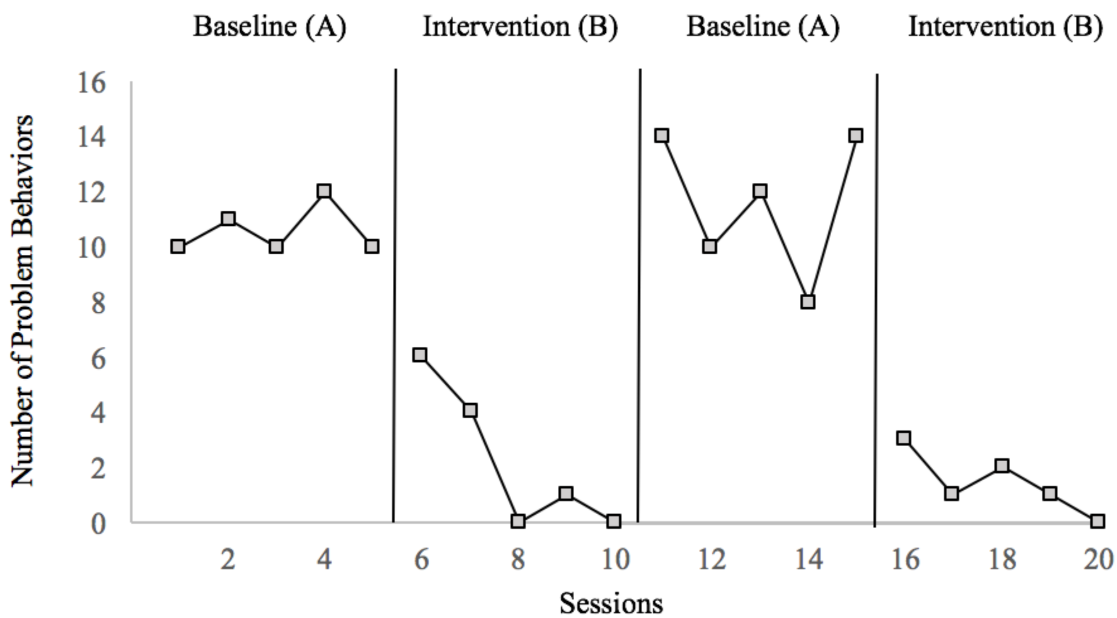

Figure 2. A-B-A-B design.

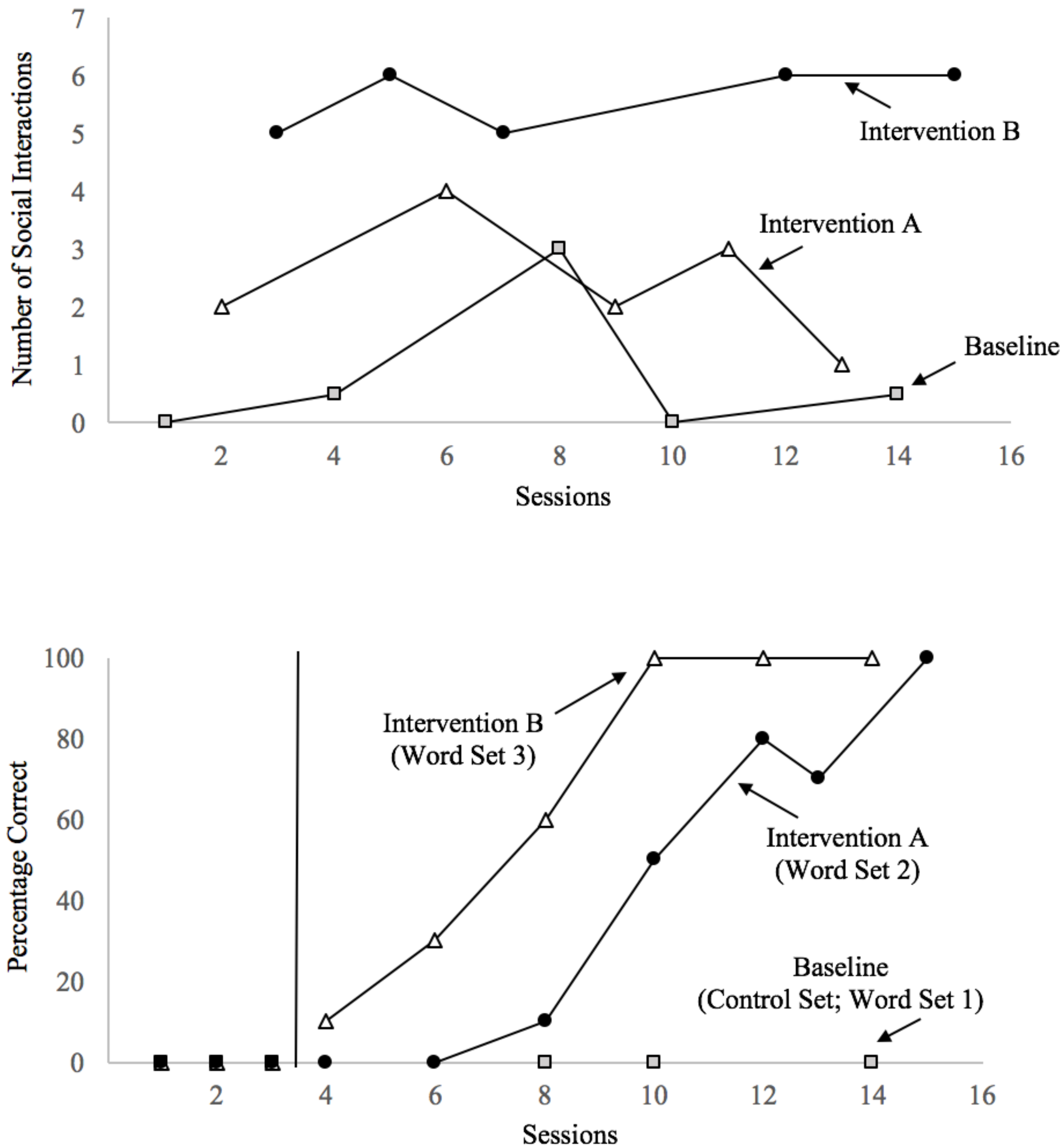

Figure 3. Alternating treatments design (top) and adapted alternating treatments design (bottom).

