Single Case Research Methodology

Applications in Special Education and Behavioral Sciences

FOURTH EDITION

Edited by
Jennifer R. Ledford and David L. Gast

Single Case Research Methodology

The fourth edition of this bestselling text provides a comprehensive discussion of single case research methodology, with updated information throughout the book, including new content on design types, design selection, social validity, fidelity, generality, visual analysis, and writing. Students, researchers, and practitioners can use this detailed reference tool to conduct single case research design studies; interpret findings of single case design studies; and write proposals, manuscripts, or systematic reviews of single case methodology research. The new text features updates relevant to contemporary guidelines about single case research and includes examples of recent and historical studies in education and behavioral sciences.

Jennifer R. Ledford is an Associate Professor in the Department of Special Education at Vanderbilt University.

David L. Gast is Professor Emeritus of Special Education in the Department of Communication Sciences and Special Education at the University of Georgia.

Single Case Research Methodology

Applications in Special Education and Behavioral Sciences

Fourth Edition

Edited by Jennifer R. Ledford and David L. Gast

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Contents

Preface

This 4th edition of *Single Case Research Methodology* was edited to include information regarding contemporary developments in single case design, while retaining an emphasis on lessons learned from more than 50 years of work by early single case research scholars, including the work of Dr. David Gast, the driving force behind this text (first published in 2010) and its predecessor, *Single Subject Research in Special Education* (along with Dr. James Tawney, 1984). His work began at the University of Kansas Department of Human Development and Family Life, where he worked among some of the preeminent early behavioral researchers, including Drs. Joseph Spradlin, Sebastian Striefel, James Sherman, Donald Baer, and Montrose Wolf. He continued the mentorship model, in which professors worked closely alongside graduate students to conduct meaningful applied research, at the University of Kentucky (1975–1989) and then the University of Georgia (1990–2016), where we met and I conducted my first research synthesis and single case experimental design study. It was here first, and then at Vanderbilt University, where I worked with Dr. Mark Wolery, where I was taught the intricacies and importance of single case research design for researchers and practitioners. I continue to be humbled and excited to work with and in the shadow of so many great single case methodology researchers and to help share what I've learned with the next generation of single case scholars.

Our goal in editing this edition, as with the previous editions, is to present a thorough, technically sound, user-friendly, and comprehensive discussion of single case research methodology. We intend for this book to serve as a detailed reference tool for students, researchers, and practitioners who intend to conduct single case studies; interpret findings of these studies; or write proposals, manuscripts, or reviews of single case research. We expect readers will come from a variety of disciplines in social, educational, and behavioral science including special and general education; school, child, clinical, and neuropsychology; speech, occupational, recreation, and physical therapy; and social work. In the book, we present a variety of single case research studies with a wide range of participants, for a range of purposes, in various settings. As in previous editions, much of the work originates in the fields of special education and behavior analysis, although, increasingly, other fields are also represented (e.g., school psychology, speech language pathology).

The organization of this edition is somewhat different from previous editions, with 19 chapters instead of the traditional 14. The biggest organizational departure from previous editions is that we discuss research designs according to three primary condition ordering types (sequential introduction and withdrawal, time lagged, and rapid iterative alternation) and have one chapter dedicated to the analysis of each primary type. We added content about generality, validity, fidelity, and choosing design types. We expanded content related to writing about single case design,

including one chapter on writing proposals and reports and a second dedicated to writing systematic reviews. The guidelines presented in this text are intended to assist you in the design, analysis, implementation, and dissemination of meaningful and rigorous single case research. We hope this text helps you to conduct work that matters to you and provides meaningful information to move your field forward—good luck!

Jennifer R. Ledford

Author Bios

Jennifer R. Ledford is Assistant Professor at Vanderbilt University. She was introduced to single case design by Dr. David Gast at the University of Georgia and was further mentored by Dr. Mark Wolery during her doctoral program at Vanderbilt University. Her research interests include improving the use and synthesis of single case design research and instructional practices for young children with disabilities in classroom settings. She urges single case researchers to remember that you cannot answer all the interesting questions in a single study—ask and answer one question well and you'll have done a service to the field.

David L. Gast is Professor Emeritus at the University of Georgia. He earned his doctorate from the University of Kansas in 1975. Prior to joining the faculty at the University of Georgia he collaborated extensively with Dr. Mark Wolery at the University of Kentucky (1975–1989). His research interests include use of errorless instructional strategies and use of single case experimental designs to evaluate clinical and educational practices. Most of his studies used multiple probe designs across behaviors because of their practicality when conducting research in applied settings. With this and other single case experimental designs, the effectiveness of an intervention can be determined early in the study by monitoring data trends and, if necessary, making modifications to the independent variable to achieve the educational or clinical objective.

Kevin M. Ayres is Professor at the University of Georgia where he also received his doctoral degree. He was introduced to single case design by Dr. David Gast at the University of Georgia. Dr. Ayres's current research interests primarily include focus on evaluations of behavior analytic practices in classroom contexts as well as evaluating parameters of reinforcement as they relate to response allocation. He likes to remind his students that experiments require careful and thoughtful choices that are steered by a mix of experience and a risk reward calculus. In the end, even if the experiment does not "work," a good scientist still learns something.

Kathryn M. Bailey is a doctoral student in Special Education at Vanderbilt University. She was mentored in single case design by her advisors, Joseph Lambert and Ann Kaiser, during her master's and doctoral program. Kathryn's research interests include the relations between language and behavioral development, language intervention for children with disabilities, and caregiver-mediated interventions. Her favorite part of single case design is the ongoing process of blending scientific rigor with practical considerations for children with disabilities and their caregivers.

Erin E. Barton is the owner and Lead Consultant at Barton Consulting, LLC. She received her doctoral degree from Vanderbilt University. Mark Wolery was her advisor and introduced her to single case design. Her research interests include identifying practices that support young children's full participation in the settings in which they live and play. Her favorite aspect of conducting single case research is the dynamic nature of the designs.

Eric Alan Common is Associate Professor at the University of Michigan-Flint. He received his doctoral degree from the University of Kansas. His research focuses on social-emotional and behavior preventive and intervention in educational settings delivered through Comprehensive, Integrated, Three-Tiered (Ci3T) Models of Prevention and school-based applied behavior analysis. His favorite single case design is the changing criterion design because he finds the research questions and procedures innovative, inherently humanistic, and socially just.

Tara Fahmie is Associate Professor at the University of Nebraska Medical Center. She received her doctoral degree from the University of Florida and became passionate about single case design under the teachings and mentorship of Drs. Hank Pennypacker, Gregory Hanley, and Brian Iwata, among others. Her research interests include the prevention, assessment, and treatment of severe problem behavior in individuals with intellectual and developmental disabilities. Her two favorite things about single case design are that it promotes an understanding of individual variation and allows moment-to-moment decision making in service of personalized goals.

Joseph M. Lambert is Assistant Professor in the Department of Special Education at Vanderbilt University. He received his doctoral degree from Utah State University and was introduced to single case design by Drs. Bill Ahearn and Sarah Bloom. His research interests include practitioner training, functional analysis, function-based interventions, and environmental factors that influence the maintenance of these interventions. His favorite single case design is the multielement-alternating treatments design (ME-ATD) because of its efficiency. His favorite part of conducting single case design is the curiosity it requires. Specifically, the process of contemplating why people do what they do.

Justin D. Lane is Associate Professor of Special Education at the University of Kentucky. Mark Wolery introduced him to single case design at Vanderbilt University and later David Gast at the University of Georgia. His research interests include evaluating and refining interventions for young children with complex communication needs and coaching parents and educators to embed naturalistic language interventions in play. His advice for students planning a single case design study is to remember the saying, "Clear is kind." That is, students should focus on clearly describing their amazing ideas so others can learn from and replicate their work with precision.

Kathleen Lynne Lane is a Roy A. Roberts Distinguished Professor in the Department of Special Education at the University of Kansas and Associate Vice Chancellor for Research. She earned her doctoral degree from University of California, Riverside. Her research interests focus on designing, implementing, and evaluating Comprehensive, Integrated, Three-tiered (Ci3T) models of prevention to (a) prevent the development of learning, behavior, and social and emotional well-being challenges and (b) respond to existing instances, with an emphasis on systematic screening. She appreciates the beauty of single case research design: The rigor, flexibility, and transparency.

Blair P. Lloyd is Associate Professor of Special Education at Peabody College, Vanderbilt University (where she also received her doctoral degree). Her research interests include school-based behavioral assessment and intervention methods for students with emotional/behavioral disorders; the intersection of Applied Behavior Analysis and School Mental Health; and issues in observational measurement. Blair's favorite part of single case research is its emphasis on transparent visual data display—and how one well-constructed graph can tell the story of an experiment from beginning to end.

Hedda Meadan-Kaplansky is the Margaret Valpey Professor in the Department of Special Education and Health Innovation Professor in the Carle Illinois College of Medicine at the University of Illinois Urbana Champaign. She received her doctoral degree from the University of Illinois. Her research and scholarship address the social communication behavior of young children with disabilities, including design and testing of interventions to enhance children's communicative functioning and caregivers' and professionals' use of recommended practices. A word of advice she would give students planning single case design study is to give special attention to and assess, through multiple methods and sources, the social validity of their intervention.

Natalie S. Pak is an Assistant Professor in the Department of Communication Sciences and Disorders at the University of South Florida. She received her doctoral degree from Vanderbilt University. Her research interests include family-centered early language interventions for children who are dual language learners and/or who use augmentative and alternative communication. Her favorite part of single case design research is the ability to see the data over time for every participant.

Amy D. Spriggs is Associate Professor at the University of Kentucky. She received her doctoral degree from University of Georgia where she was introduced to single case design by David Gast. Her research interests include using everyday technology to increase independence in individuals with significant support needs, promoting generalization, empirically evaluating generalization, and evaluating the rigor of single case research. A word of advice she would give to students planning a single case design study is to ask questions. This is a collaborative field and two heads are always better than one!

Robyn Tate is Emeritus Professor at the University of Sydney, Australia. She received her doctoral degree from the University of Newcastle, Australia. Her research interests include the methodology of single case designs and evidence-based clinical practice in people with neurological conditions. Her favorite single case design is probably the concurrent multiple-baseline design across behaviors/settings because it is appropriate for both withdrawable and non-withdrawable interventions and more feasible in the clinical setting than concurrent designs. A word of advice she would give to students planning a single case design is to use field notes at the implementation stage to document unforeseen events that might influence outcomes.

Kathleen N. Tuck is Assistant Professor of Special Education at the University of Kansas. She received her doctoral degree from Vanderbilt University. Her research interests include supporting educator implementation of evidence-based instructional practices to promote the engagement of children with language and behavioral support needs in inclusive classrooms and enhancing the methods and measurement of single case research design studies and syntheses. Her favorite part of conducting single case design research is the ability to rigorously and flexibly adapt interventions and designs based on child or educator preferences and experiences.

Katie Wolfe is Associate Professor at University of South Carolina. She received her doctoral degree from Utah State University. Her research interests include data-based decision making, visual analysis, and parent training. A word of advice she would give to students planning a single case design study is to be flexible—things probably won't go according to plan!

Section 1

Introduction to Research and Measurement

Before conducting a single case design study, you must understand the purposes and foundations of research in general, and the importance of reliable, accurate, and valid measurement. The first section of the book focuses on these topics.

In this section, we begin with a broad introduction to research in general in Chapter 1. We briefly discuss different research paradigms (qualitative, between-groups, single-case) and different approaches to conducting single case research (inductive/deductive, dynamic/static, process/procedure). The overarching questions answered in this chapter are: (1) What is research? (2) What are the different approaches to doing it?

Then, in Chapter 2, we discuss replication and how we draw conclusions across multiple experiments. Questions relevant for this chapter are: (1) What is external validity, and how is it different than ecological and construct validity? (2) What are the different types of replication, and how do they relate to internal versus external validity? (3) What factors should you consider when you are conducting an experimental replication?

In Chapter 3, we focus in on single case research specifically, by identifying confounding factors (i.e., threats to internal validity) whose impacts we must attempt to mitigate when we conduct single case research studies. Questions critical for this content include: (1) What is internal validity, and how is that related to identifying functional relations? (2) What are several major threats to internal validity, and how do we control for and/or detect them?

Chapter 4 focuses on the "why" and "how" of measuring behavior. We discuss continuous and discontinuous methods, how to choose a system appropriate for the dimension of interest, and the strengths and weaknesses of major systems. Questions you should be able to answer after reading this chapter include: (1) What are some examples of reversible versus nonreversible behaviors, long- versus short-duration behaviors, and free operant versus trial-based behaviors? (2) What are the major dimensions of interest, and how can you (a) measure them with continuous systems and (b) estimate them with discontinuous systems?

Next, we focus on the validity, accuracy, and reliability of data collection in Chapter 5. Much of the chapter is dedicated to interobserver agreement, the primary method by which single case researchers evaluate reliability. Questions relevant for this chapter include: (1) What are some steps for ensuring data reliability, before, during, and after measurement? (2) What are the differences between point-by-point and gross agreement? (3) Why do we collect interobserver agreement (IOA) data?

DOI: 10.4324/9781003294726-1

In Chapter 6, we focus on planning and ensuring reliability of experimental procedures. We discuss using the theory of change to identify active ingredients and likely impacts of your intervention, ways to conceptualize independent variables in complex experiments, and the importance of measuring independent variable implementation. Critical questions from this chapter include: (1) Why do you need to understand your theory of change? (2) What is the difference between procedural fidelity and treatment fidelity, which is superior, and why? (3) What is the difference in purpose for formative and summative analysis of fidelity?

The content in Chapter 7 is focused on generality of behavior change. We discuss generality and social validity and explain the importance of using multiple sources of data for evaluating outcomes of single case research. We discuss the complexity of the typical use of terms "generalization" and "maintenance." Questions you should be able to answer after reading this chapter are: (1) What are generality and social validity, and why are they important? (2) What are some relevant ways to measure generalization? (3) What are the different ways in which the term "maintenance" is used? (4) How should you know if maintenance and generalization would be expected in your study?

1
Research Approaches

David L. Gast and Jennifer R. Ledford

Important Terms

basic research, applied research, independent variables, dependent variables, participants, participatory action research, experimental research, non-experimental research, evidence-based practice, nomothetic, idiographic, dynamic research, static research, baseline, intervention

Table of Contents

Callout

DOI: 10.4324/9781003294726-2

The goal of science is to advance knowledge. One process by which we advance knowledge is via research—the systematic investigation and manipulation of variables to identify associations and understand processes. The effectiveness of educational and behavioral interventions is dependent on the use of evidence derived from research, but this process is not straightforward. This complexity results in a "research to practice gap" problem that suggests that research outcomes are not necessarily directly applicable to problems of practice. For example, outcomes of research studies have been reported to be non-replicable (Open Science Collaboration, 2015); to be dependent on counterfactual conditions (Lemons et al., 2014); to fail to generalize to outside of research contexts, in applied or authentic settings (Spriggs et al., 2016); and to be largely inapplicable to "real" problems faced by practitioners (Snow, 2014). How then does research contribute to the advancement of knowledge, and does it do so in a useful manner? In this chapter, we introduce the concepts of applied research and evidence-based practice, describe different ways to characterize research, and explain several research approaches and their corresponding rationales and assumptions.

Applied Research

Basic research is concerned with the advancement of knowledge that may or may not have immediate and specific application to practical concerns. **Applied research** involves systematic investigation related to the pursuit of knowledge in practical realms or to solve real-world problems. For example, a great deal of basic research has demonstrated that "resurgence" (i.e., the return of previously eliminated behavior when reinforcement conditions for current behavior worsen) consistently occurs across species, including humans (Kimball et al., 2023). The findings of this body of basic research guided translational researchers to demonstrate how the phenomenon could manifest following effective intervention; for example, when challenging behavior resurges after communicative responses are not reinforced (e.g., Volkert et al., 2009). Because translational research paradigms demonstrated the phenomenon's relevance to socially important situations (i.e., the assessment and treatment of challenging behavior), applied researchers have now begun to develop procedures intended to "inoculate" treatment outcomes against the environmental determinants of resurgence (Banerjee et al., 2022; Bloom & Lambert, 2015; Fuhrman et al., 2021; Neely et al., 2020).

In applied research evaluated with single case designs, we are most interested in determining the relation between **independent variables**—the variables manipulated by researchers (i.e., intervention) and **dependent variables**—the variables we expect to change given the manipulation (e.g., percentage of time engaged in challenging behavior), to solve problems of practice. For example, we might evaluate the impact of a specific coaching model (the independent variable) on the percentage of correct implementation of a naturalistic intervention (the dependent variable; Quinn et al., 2021) or the impact of using a differential reinforcement intervention package (the independent variable) for increasing time spent in a nonpreferred environment to reduce elopement (the dependent variable; Lambert et al., 2017). This book is primarily focused on the use of single case design in applied research, although single case designs can also be used to evaluate basic and translational questions. In applied research, we refer to the people who choose to participate in the study as **participants**, although historically they were referred to as subjects—because we assume that participants *willingly volunteer* rather than being *subjects* of the studies (Boynton, 1998).

Integrating Science into Educational and Clinical Practice

The purpose of a research project is to produce generalizable knowledge about relations between variables that can be used by others to expand understanding of a given phenomenon.

Applied research also has this goal, although some difficulties may arise when attempting to balance this goal with the additional goal of improving outcomes for participants. That is, the goal to establish confidence in relations is sometimes at odds with providing the best services for a given participant. One example is that we may need to collect a significant amount of data in non-intervention contexts to increase experimental rigor and confidence in the relation between the independent and dependent variables. However, increasing the amount of time a participant spends in counter therapeutic environments is generally objectionable when considering only the participant's best interests. We will discuss the ethical ramifications of decisions like this in Chapter 16. We refer to practitioners who engage in research as scientist-practitioners (a label coined by Barlow and colleagues in 1984 to describe interventionists who make data-based decisions an integral part of their practice).

When research is conducted under highly controlled conditions, as is often the case in studies using single case designs, the ability of those working in "typical" or "authentic" community settings to replicate conditions may be unclear. That is, interventions found to be effective in resource-rich controlled settings may not be able to be carried out at the same level of fidelity, thus affecting the outcome of the intervention. Snow (2014) suggested educational research should include more collaboration with practitioners, to address applied problems and enhance the pertinence of research. This position is not new, and that single case designs are particularly well suited to answer these applied problems has been acknowledged for decades (Barlow et al., 1984; Borg, 1981; Odom, 1988; Tawney & Gast, 1984). We suggest that there are important similarities between practice and research, and provide guidance for practitioners interested in conducting research in applied settings in Callout 1.1.

Callout 1.1 Questions for Scientist-Practitioners

Successful practitioners must demonstrate that they can bring about positive behavior change in their clients. Practitioners who collect data on client or student behavior in response to the treatments they implement can show behavior change that occurs over time. However, sometimes behavior change may be the result of other factors, rather than the treatment itself (e.g., other treatments or experiences). The utilization of single case designs allows practitioners to go one step further than showing behavior change—to establish a causal link between his or her practices and the child's behavior change. That is, single case designs can help practitioners increase the confidence that the treatment they implemented *and only that treatment* caused behavior to change. Given the potential for single case design to enhance conclusions drawn by practitioners, and the guidelines suggesting practitioners use scientific evidence and data-based decision making, the use of the scientist-practitioner model is valuable. Given the potential difficulties, and to ensure that a planned applied project has not only scientific value, but also practical value for participants, Eiserman and Behl (1992) suggested considerations for special educators conducting research, which we have summarized and adapted below:

1. Is the dependent variable meaningfully related to educational or therapeutic goals?
2. Does available research suggest that the participant is likely to benefit?
3. Are procedures and goals in line with the policies of the institution and with objectives of the clients (e.g., school, clinic, family)?
4. Is the answer to the question of interest to all relevant stakeholders?
5. Can the research be completed given the participants' other scheduled activities, or is there an acceptable modification of these activities that allows for completion?

6. How does the research interact with ongoing therapeutic, educational, or leisure activities?
7. How much time is required for participation, and is that amount of time justifiable, given participant needs?
8. Do participants agree to participate, and are they given opportunities to continue to agree (assent) or to withdraw their agreement (dissent)?
9. Do other important stakeholders support participation?
10. Are personnel and materials resources sufficient for the project?
11. Are there any ethical concerns that should be addressed prior to initiation of the project?

Participatory Action Research

One framework that might be particularly helpful for considering the relationship between applied research and real-world contexts is **participatory action research**. This framework explicitly acknowledges the critical importance of the experiential knowledge and values of people who are to be impacted by any research. While the primary motivation for direct stakeholders (e.g., parents, teachers, children) is to solve a problem that directly impacts them, the primary motivation for researchers is the pursuit of knowledge for the greater good. These differing contingencies are sometimes at odds with one another (Pritchett et al., 2022).

When using this framework, researchers (1) build relationships with those potentially impacted by their research (e.g., teachers and students in schools), (2) generate a common understanding of any problems that exist, (3) generate data about the problem and analyze those data, (4) plan solutions, and (5) take action (Cornish et al., 2023). These steps align quite well with single case research, given the ability of single case research to be dynamic and individualized (thus allowing for steps 1, 2, and 4), the need to collect baseline and intervention data (steps 3 and 5), and the ability to modify action steps (i.e., procedures) as needed (steps 4 and 5).

It is important to note that single case design does not necessarily follow this framework; participatory action research requires that people who are affected by a particular issue take a leading role in the trajectory described above (Kindon et al., 2007). Researchers serve as methodology experts but acknowledge the content and experiential expertise of other stakeholders. A project that is designed by researchers to answer a question that they are interested in may include important steps during which they get stakeholder feedback but would not be considered participatory action research. There is a large continuum of research approaches that range from entirely researcher-driven and with little participant input to entirely practice-driven with researchers serving only as methodological experts. Many structures (e.g., funding agencies) uphold a more researcher-driven research process, which requires that questions and processes are well-described and justified prior to beginning a project. Although there is no "wrong way" to do research, it is important to understand that the extent to which research matters in the real world may be considerably impacted by decisions about the ways in which non-researcher stakeholders are included in the process.

Evidence-Based Practice

Guidelines in the Individuals with Disabilities Education Improvement Act (IDEIA) and Every Student Succeeds Act (ESSA) mandate the use of evidence-based practice (alternately,

"scientific, research-based intervention"; IDEIA; or "empirically supported practice"; Ayres et al., 2011). Similarly, professional organizations like the American Psychological Association (APA), American Speech-Language Hearing Association (ASHA), and the Behavior Analysis Certification Board (BACB) have standards requiring the use of evidence-based interventions. Though the term is relatively new, the idea that research should guide practice is not, particularly in the field of applied behavior analysis (cf., Baer et al., 1968). A narrow view of **evidence-based practice** (Spencer et al., 2012) refers to intervention procedures that have been scientifically verified as being effective for changing a specific behavior of interest, under given conditions, and for particular participants (Horner et al., 2005; Steinbrenner et al., 2020). The APA (2005) adds to its definition of evidence-based practice the integration of research evidence with clinical expertise and context (e.g., the preferences and values of stakeholders). Similarly, ASHA adds both clinical expertise and evidence—which can refer to both external evidence (e.g., published research) and internal evidence (e.g., data gathered from a particular case). Even reviews that have focused on identified practices as evidence-based due to scientific support have acknowledged that the identification of effective practices must take into consideration context and the expertise of practitioners (e.g., Steinbrenner et al., 2020).

In behavioral sciences, "trustworthiness" or credibility of research findings is based on the rigor of the scientific method employed and the extent to which the research design controls for alternative explanations. The scientific method requires investigator objectivity, reliability of measurement, and independent replication of findings. Given the components described above, quantitative research, including single case research (as well as other types of scientific inquiry) is *necessary but not sufficient* for identifying evidence-based practices. Moreover, different research questions or objectives require different research approaches—no one research method or design is appropriate for answering all research questions, and research evidence must be synthesized with clinical expertise and client values to improve implementation of evidence-based practice in typical contexts.

Characterizing Designs

Research designs can be categorized according to a range of features, including attributions of causality, assumptions about generalizability, and flexibility of procedural rules (e.g., ability to change based on response to intervention). Designs can also be grouped by general approach, including several quantitative approaches as well as qualitative methods. We briefly describe these categorizations below.

Attributions of Causality

The act of intentionally manipulating an environmental variable to see if there is a measurable change in some outcome while controlling for other probable reasons for change differentiates **experimental research** from **non-experimental research**. Appropriately utilized single case designs can be categorized as experimental, in addition to group comparison approaches, such as randomized controlled trials (Horner et al., 2005; Ledford et al., 2023). Experimental studies include (1) descriptions of the target behavior(s), (2) predictions regarding what impact the independent variable will have on the dependent variable(s), and (3) appropriate tests to see if the prediction is correct. One characteristic that differentiates an experimental design study from a non-experimental design study is the extent to which the design controls for threats to internal validity—variables other than the planned independent variable that could result in changes in the dependent variable. Correlational, descriptive, and qualitative research designs

are non-experimental. Only single case designs that adequately control for likely threats to internal validity can be considered to be experimental.

Assumptions about Generalizability

Nomothetic research approaches are generally based in the natural sciences and are characterized by attempting to explain associations that can be generalized to a population with certain characteristics. **Idiographic research approaches**, common in the humanities, attempt to specify associations that vary based on certain characteristics or contingencies present for the participant or case of interest. Both nomothetic and idiographic approaches are valid, depending on the research question of interest (Ottenbacher, 1984). Traditional group design approaches generally apply assumptions applicable to nomothetic research while traditional qualitative approaches generally apply assumptions applicable to idiographic approaches. Single case design was historically considered to be idiographic in nature, but increasingly has been used to draw conclusions using nomothetic approaches (e.g., hypothesis-testing versus hypothesis-generating data). These terms are not quite synonymous with, but are related to inductive research approaches versus deductive research approaches. When using an **inductive approach**, researchers collect data, analyze patterns, and establish theory. When using a **deductive approach**, researchers begin with a hypothesis, collect data, and analyze those data with an emphasis on whether their hypothesis was correct.

Relatedly, single case often uses a **dynamic research** approach. That is, researchers might alter a planned independent variable based on response to intervention. Other research approaches, including group designs like randomized controlled trials, are generally considered static research approaches. **Static research** approaches generally test a pre-determined hypothesis, and no changes are made to the approach based on data. In some cases, single case research can also be static—that is, researchers can choose to evaluate an intervention without making data-based changes. The lines between static and dynamic approaches exist in both single case research (e.g., when a researcher sets out to answer a specific question in a static design but makes dynamic changes when the study progresses) and also in group design studies (e.g., SMART designs; Chow & Hampton, 2019). It is important for researchers to determine whether they are using a dynamic or static approach prior to initiation of their study.

Process versus Procedure Questions

Researchers, especially single case researchers, might also consider whether they are interested in evaluating a **process** or a **procedure** when they develop a research study. Differentiating these two types of questions and research approaches is in its infancy but may be likely to inform future work. Evaluating a procedure via single case design tends to be associated with a more static approach, wherein researchers answer the question of "Does procedure A result in changes in behavior B?" Evaluating a process is more associated with a dynamic approach and might answer a question like "How do we use process A to get a desirable change in behavior B?" The latter type of research often involves clinical judgment and modifications based on ongoing data analysis but these decisions can be built into a static research approach (i.e., the decisions are built into the intervention approach). Both types of research are valuable—procedure research may be most easily translated into conclusions about evidence-based practice while process research may contribute to knowledge about underlying processes, variations in implementation required across contexts and participants, and complex interventions requiring

practitioner decision making. In later chapters, we will discuss complexities of conducting and evaluating each of these types of research.

Research Approach

As the book title connotes, the focus of this text is on single case design research methodology and its use by applied researchers in behavioral sciences. Despite this focus on a single type of research design, it is important to be able to compare and contrast research approaches on the basis of their research logic, strategies for controlling for threats to internal validity, and generalization of findings to individual cases. Understanding variability in research approaches will allow you to choose the appropriate type for answering your research questions. As we mentioned previously, no single research approach or design is appropriate for answering all research questions. In the sections that follow, common research approaches and designs are briefly overviewed. More detailed design descriptions and analyses are found elsewhere in such general research methodology texts as Creswell and Clark (2017), deMarrais and Lapan (2004), Farmer et al. (2022), Fraenkel and Wallen (2006) as well as recent methodological guidelines (e.g., Leko et al., 2023; Toste et al., 2023).

Qualitative Research

Qualitative research is generally considered to be ideographic and non-experimental. It does not involve manipulation of an independent variable, but instead is focused on observation of naturally occurring events. Qualitative research approaches provide a detailed, in-depth description of the case under study. The term qualitative research is an "umbrella" term that refers to several approaches with a focus on description rather than quantification of events. Three approaches have particular prominence among educational and clinical researchers who conduct qualitative research studies: Case study, ethnography, and phenomenology. The case study approach entails an in-depth and detailed description of one or more cases, while ethnography refers to the study of a specific cultural group. Both are conducted in a natural setting without an attempt to influence a specific target behavior—thus, although the names are similar, a qualitative *case study* is <u>not</u> the same as an experimental *single case study*. Sometimes confused with ethnography, phenomenology is the study of perceptions of a particular event or situation. Common activities for qualitative research studies include observations, interviews, surveys, and focus groups. For a more in-depth discussion of these and other qualitative research approaches see Glasser and Strauss (1967), Lincoln and Guba (1985), Patton (2014), and a recent article in *Exceptional Children* by the QR Collective (2023).

Descriptive and Correlational Research

Descriptive and correlational research methods are both focused on the quantitative characterization of phenomenon, without any efforts to impact their occurrence. Descriptive research includes quantification of a variable or variables of interest (e.g., To what extent are children with disabilities included in general education classrooms?), while correlational research describes relations between variables (e.g., Does socio-economic status predict the extent to which children with disabilities are included in general education classrooms?). Correlational research is different from both between-group and single case research because the relations are descriptive rather than causal—that is, correlational research allows you to determine that a relation exists between two or more variables but does not allow you to determine whether the relation is *caused* by one of the variables.

Between-Groups Research

Group research approaches are generally considered nomothetic and static. They can be experimental or non-experimental. The basic logic underlying group research is that a large number of individuals are divided and assigned to one of two or more study conditions. In the simplest version, the study includes a control condition, in which participants are not exposed to the independent variable, and treatment condition, in which participants are exposed to the independent variable. Participants could also be equally divided between two treatment groups (e.g., Treatment A and Treatment B). In some group studies more than two conditions may be compared, in which case an equal number of participants would be assigned to each of the conditions (e.g., 30 assigned to control, 30 assigned to Treatment A, 30 assigned to Treatment B). A critical variable to consider when evaluating a group design study is how participants are assigned to study conditions. The optimal method is random assignment of participants (experimental study), but this is not always possible.

The group research approach is the most common research methodology used in some areas of behavioral science and education. Group research designs are well suited for large-scale efficacy studies or clinical trials in which a researcher's interest is in describing whether a practice or policy, on average, will be effective for a specific population. With such research questions a group design methodology is recommended. Numerous designs and statistical analysis procedures are available for your consideration if you choose to study group behavior. Despite its usefulness for detecting average group effects, group comparison designs cannot be generalized to the individual. To paraphrase Barlow et al. (1984), generalization of group research findings to individuals requires a "leap of faith," the extent to which depends on the similarity of the individual to study participants for whom the intervention was effective. You must never lose sight when attempting to generalize a practice supported by group research to an individual, that some participants performed better, while others performed worse than the average participant. Contemporary guidelines for group design studies were recently published in *Exceptional Children* (Toste et al., 2023).

Single Case Research

Single case research can be experimental or non-experimental, ideographic or nomothetic, inductive or deductive, and static or dynamic. Single case design methodology has a long tradition in the behavioral sciences and is commonly used in behavioral sciences, special education, and school psychology (King et al., 2023; Radley et al., 2020; Shepley et al., 2023). Historically, studies using single case designs were referred to as "single subject research," but over time, the term participant replaced subject when humans involved in a study provided informed consent (Pyrczak, 2016); throughout the book we will use the contemporary term participant, although some historical references may include the term subject.

Sidman (1960) described the single case research approach in the authoritative book, *Tactics of Scientific Research*, which exemplified its application within the context of basic experimental psychology research. In 1968, Baer and colleagues elaborated on single case research methodology and how it could be used in applied research to evaluate intervention effectiveness with individuals. Since that time numerous articles, chapters, and books have been written describing single case design methodology and its use in a number of disciplines, including psychology (Bailey & Burch, 2002; Barlow & Hersen, 1984; Johnston & Pennypacker, 1993, 2009; Kazdin, 1998, 2020; Kratochwill & Levin, 1992, 2014; Skinner, 2004), special education (Gast, 2005; Kennedy, 2005; Richards et al., 1999; Tawney & Gast, 1984), occupational therapy (Lane et al., 2017), literacy education (Neuman & McCormick, 1995), communication sciences (McReynolds & Kearns, 1983; Schlosser et al., 2018), and therapeutic recreation (Dattilo et al., 2000).

Single case research is a quantitative experimental approach in which study participants serve as their own control, a principle known as baseline logic (Sidman, 1960). In the simplest single case study, each participant is exposed to both a "control" condition, generally referred to as the baseline condition, and an intervention condition. In a **baseline condition**, participants experience environmental arrangements that are not expected to result in improvements in targeted outcomes; these can be conditions without any intervention (e.g., assessment of whether a child can complete a task with no help) or conditions that represent business-as-usual (e.g., the environment remains unchanged from its typical state, such as measuring engagement during a typical math activity in a classroom). Each participant also usually participates in an **intervention condition** in which they are exposed to environmental arrangements (e.g., contingencies, materials, teaching) that are expected to result in desirable behavior change. There are some variations on this typical set-up, such as when participants are exposed only to assessment conditions in a design (e.g., as in the case of a functional analysis) or when they are exposed to only intervention conditions (e.g., the two conditions being compared are both treatments, rather than one being a treatment and one being a non-treatment condition).

Baseline logic is different from group design logic in which similar or matched participants are assigned to one of two or more study conditions (control *or* intervention). In studies using single case designs, each participant participates in *both* conditions of interest (e.g., baseline and intervention). In group design, posttest data are collected at an a priori specified time point (e.g., after three weeks of intervention), and are analyzed using statistical methods comparing the average performance of participants assigned to one condition to the average performance of participants assigned to other conditions. In single case research, data are collected regularly *while the intervention is occurring*, and intervention conditions are generally continued until a performance criterion is met or until progress is apparent via visual analysis of graphed data (although there are instances where a set time period is used instead). The use of visual analysis of graphic data for individual participants make single case design studies ideal for applied researchers and practitioners who are interested in answering research questions and/or evaluating interventions designed to change the behavior of individuals.

Mixed Methods Approaches

Mixed methods approaches to answering research questions are increasingly popular although their use in special education and behavioral sciences is relatively uncommon (Corr et al., 2021). When using mixed methods, researchers combine multiple quantitative and/or qualitative methods. Several types of mixed methods research have been identified, and include explanatory sequential designs, exploratory sequential designs, and convergent designs (Creswell & Clark, 2017). Explanatory sequential designs begin with a quantitative analysis and are followed by qualitative analyses that provide insight into the "why" and "how" for the quantitative outcomes. One example would be using single case methods to measure whether teachers implemented certain procedures more accurately during intervention conditions as compared to baseline conditions (quantitative) followed by qualitative methods for explaining why the procedures were not implemented as expected. Exploratory sequential analyses follow the opposite sequence—they begin with a qualitative component, which informs a later quantitative component. For example, a researcher might conduct focus groups with teachers, which informs the development of an intervention (qualitative) whose effects are measured with a single case design (quantitative). Convergent designs are used to compare the results of quantitative and qualitative analyses to allow for a more complete picture for a given question. For example, researchers could measure social networks in classrooms via quantitative measurement of interactions during free play, while also using interviews with children where they

are asked to explain who they play with and why. As discussed in Chapter 7, the use of qualitative research methods alongside quantitative single case methods aligns particularly well with social validity questions related to single case research outcomes (Snodgrass et al., 2022). That is, when designed well, mixed methods studies can answer questions that go beyond "Did this intervention work?" (a traditional quantitative single case question) to include "Why did this intervention work?"; "How did participants feel about this intervention?"; and "How can we improve this intervention for current participants and in the future?"

Conclusion

In this chapter, we introduced various ways of categorizing research and described the basis of single case research. We defined a number of important terms that will continue to be used throughout the text, including *dependent variables*, *independent variables*, *baseline logic*, *baseline conditions*, and *intervention conditions*. In later chapters, we will focus on single case research methods, including further expansion on the different types of single case designs that can be used to answer a wide variety of questions across many different contexts.

References

American Psychological Association. (2005). Report of the 2005 presidential task force on evidence-based practice. Washington, DC: Author.

Ayres, K. M., Lowrey, A., Douglas, K. H., & Sievers, C. (2011). I can identify Saturn but I can't brush my teeth: What happens with the curricular focus for students with severe disabilities shifts. *Education and Training in Autism and Developmental Disabilities*, 46, 11–21.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, 1, 91–97.

Bailey, J. S., & Burch, M. R. (2002). *Research methods in applied behavior analysis*. Thousand Oaks, CA: Sage.

Banerjee, I., Lambert, J. M., Copeland, B. A., Paranczak, J. L., Bailey, K. M., & Standish, C. M. (2022). Extending functional communication training to multiple language contexts in bilingual learners with challenging behavior. *Journal of Applied Behavior Analysis*, 55(1), 80–100.

Barlow, D. H., Hayes, S. C., & Nelson, R. O. (1984). *The scientist practitioner: Research accountability in clinical and educational settings*. New York: Pergamon Press.

Barlow, D. H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.

Bloom, S. E., & Lambert, J. M. (2015). Implications for practice: Resurgence and differential reinforcement of alternative responding. *Journal of Applied Behavior Analysis*, 48(4), 781–784.

Borg, W. R. (1981). *Applying educational research: A practical guide for teachers*. New York: Longman.

Boynton, P. M. (1998). People should participate in, not be subjects of, research. *British Medical Journal (BMJ)*, 317(7171), 1521.

Chow, J. C., & Hampton, L. H. (2019). Sequential multiple-assignment randomized trials: Developing and evaluating adaptive interventions in special education. *Remedial and Special Education*, 40(5), 267–276.

Corr, C., Snodgrass, M. R., Love, H., Scott, I. M., Kim, J., & Andrews, L. (2021). Exploring the landscape of published mixed methods research in special education: A systematic review. *Remedial and Special Education*, 42(5), 317–328.

Cornish, F., Breton, N., Moreno-Tabarez, U., Delgado, J., Rua, M., de-Graft Aikins, A., & Hodgetts, D. (2023). Participatory action research. *Nature Reviews Methods Primers*, 3(1), 34.

Creswell, J. W., & Clark, V. L. P. (2017). *Designing and conducting mixed methods research*. Los Angeles: Sage Publications.

Dattilo, J., Gast, D. L., Loy, D. P., & Malley, S. (2000). Use of single-subject research designs in therapeutic recreation. *Therapeutic Recreation Journal*, 34, 253–270.

deMarrais, K., & Lapan, S. D. (Eds.) (2004). *Foundations for research: Methods of inquiry in education and the social sciences*. Mahwah, NJ: Lawrence Erlbaum Associates.

Eiserman, W. D., & Behl, D. (1992). Research participation: Benefits and considerations for the special educator. *Teaching Exceptional Children*, 24, 12–15.
Farmer, T. W., Talbott, E., McMaster, K., Lee, D., & Aceves, T. (Eds.). (2022). *Handbook of Special Education Research, Volume I*. New York: Routledge.
Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education* (6th ed.). New York: McGraw-Hill.
Fuhrman, A. M., Lambert, J. M., & Greer, B. D. (2021). A brief review of expanded-operant treatments for mitigating resurgence. *The Psychological Record*, 1–5.
Gast, D. L. (2005). Single-subject research design. In M. Hersen, G. Sugai, & R. Horner (Eds.), *Encyclopedia of behavior modification and cognitive behavior therapy* (pp. 1520–1526). Thousand Oaks, CA: Sage.
Glasser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago, IL: Aldine.
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.
Johnston, J. M., & Pennypacker, H. S. (1993). *Readings for strategies and tactics of behavioral research*. Lawrence Erlbaum Associates, Inc.
Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). New York: Routledge.
Kazdin, A. E. (1998). *Methodological issues and strategies in clinical research*. Washington, DC: American Psychological Association.
Kazdin, A. E. (2020). *Single-case research designs* (3rd ed). New York: Oxford University Press.
Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Pearson/Allyn and Bacon.
Kimball, R. T., Greer, B. D., Fuhrman, A. M., & Lambert, J. M. (2023). Relapse and its mitigation: Toward behavioral inoculation. *Journal of Applied Behavior Analysis*, 56(2), 259–493.
Kindon, S., Pain, R., & Kesby, M. (2007). *Participatory action research approaches and methods: connecting people, participation and place*. Routledge. Retrieved from: https://eprints.icstudies.org.uk/id/eprint/293/1/LT-19-06-Participatory-Action-Research-Toolkit.pdf
King, S., Wang, L., Nylen, B., & Enders, O. (2023). Prevalence of research design in special education: A survey of peer-reviewed journals. *Remedial and Special Education*, 44(6), 443–505.
Kratochwill, T. R., & Levin, J. R. (1992). *Single-case research design and analysis: New direction for psychology and education*. Hillsdale, NJ: Lawrence Erlbaum.
Kratochwill, T. R., & Levin, J. R. (Eds.). (2014). *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association.
Lambert, J. M., Finley, C. I., & Caruthers, C. E. (2017). Trial-based functional analyses as basis for elopement intervention. *Behavior Analysis: Research and Practice*, 17(2), 166.
Lane, J. D., Ledford, J. R., & Gast, D. L. (2017). Current standards in single case design and applications in occupational therapy. *American Journal of Occupational Therapy*, 71, 1–9.
Ledford, J. R., Lambert, J. M., Pustejovsky, J. E., Zimmerman, K. N., Hollins, N., & Barton, E. E. (2023). Single-case-design research in special education: Next-generation guidelines and considerations. *Exceptional Children*, 89(4), 379–396.
Leko, M. M., Hitchcock, J. H., Love, H. R., Houchins, D. E., & Conroy, M. A. (2023). Quality indicators for mixed-methods research in special education. *Exceptional Children*, 89(4), 432–448.
Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, 43, 242–252.
Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Newbury Park, CA: Sage.
McReynolds, L. V., & Kearns, K. P. (1983). *Single-subject experimental designs in communicative disorders*. Baltimore, MD: University Park Press.
Neely, L., Graber, J., Kunnavatana, S., & Cantrell, K. (2020). Impact of language on behavior treatment outcomes. *Journal of Applied Behavior Analysis*, 53(2), 796–810.
Neuman, S. B., & McCormick, S. (Eds.) (1995). *Single subject experimental research: Applications for literacy*. Newark, DE: International Reading Association.
Odom, S. L. (1988). Research in early childhood special education: Methodologies and paradigm. In S. L. Odom & M. B. Karnes (Eds.), *Early intervention for infants and children with handicaps* (pp. 1–22). Baltimore, MD: Paul H. Brookes.
Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716–1–aac4716–8.
Ottenbacher, K. (1984). Nomothetic and idiographic strategies for clinical research: In apposition or opposition? *The Occupational Therapy Journal of Research*, 4, 198–212.
Patton, M. Q. (2014). *Qualitative research & evaluation methods: Integrating theory and practice*. Los Angeles: Sage Publications.

Pritchett, M., Ala'i-Rosales, S., Cruz, A. R., & Cihon, T. M. (2022). 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*, 15(4), 1074–1092.

Pyrczak, F. (2016). *Making sense of statistics: A conceptual overview*. London: Routledge.

QR Collective. (2023). Reflexive quality criteria: Questions and indicators for purpose-driven special education qualitative research. *Exceptional Children*, 89(4), 449–466.

Quinn, E. D., Kaiser, A. P., & Ledford, J. (2021). Hybrid telepractice delivery of enhanced milieu teaching: Effects on caregiver implementation and child communication. *Journal of Speech, Language, and Hearing Research*, 64(8), 3074–3099.

Radley, K. C., Dart, E. H., Fischer, A. J., & Collins, T. A. (2020). Publication trends for single-case methodology in school psychology: A systematic review. *Psychology in the Schools*, 57(5), 683–698.

Richards, S. B., Taylor, R. L., Ramasamy, R., & Richards, R. (1999). *Single subject research: Applications in educational and clinical settings*. San Diego, CA: Singular Publishing Group.

Schlosser, R. W., Belfiore, P. J., Sigafoos, J., Briesch, A. M., & Wendt, O. (2018). Appraisal of comparative single-case experimental designs for instructional interventions with non-reversible target behaviors: Introducing the CSCEDARS ("Cedars"). *Research in Developmental Disabilities*, 79, 33–52.

Shepley, C., Shepley, S. B., & Spriggs, A. D. (2023). On the history of single-case methodology: A data-based analysis. *Journal of Behavioral Education*. Advanced online publication.

Sidman, M. (1960). *Tactics of scientific research—evaluating experimental data in psychology*. New York: Basic Books.

Skinner, C. H. (2004). Single-subject designs for school psychologists. *Journal of Applied School Psychology*, 20, 2.

Snodgrass, M. R., Chung, M. Y., Kretzer, J. M., & Biggs, E. E. (2022). Rigorous assessment of social validity: A scoping review of a 40-year conversation. *Remedial and Special Education*, 43(2), 114–130.

Snow, C. E. (2014). Rigor and realism: Doing educational science in the real world. *Educational Researcher*, 44, 460–466.

Spencer, T. D., Detrich, R., & Slocum, T. A. (2012). Evidence-based practice: A framework for making effective decisions. *Education and Treatment of Children*, 35(2), 127–151.

Spriggs, A. D., Gast, D. L., & Knight, V. F. (2016). Video modeling and observational learning to teach gaming access to students with ASD. *Journal of Autism and Developmental Disorders*, 46, 2845–2858.

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*.

Tawney, J. W., & Gast, D. L. (1984). *Single subject research in special education*. Columbus, OH: Charles E. Merrill.

Toste, J. R., Logan, J. A., Shogren, K. A., & Boyd, B. A. (2023). The next generation of quality indicators for group design research in special education. *Exceptional Children*, 89(4), 359–378.

Volkert, V. M., Lerman, D. C., Call, N. A., & Trosclair-Lasserre, N. (2009). An evaluation of resurgence during treatment with functional communication training. *Journal of Applied Behavior Analysis*, 42(1), 145–160.

2

External Validity and Generalizable Knowledge

Joseph M. Lambert, Jennifer R. Ledford, Tara Fahmie, and David L. Gast

Important Terms

external validity, within-study replication, across-study replication, status variables, functional characteristics, endogenous implementers, ecological validity, construct validity

Table of Contents

In the previous chapter, we discussed purposes of research and the various paradigms and assumptions associated with different types of research. In this chapter, we specifically focus on single case research design and the importance of replication *across* studies. In the next chapter, we will focus again on replication, but with a focus on replication *within* studies.

DOI: 10.4324/9781003294726-3

External Validity

External validity refers to the extent to which a relation demonstrated in a specific context holds in other contexts (Shadish et al., 2001). For example, can the relation identified in one or more studies be expected to hold in other settings (e.g., classrooms versus clinics), for other types of participants (e.g., with different characteristics or skills, of different ages, who have different histories of intervention), or with different interventionists (e.g., with researchers who did not develop the intervention, with researchers from a different institution who were trained by different people, with non-researchers such as teachers or other service providers). It can also refer to applicability to other measured behaviors and with different treatment features (e.g., dosage, frequency, components). It is worth noting explicitly that external validity is applied to *relations*. That is, a given intervention or study cannot be deemed externally valid.

A common criticism directed at single case research methodology is that findings cannot generalize beyond the individual—there simply are too few participants in studies that employ single case designs. By contrast, group research methodology, in which a researcher randomly assigns many participants to two or more groups, is thought to more directly establish external validity. Few would argue that findings generated by large group research—compared to those generated by single case research—generalize better to other large unstudied groups. Of course, this is only true if individuals in the unstudied group are "similar" to participants in the studied group. Wolery and Ezell (1993) point out, "The more similar the two populations, the greater the likelihood of accurate generalizations, and thus the greater the likelihood that findings will be replicated" (p. 644).

These positions regarding research methodology are reasonable. However—what if your interest is in generalizing findings to a specific individual, rather than a group of individuals? Remember, in large group research the data reported are measures of central tendency; thus, there are always individuals within the group who perform better and worse than the average participant. Seldom do these studies provide detailed descriptions of individual participants nor do they often report how individual participants responded to the independent variable. Their focus is on the group, not the individual. It might be reasonable then, to expect that findings of group design research generalize more readily to groups of people, while findings of single case design research generalize more readily to individuals.

Replication

Before we discuss strategies for establishing external validity through single case research, it is likely worth exploring the concept of replication. Johnston and Pennypacker (2009) defined a *replication* as a repetition of any parts of an experiment and a *reproduction* as a repetition of results, usually as an outcome of repetition of procedures (p. 241). When replication yields reproduction, it is possible to accumulate at least one of two types of evidence. First, replication that yields reproduction allows scientists to establish the reliability, or reproducibility, of a finding. It also allows them to rule out incidental contact with extraneous confounds and/or experimental artifacts as alternative explanations for observed effects. That is, replications that yield reproduction contribute to a study's believability because it offers confidence that changes in an independent variable are responsible for changes in a dependent variable (see Chapter 3). The strategies for producing this type of evidence include within-session, within-phase, and within-experiment replication (see Table 2.1) and entails repeating a procedure exactly as it was previously implemented to answer the question: "When I repeat this procedure, will I get the same outcome?" (Johnston & Pennypacker, 2009). This type of replication has been referred to as a *direct replication* by both Sidman (1960) and Kazdin (2010) and is the focus of Chapter 3.

Table 2.1 Contributions of Various Types of Replications

Contribution			**Replication Types**		**Classification**		
			Johnston & Pennypacker (2009)		**Sidman (1960)**	**Kazdin (2009)**	**This text**
	Reliability of Effect	Internal Validity	Within-session replication	Repetition of a basic element of a procedure throughout each session (e.g., multiple opportunities to respond presented in a discrete trial format).	Direct Replication	Direct Replication	Within-Study Replication
			Within-phase replication	Repetition of the same condition many times in succession throughout a phase.			
			Within-experiment replication	Repetition of an entire phase over the course of an experiment.			
Generality of Effect		External Validity	Within-literature replication	Repetition of an earlier experiment, usually by other researchers, in which all known critical features are held constant.	Systematic Replication		Across-Study Replication
				Repetition of an earlier experiment, usually by other researchers, in which some elements of the previous experiment are held constant and others are intentionally altered.		Systematic Replication	
			Across-literature replication	Repetition of phenomena under different conditions across different fields of science.			

Note: The gray-filled cell represents ambiguity, given that replication across multiple participants in a single study could contribute to generality, but that within-study replications do not necessarily include multiple participants. Black-filled cells represent no contribution to generality of effect.

A second function of replication allows scientists to establish the generality of a given effect. That is, replication can help scientists determine whether a specified relation is constrained to a single individual or circumstance, or if it represents a generalizable principle which might contribute to scientific knowledge. The strategies for producing this type of evidence include *within-literature* and *across-literature replications* (see Table 2.1) and entail repeating some elements of a previously implemented procedure while intentionally modifying others to answer the question: "If I change the procedure, will I get the same outcome?" (Johnston & Pennypacker, 2009). This second function has been referred to as a *systematic replication* by both Sidman (1960) and Kazdin (2010).

However, Kazdin (2010) distinguished two types of *within-literature* replication and classified each differently. First, when researchers are only interested in reproducing a previously established outcome and do not seek to explore which elements of the original procedure were critical to producing the original outcome, or how a parametric manipulation of the independent variable might alter the outcome, then they replicate as many elements of the original procedure as possible. By contrast, when researchers are uncertain about the status (i.e., critical vs. non-critical) of procedural details present in a previous experiment, or when they seek to expand understanding of *how* an independent variable controls a dependent variable through parametric manipulation, they will replicate some elements of the original procedure while intentionally changing other elements. Kazdin referred to the former as a direct replication and the latter as an indirect replication. Although somewhat inconsistent with earlier conceptualizations of a direct replication (e.g., Sidman, 1960), this newer framing can be useful for communicating a study's intended contribution and is often adopted by scientists who employ, and journals that publish, single case design. Despite subtle differences in terminological nuance, all *within-literature* and *across-literature* replications serve the dual function of establishing the reliability of a phenomenon while simultaneously probing its generality (Sidman, 1960).

Given disagreement on exact terminology that has persisted across years, we elect to use the terms **within-study replication** to refer to replication that occurs in the context of a single case design study and primarily impacts internal validity and **across-study replication** to refer to replication across more than one study that impacts external validity.

Parsing Critical from Non-Critical Features via Across-Study Replication

Importantly, external validity can be evaluated for a group of studies, but not for an individual single case design study (Birnbrauer, 1981). That is, variability across attempts (e.g., number and types of differences between studies) determines the range of generality established through single case design. Sidman (1960, p. 111) noted that "replication demonstrates that the finding … can be observed under conditions different from those prevailing in the original experiment." As mentioned above, across-study replication can: (1) demonstrate the reliability of an effect, and (2) extend the generality of a finding. They are also useful for identifying exceptions to such generality. Perhaps one of the best examples of systematic across-study replication via single case design is a long series of studies conducted by David Gast and Mark Wolery (and colleagues) related to the effectiveness and efficiency of time delay prompting procedures. Over years, they conducted many studies, systematically expanding evidence of generality, including variation in participants, instructional arrangements (i.e., individual, group), and target behaviors (e.g., discrete academic behaviors, chained self-help skills), as well as less commonly studied parameters like fidelity errors. They published many studies—here we cite a representative but small subset that exemplifies the overall findings that this procedure was effective for teaching a variety of behaviors to individuals with disabilities, in a

variety of contexts (Doyle et al., 1990; Gast et al., 1990, 1991; Holcombe et al., 1994; Schuster et al., 1988; Wolery et al., 1997, 2002). Another smaller-scale example is a series of studies conducted by Vanderbilt students under the direction of M. L. Hemmeter and Jennifer Ledford, providing more limited evidence of the generality of peer-mediated intervention *Stay-Play-Talk* for establishing proximal play between children in early childhood classrooms (Milam et al., 2021; Osborne et al., 2019; Severini et al., 2019; Soemarjono, 2022; Tang, 2023; Taylor, 2023). This group of studies, along with studies done by unaffiliated researchers, suggests that *Stay-Play-Talk* will result in changes in proximal play behaviors (for a review, see Ledford & Pustejovsky, 2023) but the across-study replications were less comprehensive than the previous series by Gast and Wolery because most implementers were researchers, rather than a mix of researchers and various endogenous implementers.

When conducting across-study replications, it is important to identify potential critical and non-critical features of the intervention, implementers, participants, and settings. These can be established using your theory of change (see Chapter 6) and will guide the direction of your replication attempt. For example, if you identify a specific intervention component as a critical feature, you would include that feature in all replication attempts. If it is potentially non-critical, it may be worthwhile to vary the feature to answer the question about criticalness.

We agree with previous researchers (Cronbach et al., 1980; Cronbach & Shapiro, 1982) who have argued that scientists should not be responsible for answering questions that other researchers or practitioners have about a topic and that no one context is more important than another. While applicability across multiple contexts is useful for understanding the boundaries of generality, there is no "appropriate" amount of external validity that can be established for a given relation. The external validity of relations is viewed along a continuum in which the number of variables that change between studies will determine the extent of generality established. A relation that only holds in limited contexts for a specific *type* of participant is quite important to a practitioner when that context and participant type matches their context and client.

Tactics for Maximizing the Impact of an Across-Study Replication

There are no universally accepted guidelines indicating how many variables a researcher should modify during an across-study replication, nor how such experiments should be arranged. Sidman (1960) called systematic replication a gamble, one that if successful, would "buy reliability, generality, and additional information" (p. 112). On the one hand, large changes that yield a reproduction of previous outcomes can speak volumes about the generality of a previously established effect. On the other, when there are a substantial number of changes to a protocol and the new procedure fails to reproduce the original effect, it can be challenging to determine which of the modified or omitted variables was critical to a procedure's original success(es).

If designed properly, whatever the outcome of an across-study replication attempt, our understanding of the phenomenon being studied can be enhanced. The key to success may be to remain conservative in scope and to modify only a few variables at a time (Johnston & Pennypacker, 2009; Kazdin, 2010). In so doing, even failures to replicate can be instructional because it is easier to determine how the omission of a previously unknown critical variable impacts the overall outcome of an intervention; thus, leading to the discovery of limitations of current interventions and the discovery of new interventions.

Relatedly, Sidman (1960) recommended that researchers initiate across-study replications including extensions by first replicating the original experiment in question, with the intent of reproducing the original outcome. When a direct replication fails to reproduce the original

effect, information about the generality of said effect is offered. By contrast, when a direct replication reproduces the original effect, this outcome can serve as a baseline condition which can then be used to extend our understanding of *how* controlling variables alter participant performance.

For example, consider a researcher who wants to explore the durability of treatment effects following functional communication training (FCT; Carr & Durand, 1985)—which most often includes reinforcer available for 100% of opportunities—to situations in which reinforcement for a communicative response is reduced to 10% of opportunities. Following Sidman's recommendation, they would first directly replicate the original FCT procedure (i.e., reinforce communicative responses across 100% of opportunities) and reproduce the original outcome (i.e., a precipitous decrease in challenging behavior and an increase in independently emitted communicative responses). Then, they would explore how (if at all) a 90% decrease in reinforcer availability might impact baseline performance. In so doing, the baseline demonstration (1) serves as a direct replication and offers additional evidence of the reliability of the original effect, and (2) offers the research team a degree of credibility which allows them to more convincingly draw conclusions about the interpretability of both positive and/or null effects resultant from their subsequent manipulation. For example, if the above-mentioned researcher found that FCT was ineffective when reinforcement was only available 10% of the time, the baseline demonstration would allow them to assert that the degradation in treatment outcome was due to the corresponding decrease in reinforcer availability, rather than a lack of talent. That is, because the researcher first demonstrated that they could produce the original effect, critics could not claim that the lack of effect was attributable to the researcher's incompetence. Of course, this strategy does not always align with the intended research questions—for example, if a researcher was interested in whether FCT was effective when reinforcers were provided for 80% of opportunities *from the beginning of intervention*, the authors could not first replicate traditional FCT. Nonetheless, when procedures consistently produce stable outcomes, systematic lines of inquiry that employ this "baseline" strategy can amass a considerable amount of evidence supporting the reliability of the original relation while simultaneously exploring the parameters and boundary conditions that establish such relations.

Failures to replicate should "spur further research rather than lead to a single rejection of the original data" (Sidman, 1960, p. 74). "Science progresses by integrating, and not by throwing out, seemingly discrepant data" (Sidman, 1960, p. 83). In this regard, as an applied researcher, your responsibility is to identify modifications to the original intervention, or identify an alternative intervention, that will be beneficial to the participant. It is not generally acceptable to simply note that there was a failure to replicate and move on; this is partially because identifying and evaluating an alternative *effective* behavior change procedure will provide information about *why* the original procedure was ineffective and provides evidence that the behavior was amenable to change and that data collection procedures were sufficiently sensitive and valid.

Considerations Relevant to External Validity

Below we discuss the four features that are commonly discussed in relation to the external and ecological validity of relations (Shadish, Cook, & Campbell, 2001)—participants, contexts (settings), dependent variables, and intervention features.

Participant Characteristics

When an educator or practitioner considers using an intervention with a student or client based on one or more single case studies, they may ask: "What individual characteristics or

variables should I consider in determining the likelihood that the intervention under consideration will be successful with my students or clients?" There are several variables that they might consider, including status variables and functional characteristics. **Status variables** are participant descriptors including gender, age, race, ethnicity, disability, academic achievement, grade level, educational placement, and geographic location (Research Committee of the Council for Learning Disabilities; Rosenberg et al., 1992). This type of descriptive information is common and expected in research reports, but is it sufficient for determining whether an intervention will generalize to an individual with similar status variable descriptors? Wolery and Ezell (1993) assert that status variables are only "part of the picture" for determining external validity, and "that failure to replicate in subsequent research or in clinical and educational settings is undoubtedly related to many other variables than the precise description of subject characteristics" (p. 643). In a brief review of constant time delay (CTD) research they found that despite consistent findings across several studies, procedural modifications were necessary even though participants "were nearly identical on status variables." They concluded that status variables were, if not unimportant, at least not *the most important* considerations for intervention success.

But if status variables are not the best predictors of generalization, what variables are? We suggest that functional characteristics are likely better predictors. **Functional characteristics** are features that are particularly relevant to the relation being studied; that is, they are functionally related to the independent and dependent variables. For example, when studying naturalistic developmental behavioral interventions (NDBIs) for young children, the extent to which children initiate communication, engage in conventional play behaviors, and actively avoid proximity with others may be functional characteristics that impact intervention success more so than age, disability status, or race/ethnicity. In a study designed to increase conversation for young children with autism, Bateman et al. (2023) reported important demographic/status variables (e.g., age, diagnosis, home language spoken) but also reported each participant's typical expressive and receptive language levels ("vocally labeled a wide variety of two-dimensional and three-dimensional stimuli and receptively identified many common items … imitated adult actions, followed many 1-step directions, and was beginning to respond appropriately and correctly when asked personal information questions," p. 167). This information is likely critical for determining the characteristics of autistic children likely to benefit from the intervention. (We note here that we use both person-first and identify-first language in this text, given the variability in preferences for the subset of the population who have had the opportunity to share their opinions (e.g., Bury et al., 2023; Kenny et al., 2016).)

One way to assess functional characteristics is to describe the characteristics of participants' pre-intervention environments (e.g., response contingencies, number of opportunities to respond) and their behavior patterns. For example, during a large group activity in a classroom, with multiple opportunities for choral responding and social praise for correct answers, assume two young children (JD and Kenton) respond often and correctly and two young children (Kyson and Myles) respond rarely. In this case, all are 4-year-old males but baseline responding is consistently different for Kyson and Myles, perhaps indicating that intervention is required. This, however, is not enough to confirm that the *same* intervention is likely to result in behavior change. For example, during teacher interviews you might learn that Kyson's academic skills are advanced, but his motivation is low (indicating potential need for a reinforcement-based intervention), while Myles has more difficulty with acquiring the academic skills targeted during the large group activity (indicating a potential need for a focused academic intervention). Thus, information about the baseline performance of participants can, and should, be gleaned from multiple sources and used to determine the extent to which participants are similar on critical variables potentially impacting intervention success. Another

example of this critical concept is that high rates of challenging behavior in a baseline condition for two individuals do not necessarily implicate that the same intervention would be successful for reducing those behaviors—instead the function of the behavior (along with other contextual information) is more likely to lead to appropriate intervention selection. To determine what variables are critical, you must gain expertise in the relation being studied, including specifying a theory of change for your independent variable (see Chapter 6).

Context Characteristics

Experiments, including single case experiments, are often performed in atypical settings or activities. For example, even work described as *naturalistic* often takes place in research clinics (e.g., NDBIs, Windsor & Ledford, 2023) and work conducted in inclusive settings often takes place during activities carefully designed and controlled by researchers (e.g., pull-out instructional sessions with a researcher; Eyler & Ledford, 2023; Chazin & Ledford, 2021). We will refer to settings, activities, materials, and social partners as "context characteristics."

Not all experiments include questions about applicability in typical settings, and some researchers have pointed out that findings in a "typical" setting are no more generalizable (i.e., to other, different settings) than those in laboratory contexts (Birnbrauer, 1981). However, variability in context characteristics across multiple studies does provide evidence that the relations identified in one or more studies are robust to context differences. Thus, it is critical to describe the important, functional characteristics of contexts and activities so that others understand the evidence of generality that is present in a given body of work. For example, in the study by Bateman et al. (2023) discussed above, researchers describe the type of classroom, number of children present and their disability status, the activity type and seating arrangement, and procedures for redirecting non-participants—all of this information is important for replicability and understanding the context in which the relation can be expected.

Dependent Variables

Relations between dependent and independent variables can also vary by measurement and features of the dependent variable. For example, the relation between Intervention A and the rate of challenging behavior may vary based on the sensitivity of the measurement system and the topographical characteristics of the measured behavior (e.g., whether the behaviors are defined to include aggression, disruption, self-injury, etc.). Similarly, the relation between Intervention B and academic performance may be different based on the similarity of teaching targets to measurement, the breadth and number of targets, and the domain of behaviors (e.g., math, reading). The more variable the dependent variable features are across studies, the greater the evidence for generality of the relation with the larger construct of interest.

Intervention Features

The relation between an independent variable and dependent variable can also vary based on differences in implementation of the independent variable present across studies. Common ways interventions may vary include:

- **Frequency and dosage of intervention implementation**. The frequency and dosage of intervention implementation may impact relations between the intervention and dependent variables. The relation between frequency/dosage and outcomes is rarely experimentally established, but a body of work that shows relations are consistent even when variability in dosage exists provides evidence of external validity in this domain.
- **Variations in procedures**. Generally, interventions conducted across studies are not implemented in identical ways. For example, one researcher may use poker chips as tokens for correct responding during an instructional session, while another may provide brief

access to a preferred toy. Generality across differences in implementation provide evidence that the relation is robust to variations in procedures. Sometimes, a general framework is used to establish a relation between a given dependent variable (e.g., rate of challenging behavior) and a *process of intervention selection and implementation* rather than a specific procedure. This type of research provides evidence that *decisions* made using a framework, which vary by participant, can lead to similar outcomes. These studies may provide *within-study* evidence of external validity in relation to procedural variations.

- **Implementer role, training, and qualifications**. Evidence that a given relation holds with a variety of implementers may be valuable in predicting its utility outside the research context. For example, enhanced milieu teaching, an NDBI, has been implemented by teachers, parents, and graduate students (increasing evidence of external validity in relation to implementer status), but only after considerable training and coaching from skilled researchers (providing limited evidence of external validity in relation to variations in intensity of implementer training; Quinn, Kaiser & Ledford, 2021; Roberts et al., 2014; Wright & Kaiser, 2017).
- **Fidelity of intervention implementation**. Relations that hold even when the intervention is not implemented as planned are highly valuable, since it may be difficult for **endogenous implementers** (i.e., individuals normally present in an individual's typical environment) to consistently implement some interventions with high fidelity. Some relations may be durable in the presence of fidelity failures while others could require high-fidelity implementation. Data regarding the extent to which intervention implementation occurred as planned is critical for identifying relations between fidelity and outcomes (see Chapter 6).

Related Constructs

Ecological Validity

Ecological validity refers to the extent to which study features relate to *real-world* contexts, and as such, is considered an important aspect of external validity. Studies with a high degree of ecological validity (e.g., studies conducted in a typical setting with endogenous implementers) provide greater confidence that discovered relations will hold in those relevant contexts; by contrast, studies with a low degree of ecological validity (e.g., studies conducted in an austere research space with unfamiliar researchers as implementers) do not establish the generality of relations to other non-research contexts. When single case research is conducted to improve the everyday life of human participants, it is paramount to establish that demonstrated relations hold under typical and relevant conditions. As such, ecologically valid research is commonly desired among behavioral researchers. However, it is not accurate to assume that relations demonstrated under highly contrived conditions cannot contribute to robust and generalized phenomena; rather, across-study replication along the continuum of ecological validity (from basic, to translational, to applied research) has proven a successful strategy for the discovery of phenomena with high degrees of generality and has formed the foundation of the science of human behavior (Fahmie et al., 2023).

Construct Validity

Another concept somewhat related to generality and external validity is that of **construct validity**, which refers to the extent that features of a study are representative of the actual concepts of interest. Generally, construct validity has been conceptualized as pertaining primarily

Table 2.2 Examples of Threats to Construct Validity

Domain	Threats to Construct Validity		
	Inadequate Explication of Constructs	*Inaccurate or Non-Conventional Explication of Constructs*	*Construct Confounding*
Participants	Authors describe participants as being "at risk" but do not describe characteristics leading to this designation.	Authors describe participants with autism but include infant siblings without official autism diagnoses.	Authors describe children as having challenging behavior, but they are identified by teacher report, which is confounded with race and culture.
Settings	Authors report their study is conducted in a "school" but fail to describe that it is a private program with 1:1 staffing and highly-trained, research-involved implementers.	Authors describe the setting as "free play in the classroom" but provide only one toy and no peers are present.	Authors describe two settings for measurement as *segregated* and *inclusive* but do not specify supports and structure in each setting likely to be related to intervention success variation (e.g., one-to-one support in one setting but not the other).
Treatment Variables	Authors report the use of *visual supports* but provide no information about topographical or functional characteristics of these visuals.	Authors describe the use of response interruption and redirection (RIRD) but only use interruption rather than providing redirection to a meaningful activity.	Authors describe the use of a reinforcement-based intervention but fail to describe that one active ingredient of the intervention was more consistent use of response cost (a punishment-based procedure).
Measurement Variables	Authors measure *challenging behavior* but do not provide operational definitions of the behaviors included.	Authors use *challenging behavior* to refer to a variety of behaviors, including "precursor behaviors" which would not conventionally be considered to be problematic.	Authors measure stereotypy across conditions, but in one condition, they do not include stereotypy occurring during active intervention segments.

to dependent variables. However, as with external validity, Shadish et al. (2001) describe four areas in which construct validity issues can occur: Participants, settings, dependent variables, and intervention features. One way to think about construct validity is to ask: Are the labels used in my study representative of the ideas I'm attempting to convey, or are they misaligned, incomplete, or not sufficiently specific? Construct validity is not necessarily tied to the relation identified, although difficulties with construct validity can impact conclusions drawn about the relation. Here, we identify two threats to construct validity most applicable to single case design, using terms suggested by Shadish et al. (2001); and describe one additional problem not identified by Shadish (inaccurate or non-conventional explication of constructs). In Table 2.2, we describe examples of how these problems can occur across the four types of study features.

- **Inadequate explication of constructs**: This problem refers to the insufficient description of a study feature. Insufficient descriptions of participants, settings, dependent variables, and intervention characteristics limits correct interpretation of the concepts that authors intended to convey in their studies. For example, an author who describes participants simply as *children with autism* prevents readers from understanding functional characteristics of participants likely to benefit from the intervention given the heterogeneity associated with autistic individuals.
- **Inaccurate or non-conventional explication of constructs**: This problem, not described by Shadish et al. (2001), refers to the use of a term that is applied non-conventionally

rather than non-adequately. For example, authors might use different terms for the same intervention, or the same term for interventions that are quite different. This problem has been identified as a serious impediment to synthesizing relations across studies (cf. Ledford et al., 2021). Relatedly, authors can use terms such as *challenging behavior* to refer to a variety of behaviors, including "precursor behaviors" which would not conventionally be considered to be problematic.

- **Construct confounding**: This problem occurs when a described feature of a study is confounded with an unconsidered feature. For example, authors of a study may suggest that Intervention A is effective for changing the behavior of children with challenging behaviors. However, the components of the described intervention (one construct) co-occurred with an uncontrolled additional feature—positive and responsive adult interactions (a separate construct). Thus, it is possible that the second construct, rather than the intended one, could be partly or wholly responsible for the intervention effect.

We note that some items described by Shadish as related to *construct validity* are conceptualized as being threats to *internal validity* and described in Chapter 3 (experimenter expectancies and treatment diffusion, as related to procedural infidelity; novelty and disruption effects as related to adaptation; reactivity to experimental situation as Hawthorne effects). The differences are subtle—there is a construct validity issue when there is a concern about the extent to which study features match constructs of interest, but an internal validity issue when something about the study features impede our ability to draw causal conclusions. Construct validity has not been explicitly discussed in previous versions of this textbook (Ledford & Gast, 2018) or other well-regarded single case texts. Thus, we provide considerations for construct validity as preliminary guidance for the field. We acknowledge that additional work in this area, including specific guidance for avoiding these threats, would be beneficial for the field.

Recommendations for Across-Study Replications

If you wish to initiate an across-study replication attempt, we suggest you proceed using the following general steps. More information is provided about reviewing and summarizing literature in Chapters 18 and 19.

1. Identify studies that relate to your research interest(s) or question(s) via electronic search, author search, or ancestral search (see Chapter 19). It may be helpful to review recently published literature reviews and meta-analyses on your topic for a comprehensive reference list of empirical investigations that addressed the same or similar research question(s).
2. Organize information about the studies, including ways they are similar and different across domains (e.g., participants, settings, dependent variables, intervention features).
3. Read and list researchers' suggestions for future research on the topic. These are commonly found in the discussion section of research reports.
4. Write your research question(s), if you haven't already, considering previous research, your own clinical or educational expertise, and practical resource constraints (e.g., access to participants, daily schedule, availability of materials, control of contingencies).
5. Identify potential critical and non-critical features of the intervention, critical functional characteristics of participants, setting features that might impact replication, and measurement issues that have not been addressed in previous work.
6. Determine which features will remain the same as in previous work and which feature(s) will vary, using your theory of change (Chapter 6).

7. Write and revise a research proposal, explicitly stating that the study is a replication attempt, and report the specific differences between your proposed study and those that have preceded it.
8. Conduct the study according to your written protocols and note whether relations identified in previous studies hold given your specific variations. In cases of "failure to replicate," your ability to implement a successful variation of, or alternative to the original intervention, will advance understanding of the reliability, generality, and limitations of the relation of interest.

Conclusion

Careful consideration of the constructs of interest and replication of previously identified relations are essential for evaluating generality. Across-study replication is ongoing, never over, as a failure to replicate may be just around the corner. When outcomes are not reproduced as expected, a limitation to the external validity of the relation is revealed. Applied behavioral researchers approach such failures as a challenge and attempt to identify their cause, as well as to identify modifications to the original intervention that will bring about the desired behavior change. Through the replication process, the science of human behavior is advanced and our ability to design effective and efficient instructional and treatment programs enhanced.

References

Bateman, K. J., Wilson, S. E., Gauvreau, A., Matthews, K., Gucwa, M., Therrien, W., … & Mazurek, M. (2023). Visual supports to increase conversation engagement for preschoolers with autism spectrum disorder during mealtimes: An initial investigation. *Journal of Early Intervention*, 45(2), 163–184.

Birnbrauer, J. S. (1981). External validity and experimental investigation of individual behavior. *Analysis and Intervention in Developmental Disabilities*, 1, 117–132.

Bury, S. M., Jellett, R., Spoor, J. R., & Hedley, D. (2023). "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*, 53(2), 677–687.

Carr, E. G., & Durand, V. M. (1985). Reducing behavior problems through functional communication training. *Journal of Applied Behavior Analysis*, 18(2), 111–126.

Chazin, K. T., & Ledford, J. R. (2021). Constant time delay and system of least prompts: Efficiency and child preference. *Journal of Behavioral Education*, 30(4), 684–707.

Cronbach, L. J., & Shapiro, K. (1982). *Designing evaluations of educational and social programs*. San Francisco: Jossey-Bass.

Cronbach, L. J., Ambron, S. R., Dornbusch, S. M., Hess, R. D., Hornik, R. C., Phillips, D. C., … & Weiner, S. S. (1980). *Toward reform of program evaluation* (p. 3). San Francisco: Jossey-Bass.

Doyle, P. M., Gast, D. L., Wolery, M., Ault, M. J., & Farmer, J. A. (1990). Use of constant time delay in small group instruction: A study of observational and incidental learning. *The Journal of Special Education*, 23(4), 369–385.

Eyler, P. B., & Ledford, J. R. (2023). Efficiency and child preference for specific prompting procedures. https://osf.io/dpq5w/

Fahmie, T. A., Rodriguez, N. M., Luczynski, K. C., Rahaman, J. A., Charles, B. M., & Zangrillo, A. N. (2023). Toward an explicit technology of ecological validity. *Journal of Applied Behavior Analysis*, 56(2), 302–322.

Gast, D. L., Doyle, P. M., Wolery, M., Ault, M. J., & Baklarz, J. L. (1991). Acquisition of incidental information during small group instruction. *Education and Treatment of Children*, 14(1), 1–18.

Gast, D. L., Wolery, M., Morris, L. L., Doyle, P. M., & Meyer, S. (1990). Teaching sight word reading in a group instructional arrangement using constant time delay. *Exceptionality: A Special Education Journal*, 1(2), 81–96.

Holcombe, A., Wolery, M., & Snyder, E. (1994). Effects of two levels of procedural fidelity with constant time delay on children's learning. *Journal of Behavioral Education*, 4, 49–73.

Johnston, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). New York: Routledge.

Kazdin, A. (2010). *Single-case research designs: Methods for clinical and applied settings*. New York: 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

Ledford, J. R., Lambert, J. M., Barton, E. E., & Ayres, K. M. (2021). 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.

Ledford, J. R., & Gast, D. L. (2018). *Single case research methodology*, 3rd ed. New York: Routledge.

Ledford, J. R., & Pustejovsky, J. E. (2023). Systematic review and meta-analysis of stay-play-talk interventions for improving social behaviors of young children. *Journal of Positive Behavior Interventions*, 25(1), 65–77.

Milam, M. E., Hemmeter, M. L., & Barton, E. E. (2021). The effects of systematic instruction on preschoolers' use of Stay-Play-Talk with their peers with social delays. *Journal of Early Intervention*, 43(1), 80–96.

Osborne, K., Ledford, J. R., Martin, J., & Thorne, K. (2019). Component analysis of stay, play, talk interventions with and without self-monitored group contingencies and recorded reminders. *Topics in Early Childhood Special Education*, 39(1), 5–18.

Quinn, E. D., Kaiser, A. P., & Ledford, J. (2021). Hybrid telepractice delivery of enhanced milieu teaching: Effects on caregiver implementation and child communication. *Journal of Speech, Language, and Hearing Research*, 64(8), 3074–3099.

Roberts, M. Y., Kaiser, A. P., Wolfe, C. E., Bryant, J. D., & Spidalieri, A. M. (2014). Effects of the teach-model-coach-review instructional approach on caregiver use of language support strategies and children's expressive language skills. *Journal of Speech, Language, and Hearing Research*, 57(5), 1851–1869.

Rosenberg, M. S., Bott, D., Majsterek, D., Chiang, B., Bartland, D., Wesson, C., Graham, S., et al. (1992). Minimum standards for the description of participants in learning disabilities research. *Learning Disabilities Quarterly*, 15, 65–70.

Schuster, J. W., Gast, D. L., Wolery, M., & Guiltinan, S. (1988). The effectiveness of a constant time-delay procedure to teach chained responses to adolescents with mental retardation. *Journal of Applied Behavior Analysis*, 21(2), 169–178.

Severini, K. E., Ledford, J. R., Barton, E. E., & Osborne, K. C. (2019). Implementing stay-play-talk with children who use AAC. *Topics in Early Childhood Special Education*, 38(4), 220–233.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA: Cengage Learning.

Sidman, M. (1960). *Tactics of scientific research—Evaluating experimental data in psychology*. New York: Basic Books.

Soemarjono, F. (2022). Comparing stay play talk with or without reinforcement versus business-as-usual on children's duration of play and talk (Unpublished Master's thesis). Vanderbilt University.

Tang, L. (2023). *Using stay-play-talk to increase levels of initiations and responses for children with social delays* (Master's thesis). http://hdl.handle.net/1803/18085

Taylor, A. L. (2023). *An adaptation of stay-play-talk for young children with internalizing behaviors*. (Doctoral dissertation, Vanderbilt University).

Windsor, S. A., & Ledford, J. R. (2023). Naturalistic developmental behavioral interventions: A systematic review of procedures, participants, and outcomes. *Under review*.

Wolery, M., Anthony, L., Caldwell, N. K., Snyder, E. D., & Morgante, J. D. (2002). Embedding and distributing constant time delay in circle time and transitions. *Topics in Early Childhood Special Education*, 22(1), 14–25.

Wolery, M., Anthony, L., Snyder, E. D., Werts, M. G., & Katzenmeyer, J. (1997). Training elementary teachers to embed instruction during classroom activities. *Education and Treatment of Children*, 20(1), 40–58.

Wolery, M. & Ezell, H. (1993). Participant descriptions and single participant research. *Journal of Learning Disabilities*, 26, 642–647.

Wright, C. A., & Kaiser, A. P. (2017). Teaching parents enhanced milieu teaching with words and signs using the teach-model-coach-review model. *Topics in Early Childhood Special Education*, 36(4), 192–204.

3

Establishing Internal Validity via Within-Study Replication

Jennifer R. Ledford and Kevin M. Ayres

Important Terms

intra-participant replication, inter-participant replication, internal validity, history effects, history threats, maturation effects, maturation threats, facilitative testing effects, inhibitive testing effects, testing threats, instrumentation, procedural infidelity, attrition, attrition bias, sampling bias, data instability, cyclical variability, multi-treatment interference, regression to the mean, adaptation, Hawthorne Effect

Table of Contents

Callout

DOI: 10.4324/9781003294726-4

As described in Chapter 2, replication refers to repeating attempts to establish a functional relation between a specific independent variable on a dependent variable. Replication is important in all research paradigms, and, in fact, the failure to replicate has been referred to as a "crisis" in psychology, behavioral science, education, and related literatures (Coyne et al., 2016; Locey, 2020; Shrout & Rodgers, 2018). The replication rate in published single case design research is higher than that of between-groups research (Lemons et al., 2014), although failures to replicate are difficult to locate because of the "file drawer effect" which describes the reluctance or inability of researchers to publish findings that show that an intervention *does not work* under some conditions (Gage et al., 2017; Tincani & Travers, 2019). As described in Chapter 2, there is some inconsistency in terminology used across time within in the field. We will refer to replication that occurs in a single study as *within-study replication*, and that is the focus of this chapter.

Within-Study Replication

What we refer to as within-study replication was defined as *direct replication* by Sidman (1960) as "the repetition of a given experiment by the same experimenter … accomplished either by performing the experiment again with new subjects or by making repeated observations on the same subject under each of several conditions" (p. 73). He describes two relevant classes of these replications: Intra-participant direct replication and inter-participant direct replication (historically, "intra-subject" and "inter-subject"—see Chapters 1 and 16 for a discussion of the use of the term "participant" rather than "subject"). Contemporary guidelines (e.g., Ledford et al., 2023) call for every single case study to include multiple attempts at replication (see Chapters 9, 11, and 13 for more information about the various ways in which replications are built into common single case designs).

Both intra-participant and inter-participant replications refer to an investigator's attempts to repeat an experimental effect. Repeated attempts for the same participant are referred to as **intra-participant replication**. For example, researchers have evaluated whether parent training improved the extent to which parents of young children with disabilities used *matched turns* (first demonstration), *target talk* (replication), and *expansions* (another replication) with their children (Peredo, Zelaya, & Kaiser, 2018) when intervention was applied to each behavior sequentially. Thus, for each participant, there were three opportunities for demonstration of the relation between the intervention and targeted interaction skills. Repeated attempts for different participants in the same study are referred to as **inter-participant replication**. For example, researchers have evaluated whether children learn better with and prefer time delay prompting procedures or the system of least prompt strategy, and included replication of the comparison for 6–10 children in a single study (e.g., Eyler & Ledford, 2023; Chazin & Ledford, 2021). Single case design studies with more than one participant can include both intra- and inter-participant replication. See Figures 3.1 and 3.2 for examples of a replicated relation for a single participant (intra-participant replication) and multiple participants (inter-participant replication). In Figure 3.1, the comparison between baseline and intervention conditions is repeated over time with the same participant by alternating between phases (an A-B-A-B design, discussed in detail in Chapters 9 and 10). In Figure 3.2, the comparison between baseline and intervention conditions is repeated with three different participants (a multiple baseline across participants design, discussed in detail in Chapters 11 and 12).

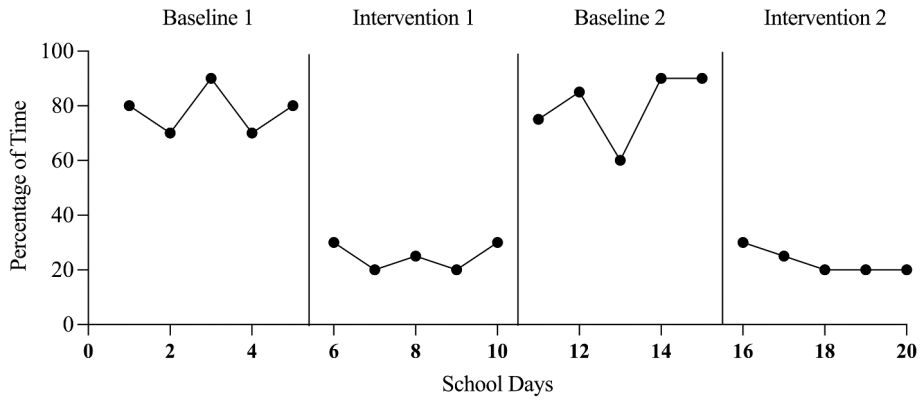

Figure 3.1 Intra-Participant Replication.

Note: This figure depicts within-participant replication of effects (i.e., intra-participant replication), with high levels of behavior in the first baseline condition (School Days 1–5), followed by a decrease in levels in the first intervention condition (School Days 6–10). This effect is then replicated with the same participant (School Days 11–20).

Callout 3.1 A Note on Terminology

Given we have already used the term *condition* multiples times, it seems important to define the term and to compare it with the similar term *phase*, since these terms are often used interchangeably and were not consistently used in the previous version of this text. We will use the term **condition** to refer to a collection of **measurement occasions** (e.g., each time point at which you measure behaviors, often divided into segments such as days, dates, or sessions) during which identical procedures are used. For example, a **baseline condition** refers to a collection of sessions during which the same non-intervention conditions are applied while an **intervention condition** refers to a collection of sessions during which the same treatment procedures are applied. These sessions need not be consecutive or adjacent; they are referred to as a single condition because of their procedural identicalness. To confuse matters, the plural of the term is sometimes used to refer to the procedures themselves (e.g., *baseline conditions* can be used interchangeably with *baseline procedures*). We will use the term **phase** to refer to a collection of sessions that occur during a given period of time. In some designs, conditions can be implemented in two different phases (e.g., baseline conditions can occur in two temporally separate phases, one before intervention implementation and one after intervention withdrawal), and in some phases, multiple conditions can be implemented (you can read more about this in Chapters 13 and 14 on rapid iterative alternation designs). In summary, a *condition* refers to a collection of sessions that are procedurally connected, while a *phase* refers to a collection of sessions that are temporally connected.

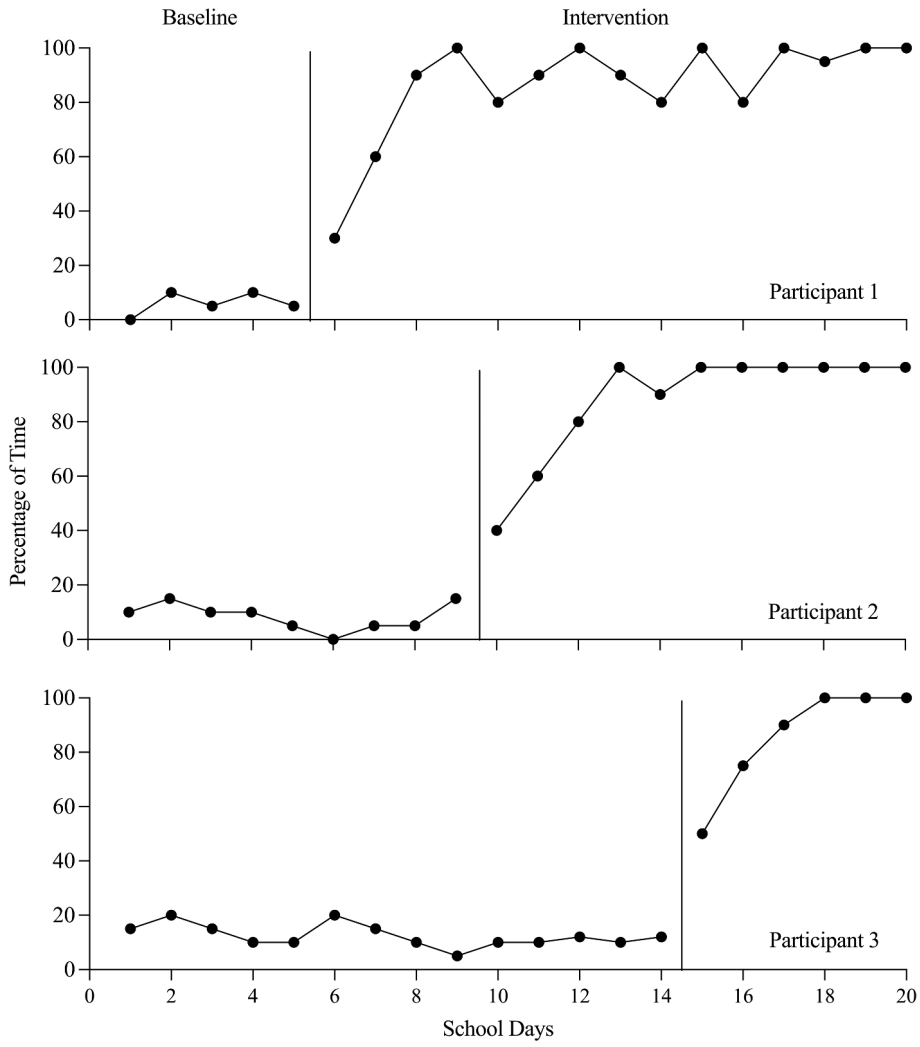

Figure 3.2 Intra-Participant Replication.

Note: This figure depicts across-participant replication of effects (i.e., inter-participant replication), with low levels of behavior in the baseline condition for Participant 1, followed by an increasing trend and high levels in the intervention condition (top panel). This effect is then replicated with two additional participants (Participant 2, middle panel; Participant 3, bottom panel).

Internal Validity

The purpose of both intra- and inter-participant replication is to provide confirmatory evidence that the relation that was initially observed is the result of a researcher's planned manipulations rather than any other cause that is external to the study. That is, within-study replication is necessary for establishing **internal validity**—the extent to which the relations observed in the study can be confidently attributed to the planned changes between conditions *and only to those changes*. For example, a researcher might collect data in a classroom across five days during a school week and observe high rates of off-task behavior for a particular child. That researcher might devise and implement an intervention the next school week and observe reduced rates of off-task behavior. However, many alternative explanations are possible for the change in behavior—perhaps the child was ill during the first week of observations and felt

well the second week. Or they may have been distracted by the researcher during the first week but acclimated to his presence the second week. Many possible and plausible explanations could exist for the difference. However, if the researcher attempted to replicate the relation by observing *without the intervention* during a third week, and again *with the intervention* during a fourth week, and the relations held (e.g., higher off-task behavior during no-intervention weeks and lower off-task behavior during intervention weeks), we would be more confident that differences in condition procedures were the cause of behavior change. Of course, even this is not foolproof—for example, the child may have been with a parent who enforced an early bedtime in weeks 2 and 4 but with a parent who had a later bedtime rule during weeks 1 and 3. This is why contemporary guidelines require specific ways of ordering the replications to ensure probable alternative explanations are ruled out (see Chapters 9, 11, and 13).

Threats to Internal Validity

Internal validity of a study depends on how well the researcher has controlled or mitigated other plausible explanations for behavior change beyond those planned experimental changes. Two concepts are important for understanding the pragmatics of internal validity. First, it is impossible to control for every possible alternative explanation. Second, a possible alternative explanation may not be an actual threat. Each possible alternative explanation should be considered in the design of your study and the analysis of other researchers' studies. The extent to which threats to validity are evaluated and controlled for, along with the presence of a sufficient number of within-study replications, will determine the level of confidence you have in the findings. You should not be disheartened to learn that just as there is no free lunch, there is no perfect experiment. Instead, there are carefully designed experiments, experiments that are executed as carefully as they were planned and that provide "adequate and proper data" (Campbell & Stanley, 1963, p. 2) for analysis. Your task is to describe what happened during the experiment and to be able to account for planned and unplanned outcomes. Below is a non-exhaustive list of threats to internal validity that may be likely in studies using single case design; many are also applicable for other experimental studies (e.g., group comparison studies). Table 3.1 lists these threats in relation to threats listed by Shadish et al. (2002, a book primarily concerned with between-groups research). We note that he identifies some threats as related to internal validity and some to statistical conclusion validity (the extent to which changes occur, without regard to causality) or construct validity (discussed in Chapter 2). We consider all to be relevant to internal validity, as they relate to single case design. Also in Table 3.1 is how the threats identified by Shadish and colleagues were categorized by Petursdottir and Carr (2018), in relation to single case design.

It is important to minimize the likelihood of threats to internal validity so that you can draw confident conclusions about the relation between the independent variable (e.g., intervention) and dependent variable in a single case study. A functional relation is established when consistent behavior change occurs, in the expected direction, when and only when condition changes occur, *and* when likely threats to internal validity have been mitigated. Specific requirements for functional relations and how to establish them via visual analysis will be discussed more in Chapters 10, 12, and 14.

History

History effects refer to events that occur during an experiment, but that are not related to planned procedural changes, that may influence the outcome. These effects, and their potential influence on conclusions about internal validity, are detected via visual inspection of graphs and (when applicable) careful session notes (e.g., describing the occurrence of events that could

Table 3.1 Threats to Internal Validity, as Categorized by Other Researchers

Threats	Terminology Used by Shadish et al. (2002)	Type (Shadish et al., 2002)	Relevance to Single Case (Petursdottir & Carr, 2018)
History	History	Internal	Relevant
Maturation	Maturation	Internal	Relevant
Testing	Testing	Internal	Relevant
Attrition	Attrition	Internal	Not Relevant
Selection	Selection	Internal	Not Relevant
Regression to the Mean	Regression Artifacts	Internal	Not Relevant
Instrumentation	Unreliability of Measures	SC	Relevant
	Instrumentation	Internal	Relevant
Procedural Infidelity	Unreliability of Implementation	SC	Relevant
	Experimenter Expectancies	Construct	Relevant
	Treatment Diffusion	Construct	Relevant
Adaptation	Novelty and Disruption Effects	Construct	Relevant
Hawthorne Effect	Reactivity to Experimental Situation	Construct	Relevant
Multi-treatment Interference	Not Addressed	–	Relevant
Data Instability	Extraneous Variance in Experimental Setting	Statistical Conclusion	Relevant
Design-related Confounds	Not Addressed	–	–

Note: Shadish defines some threats as related to construct or statistical conclusion rather than internal validity. Petursdottir & Carr (2018) identify some as relevant to single case design. We consider all potentially relevant for at least some studies. SC = Statistical Conclusion Validity.

be related to participant behavior). Generally speaking, the longer the study (both in terms of the number of experimental sessions and number of calendar days), the greater the threat due to history. Potential sources of history effects, when a study is conducted in community settings, are the actions of others (parents, siblings, peers, childcare providers) or by study participants themselves (independent online research, observational learning, serendipitous exposure via social media). For behaviors that demand immediate attention in the eyes of a significant other, there may be an attempt to intervene prior to the scheduled intervention time. For example, while a researcher is implementing a token economy to reduce problem behaviors, a parent might introduce a separate (and unplanned) punishment procedure while the study is ongoing. While the parent may intend for the additional procedures to enhance your planned intervention (and while they may do this), this unplanned "history" effect will render your results less interpretable. Also, participants may learn target content through television or learn target social behaviors through observing the consequences delivered to others; the change in behavior resulting from this learning is a history effect. Other individual-specific unplanned events (e.g., seizure the night before, fight on the school bus, medication change) or community-wide events (e.g., school-wide policy change, widespread social unrest) may temporarily alter the occurrence of the target behavior.

A history effect that has a minimal impact on confidence in outcomes is depicted in the top panel of Figure 3.3. In this hypothetical example, a participant engaged in much lower levels of challenging behavior on a school day when he was ill. Because levels of challenging behavior

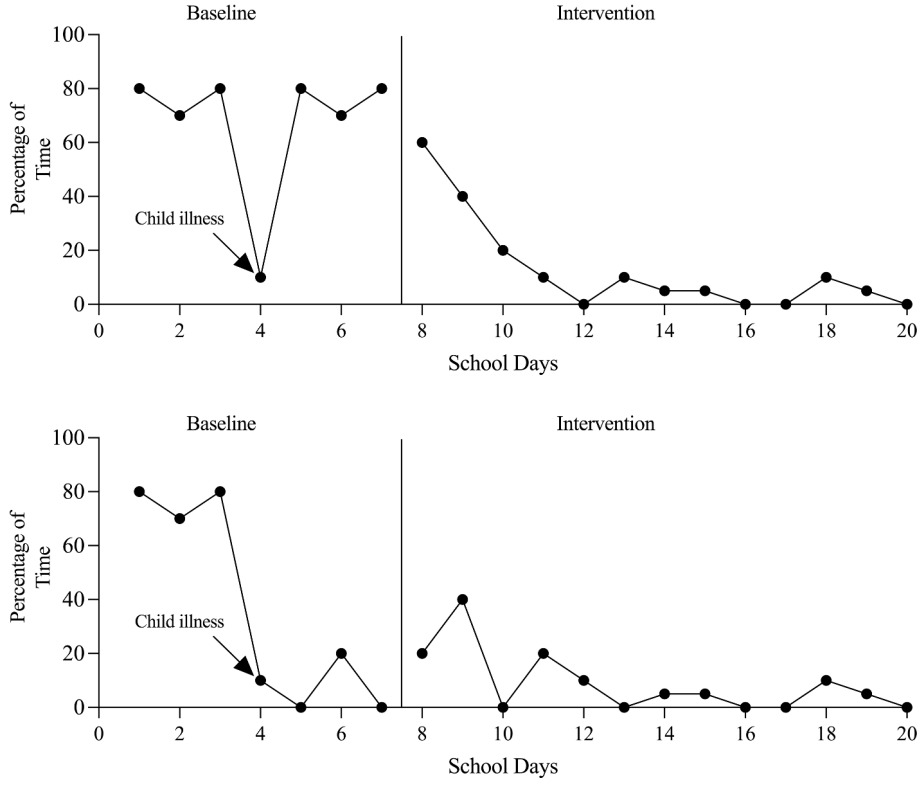

Figure 3.3 History Effects with Minimal (top) and Considerable (bottom) Impacts on Confidence in Relation between Independent and Dependent Variables.

Note: In the top panel, a potential history effect is depicted in the baseline condition. On School Day 4, a child was ill, which corresponded with unusually low levels of challenging behavior compared with the other six measurement occasions during baseline. Because levels were similar before and after the day that event occurred, our confidence in the relation between the independent and dependent levels is minimally impacted. In the bottom panel, the same decrease in level occurred when a child was ill, but levels remained low in the baseline condition following the illness. Thus, we cannot be reasonably certain that the intervention was responsible for any change in challenging behavior.

were consistently high on other days (before and after the illness), that one event has little impact on our interpretation of the effectiveness of the intervention. However, in the bottom panel of Figure 3.3, a **history threat**—an instance of a history effect that decreases internal validity and interpretability of results—is depicted. In this hypothetical example, the participant's challenging behavior does not increase in the school days following the illness. This prevents us from drawing conclusions about the effectiveness of the intervention. History threats generally cannot be prevented, although in some cases, it might be appropriate to ask participants to not engage in related outside activities that increase risk of history threats (e.g., beginning new related therapies, doing an Internet search about the intervention procedures prior to training); when these types of instructions are given to participants, they should be reported in written descriptions of the study.

It is important to note that history, as well as maturation and testing effects, may be unclear or indistinguishable from other threats via visual analysis. That is, we can hypothesize that a given event occurred (history threat) but we often cannot be certain that the hypothesized event is responsible for behavior change. Similarly, it may not be clear from graphs whether a

history event outside of the study resulted in behavior change, or whether behavior change is due to testing procedures during baseline (described below), or some other factor.

Maturation

Maturation effects refer to changes in behavior due to the passage of time and are also detected via visual inspection of graphs. In a "short" duration study maturation is not likely to influence the analysis of the effectiveness of a powerful independent variable that focuses on improving language or motor skills of a child who has a history of slow development. If the study is carried out over several months or longer with the same young child, especially if an intervention is used during which slow and gradual improvements in skills are expected, there is a greater likelihood that maturation effects result in decreased interpretability. The top panel of Figure 3.4 shows a data pattern in baseline that might indicate a *maturation effect* (i.e., a child's behavior is changing due to the passage of time), but one that does not significantly impact conclusions drawn about change that occurs when the intervention is implemented. The bottom panel shows an instance of a **maturation threat**—an instance of a maturation effect that threatens the internal validity of the study, making interpretation of intervention effects difficult. Maturation threats are primarily prevented by avoiding long-duration studies (e.g.,

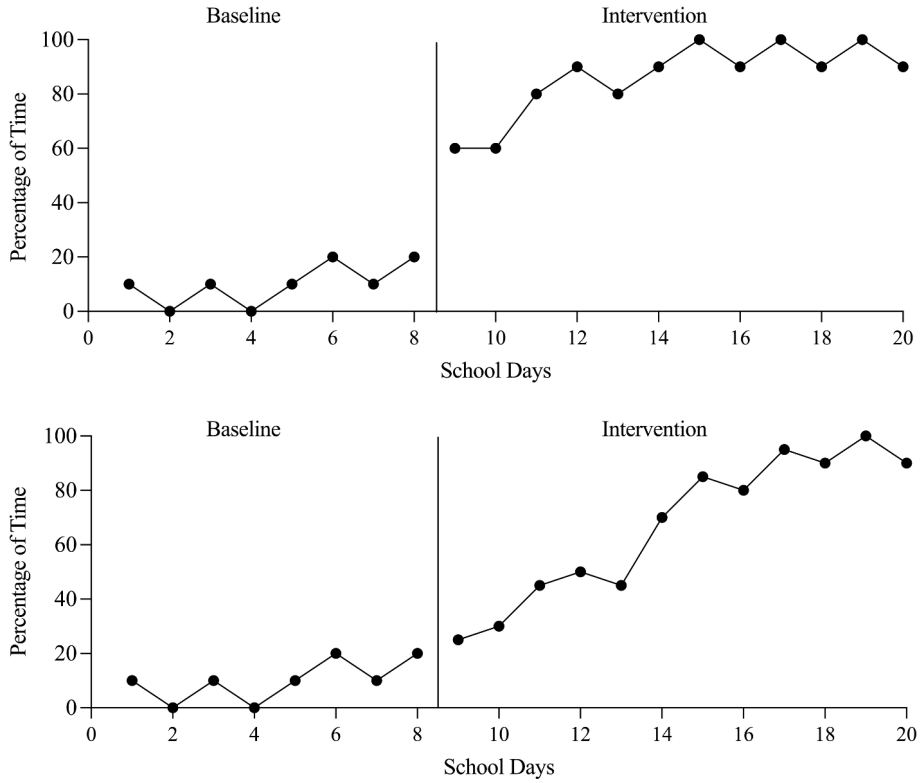

Figure 3.4 Maturation Effects with Minimal (top) and Considerable (bottom) Impacts on Confidence in Relation between Independent and Dependent Variables.

Note: In the top panel, the maturation effect is not worrisome because there is a clear and large change in behavior when the intervention condition begins. In the bottom panel, the maturation effect threatens internal validity because it is difficult to tell whether behavior gets to high levels because of maturation, the intervention, or an interaction of these factors.

scheduling sessions multiple times per week rather than one time per week, to reduce total time spent in each phase) and by assessing interventions that are likely to result in substantial and immediate improvements relative to baseline. Between-groups designs (see Chapter 1) are better suited for behaviors that are likely to change slowly over time during baseline *and* relatively slowly or with a delay during intervention.

Testing

Testing is a potential threat in any study that requires participants to respond to the same test repeatedly, especially during a baseline or probe condition; **testing effects** occur when repeated assessment tasks result in participant behavior change. Like history and maturation effects, testing effects can be detected via visual analysis of graphs.

Repeated testing may have a facilitative effect (improvement in performance over successive baseline or probe testing or observation sessions) or an inhibitive effect (deterioration in performance over successive baseline or probe testing or observation sessions) depending on how the "test" condition is designed. A test condition that repeatedly presents the same academic task, prompts correct responses through a correction procedure, or delivers reinforcement contingent upon a correct response, may result in a **facilitative testing effect**. This is, of course, beneficial for participants, but renders conclusions about subsequent intervention effects difficult to draw. Detecting facilitative testing effects in baseline often results in a decreased need for intervention; if researchers continue to intervene despite these effects, they produce a **testing threat**—which occurs when a history effect results in decreased internal validity and difficulty drawing conclusions about outcomes (see top panel of Figure 3.5). Test sessions of long duration, requiring substantial participant effort, with minimal or no reinforcement for attention and active participation may result in an **inhibitive testing effect**. If potential inhibitive effects are identified, baseline procedures can be modified to mitigate the effects to avoid a testing threat to internal validity that reduces internal validity. For example, the bottom panel of Figure 3.5 shows initial levels of behavior that ranged from 20–30% correct then decreased to 0% correct responding. This might happen, for example, if you failed to reinforce attempts or accurate responding during baseline conditions. A modified baseline (perhaps with additional instructions and opportunities for reinforcement) resulted in levels of behavior that was similar to that in initial sessions, perhaps representing a participant's "best effort."

Testing effects can be prevented by designing conditions so that they yield participants' best effort so that you neither overestimate nor underestimate the impact of the independent variable on the behavior. Facilitative effects of testing can be avoided by not reinforcing correct responses, particularly on receptive tasks; not correcting incorrect responses; and not prompting (intentionally or unintentionally) correct responses. Of course, these procedures (e.g., failing to reinforce correct responses) might be ethically or practically objectionable, and may impede later learning. Thus, you may decide that some level of facilitative procedures in baseline are acceptable. Procedural reliability checks will help with detecting procedural errors that could influence participant performance. Inhibitive effects of testing can be avoided by conducting sessions of an appropriate length and difficulty level (i.e., avoid session fatigue; intersperse known stimuli with unknown stimuli and reinforce correct responses to known stimuli; and reinforce correct responses on expressive, comprehension, and response chain tasks).

Instrumentation

Instrumentation threats refer to concerns with the measurement system that reduce confidence in outcomes (i.e., threaten internal validity); they are of particular concern in single case design studies because of repeated measurement by human observers who may make

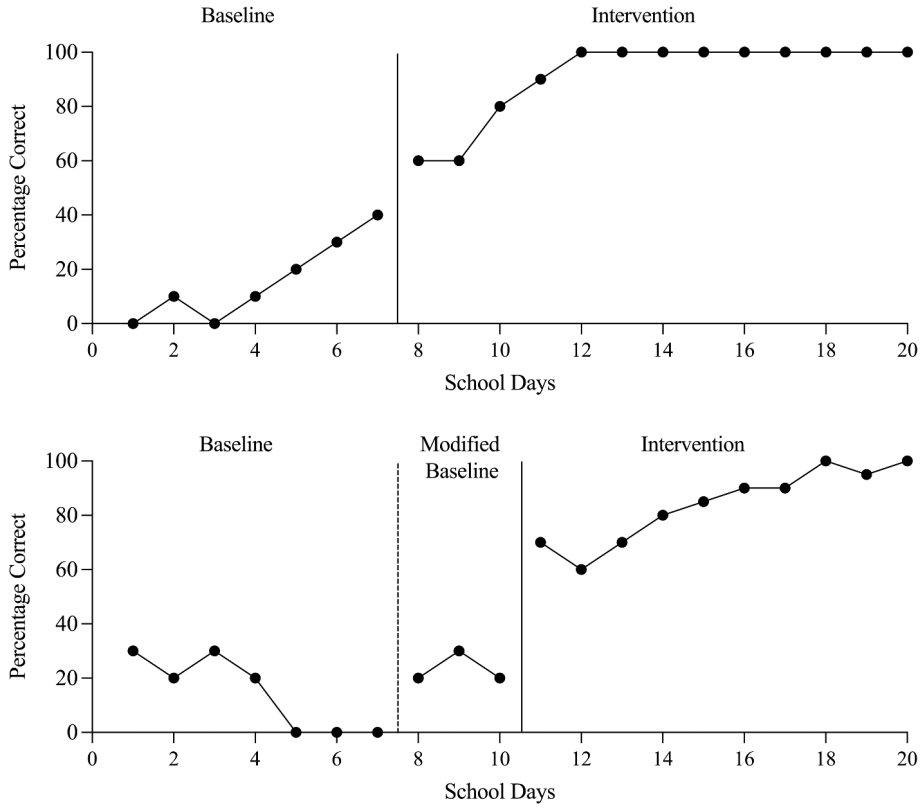

Figure 3.5 Testing Effects with Facilitative (top) and Detrimental (bottom) Effects on Baseline Performance.

Note. In the top panel, the participant could be learning the targeted behaviors due to simple repeated exposure or reinforcement for guessing in baseline, threatening internal validity because it isn't clear the extent to which changes in behavior are due to testing effects, the intervention, or a combination of those factors. In the bottom panel, failing to reinforce correct responding could have led to decreased effort for the participant and zero correct responding. Thus, a modification was required to get accurate baseline performance before initiating the intervention condition.

systematic errors. In studies using single case logic, the percentage agreement between two independent observers is the most common strategy for determining whether there is a threat to internal validity due to instrumentation (see more in Chapter 5). That is, two people observe the same event and record their decisions about how to characterize the event according to a specific set of rules. Afterward, they check to see the extent to which their decisions agree. Instrumentation threats are detected via analysis of these interobserver agreement (IOA) data. If agreement is low or changes over time, it may suggest that the data were not measured in a way consistent with the intent of researchers. You can avoid common problems by carefully defining behaviors of interest, using appropriate recording procedures, frequently checking for reliability by using a secondary observer, and visually analyzing data from both observers on the same graph (Artman et al., 2012; Ledford et al., 2012; Ledford & Wolery, 2013). Analyzing these data is covered in depth in Chapter 5.

Procedural Infidelity

Procedural infidelity refers to the lack of adherence to condition protocols by study implementers. It is detected via collection and analysis of *procedural fidelity data*—data that describe the extent to which implementers are implementing procedures as intended. Procedural infidelity might occur once (e.g., when a new research assistant has a misunderstanding about the situations under which she should reinforce a specific behavior and conducts one session using the wrong rules) or consistently (e.g., if teachers implementing your intervention consistently cannot engage in all of the planned behaviors in the complex setting of their classrooms). If the procedures of an experimental condition (baseline, probe, intervention, maintenance, generalization) are *consistently* not implemented as described in the Methods section of the research proposal or report, confidence that outcomes are related to the intervention is considerably reduced, resulting in a **procedural fidelity threat**. Further, even some momentary procedural fidelity lapses may result in uninterpretable data (e.g., Lambert et al., 2023). Procedural infidelity threats to internal validity can be avoided by defining clear rules for implementation, adequate implementer training and supports, and consistent collection and analysis of fidelity data, with modifications to training and supports provided as needed.

Selection Bias

Generally, participants are selected according to rules called *inclusion criteria*. **Selection bias** involves choosing participants using rules that are different than reported inclusion criteria and that are not apparent to non-researchers, and this is a threat when these rules are used in a way that differentially impacts the inclusion or retention of participants in a study, when compared to the population of interest. Several resources are available which discuss selection bias in group comparison designs (Pyrczak, 2016; Shadish et al., 2002). In single case research, the population refers to individuals who meet the inclusion criteria for the study and have similar functional characteristics to the participants (Lane et al., 2007; Wolery, Dunlap, & Ledford, 2011). For example, Ledford et al. (2017) included 12 children in a study to assess preference for massed versus embedded instruction, and named the following inclusion criteria: (1) ability to play developmentally appropriate games with turn-taking, (2) ability to make choices given line drawings, and (3) verbal imitation. Assume that Ledford and colleagues had 14 potential participants but decided to request consent from 12 due to resource constraints. Thus, she excluded two boys who had a history of not following directions during teacher-led activities (e.g., massed instruction) to reduce the risk of attrition. This decision leads to the potential for inaccurately identifying differential outcomes because of the purposeful exclusion of participants unlikely to perform well in one of the two conditions. As a side note, this particular hypothetical situation did not occur, but participants were chosen from a larger set of eligible students based on convenience, so sampling bias is still possible (e.g., we may have chosen students who had relatively high academic skills because students with higher support needs received more therapy and were thus available less frequently).

Attrition is the loss of participants during the course of a study, which can limit the generality of the findings, particularly if participants with certain characteristics are likely to drop out (e.g., participants who are not benefitting from the intervention). **Attrition threats** occur when participant loss (attrition) impacts the outcome of the study. Thus, when *any participant consents to participate in your study and does not complete the study*, you should always (1) explicitly report it, along with relevant information about why it occurred, and (2) include any data collected for that participant in your research report. This ensures that data from "non-responders" are not systematically excluded from published research, resulting in bias regarding evidence of intervention effectiveness. Preventing attrition may be difficult, but attrition may be less likely to occur when (1) baseline durations are effectively managed, (2) you choose

participants likely to benefit from the intervention, and (3) you are forthcoming about any difficulties that could be associated with research participation (e.g., uncomfortable baseline sessions, difficult-to-implement intervention components). All participants who are included in a study, even those who drop out and especially those who have unexpected response to intervention, should be included in all reports.

Multiple-Treatment Interference

Multiple-treatment interference (also called **multi-treatment interference**) can occur when a study participant's behavior is influenced by more than one planned intervention during the course of a study. These effects can be detected via visual analysis when appropriate designs are selected and suitable condition ordering is used (see Chapters 9 and 13). One type is sequential confounding (sequence effects), which refers to the influence of a participant's behavior that is due specifically to the order in which interventions are introduced. Another is a carryover effect, which occurs when a procedure used in one intervention condition influences behavior in an adjacent condition. It is important to note that multi-treatment interference refers to interference between *planned treatment conditions*, not from uncontrolled outside treatments (e.g., participant begins taking medication on the same day that you start your intervention condition); the latter is a history effect. You can prevent or detect multi-treatment interference by selecting appropriate designs, counterbalancing conditions (see Chapter 9), having sufficient phase lengths for reaching stability, and (when possible) providing participants with sufficient information about differences between conditions.

Data Instability

Instability refers to the amount of session-to-session change in the values of data (dependent variable); when you have variable data, it is difficult to predict the approximate value of the next data point. When you have stable data, it is relatively easy to predict a small range in which the next data would fall given no changes in condition. A **data instability threat** to internal validity occurs when the instability of data in one or more conditions makes it more difficult to draw conclusions about differences in outcomes across phases (i.e., decreased internal validity). Generally, instability threats are detected via visual analysis and prevented by changing conditions only when data are stable or when variability does not exceed expectations. When very large changes in outcomes between conditions are expected, more tolerance of variable data is allowable. When small, delayed, or variable changes in outcomes are expected, instability will result in a threat to internal validity. Thus, conservative researchers will wait until baseline data are stable, regardless of behavior change expectations, to ensure data instability threats will not materialize. The top panel in Figure 3.6 shows instability in a baseline phase with minimal influence on conclusions about behavior change (i.e., even though data are variable in baseline, the change in outcomes between phases is apparent). The bottom panel in Figure 3.6 shows instability in the baseline that makes it difficult to confidently conclude that outcome changes between phases are due to planned differences between conditions alone (i.e., a threat to internal validity).

Data instability (also referred to as variability) can result in a specific threat, referred to as **regression to the mean**. Regression to the mean refers to the likelihood that following an outlying data point, data are likely to revert to levels closer to the average value. For example, suppose you are hoping to intervene to increase behavior occurrence, and data are somewhat low (e.g., 30%) for the first three data points. For the fourth data point, values drop all the way to 0%. Some would say that this is a clear indication that intervention is needed; however, even without intervention, data are likely to improve after this outlying value. Changing conditions at this point can decrease confidence that your intervention, rather than typical variability, is

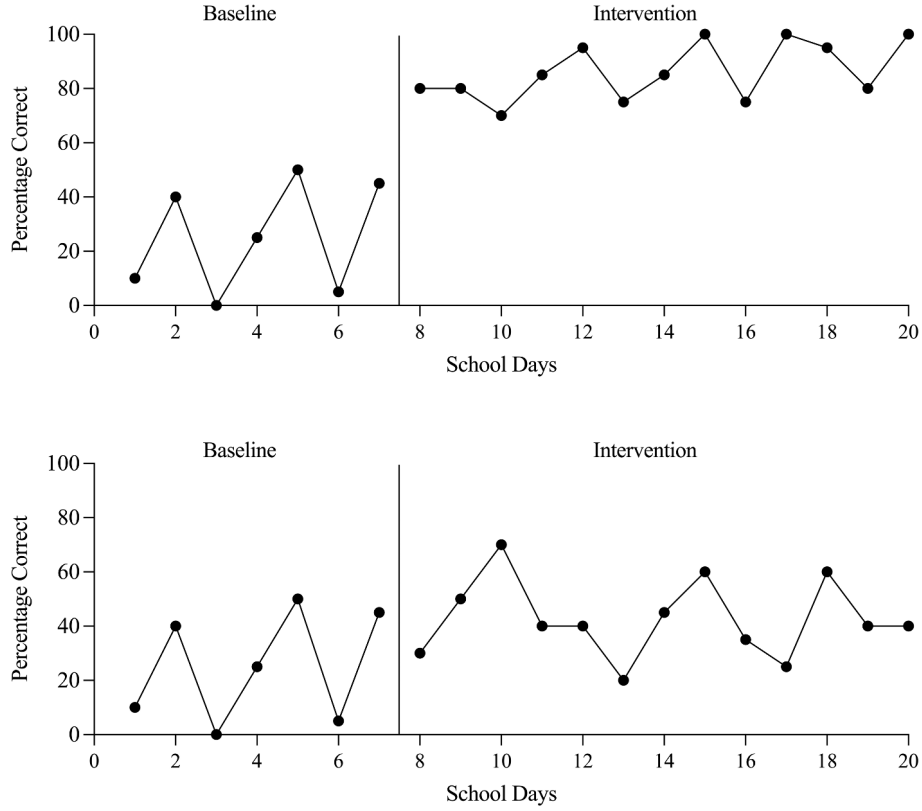

Figure 3.6 Data Instability Effects with Minimal (top) and Considerable (bottom) Impacts on Confidence in Relation between Independent and Dependent Variables.

Note: Data instability is depicted in both panels, with especially large variability in baseline conditions. Although instability (variability) is identical in both panels, confidence is less impaired in the top panel due to large changes between conditions. In the bottom panel, variability makes confidence in relations between behavior change and the intervention low.

the cause. You can avoid threats associated with regression to the mean, continue collecting data until stability is established. **Cyclical variability** is a specific type of data instability that refers to a repeated and predictable pattern in the data series over time. When phases are of equal length (e.g., five days in each condition) it is possible that your observations coincide with some unidentified natural source that may account for the variability. For example, if your experimental schedule coincides with a parent's work schedule (away from home for five days, at home for five days) you may incorrectly conclude that the independent variable is responsible for changes in behavior when in fact it may be due to the presence or absence of the parent at home. To avoid confounding due to cyclical variability it is recommended that you vary phase lengths across time.

Adaptation

Adaptation refers to a period of time at the start of an investigation in which participants' recorded behavior may differ from their natural behavior due to the novel conditions under which data are collected. **Adaptation threats** occur when adequate measures are not taken to ameliorate these effects prior to data collection beginning; this results in changes in baseline data that make changes between phases more difficult to interpret. The **Hawthorne Effect**,

which refers to participants' observed behavior not being representative of their natural behavior as a result of their knowledge that they are participants in an experiment (Kratochwill, 1978; Portney & Watkins, 2000), is a specific type of adaptation threat to validity. We recommend study participants be exposed to unfamiliar adults, settings, formats, and data collection procedures (e.g., video recording) prior to the start of a study (sometimes referred to as history training), to increase the likelihood that data collected on the first day of a baseline phase is representative of participants' "true" behavior. You can also avoid adaptation effects by reducing the obtrusiveness of your measurement procedures and using endogenous implementers rather than researchers when that is feasible.

Design-Related Confounds

One important internal validity consideration, undiscussed in previous versions of this text, and to our knowledge, generally unconsidered in single case design, is the impact of design on outcomes. That is, the changes in behavior that occur in your experiment may be due—in whole or part—to the design used to assess the outcomes, rather than the condition procedures themselves. Although this threat has not been widely discussed, some authors have noted that design-specific effects might have influenced findings. For example, Chazin and Ledford (2021) made the following point about their findings regarding efficiency differences between two prompting procedures:

> Because conditions alternated rapidly, children in the SLP superior group may have been unable to discriminate when incorrect responding would result in error correction (CTD) versus an additional learning opportunity (SLP)…researchers in future studies should pre-teach use of waiting versus responding incorrectly; they could then indicate during alternating instructional sessions that responding incorrectly is acceptable for SLP while it is not for CTD. This would ensure that only children who have appropriate prerequisite skills for both the procedures themselves and those required due to the study design (e.g., alternation of procedures) are included in the study.
>
> (underlined emphasis ours)

Thus, authors of that study acknowledged that the apparent differences between conditions in the study could have either been due to intervention effects or due to *design* effects. This is different from multi-treatment interference because it is unrelated to the interventions being implemented and instead is due to the ordering of conditions (in this case, the rapid alternation, which is required given the design type used in the study). Preventing and detecting these threats require a comprehensive understanding of the theory and processes driving behavior change in a given experiment (see discussions about Logic Models in Chapter 6) and the logic and use of different types of single case designs. We will discuss specific considerations for preventing and detecting these threats for different design types in Chapters 9, 11, and 13.

Conclusions

In this chapter, we discussed a cornerstone in all research—replication. Single case research relies on replication to increase confidence that changes in the *dependent variable* between conditions are due to changes in the *independent variable* between conditions, and only to those changes. We also introduced some potential threats to internal validity, which reduce confidence in conclusions that we can draw from an individual single case study. In subsequent chapters, we will address many of these threats more thoroughly, and provide detailed guidelines for avoiding, detecting, and minimizing these threats.

References

Artman, K., Wolery, M., & Yoder, P. (2012). Embracing our visual inspection and analysis tradition: Graphing interobserver agreement data. *Remedial and Special Education*, *33*(2), 71–77.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

Chazin, K. T., & Ledford, J. R. (2021). Constant time delay and system of least prompts: Efficiency and child preference. *Journal of Behavioral Education*, *30*, 684–707.

Coyne, M. D., Cook, B. G., & Therrien, W. J. (2016). Recommendations for replication research in special education: A framework of systematic, conceptual replications. *Remedial and Special Education*, *37*(4), 244–253.

Eyler, P. B., & Ledford, J. R. (2023). Systematic review of time delay instruction for teaching young children. *Journal of Early Intervention*, 10538151231179121

Gage, N. A., Cook, B. G., & Reichow, B. (2017). Publication bias in special education meta-analyses. *Exceptional Children*, *83*(4), 428–445.

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

Lambert, J. M., Copeland, B. A., & Alexandrova, M. (2023). Reinforcer value moderates response magnitude and persistence during extinction: A randomized trial. *Journal of Applied Behavior Analysis*.

Lane, K., Wolery, M., Reichow, B., & Rogers, L. (2007). Describing baseline conditions: Suggestions for study reports. *Journal of Behavioral Education*, 16, 224–234.

Ledford, J. R., Chazin, K. T., Harbin, E. R., & Ward, S. E. (2017). Massed trials versus trials embedded into game play: Child outcomes and preference. *Topics in Early Childhood Special Education*, 37, 107–120.

Ledford, J. R., Lambert, J. M., Pustejovsky, J. E., Zimmerman, K. N., Hollins, N., & Barton, E. E. (2023). Single-case-design research in special education: Next-generation guidelines and considerations. *Exceptional Children*, *89*(4), 379–396.

Ledford, J. R., & Wolery, M. (2013). Effects of plotting a second observer's data on ABAB graphs when observer disagreement is present. *Journal of Behavioral Education*, *22*, 312–324.

Ledford, J. R., Wolery, M., Meeker, K. A., & Wehby, J. H. (2012). The effects of graphing a second observer's data on judgments of functional relations in A–B–A–B graphs. *Journal of Behavioral Education*, *21*, 350–364.

Lemons, C. J., Fuchs, D., Gilbert, J. K., & Fuchs, L. S. (2014). Evidence-based practices in a changing world: Reconsidering the counterfactual in education research. *Educational Researcher*, 43, 242–252.

Locey, M. L. (2020). The evolution of behavior analysis: Toward a replication crisis?. *Perspectives on Behavior Science*, *43*, 655–675.

Peredo, T., Zelaya, M., & Kaiser, A. (2018). Teaching low-income Spanish-speaking caregivers to implement EMT en Español with their young children with language impairment: A pilot study. *American Journal of Speech-Language Pathology*, *27*(1), 136–153.

Petursdottir, A. I., & Carr, J. E. (2018). Applying the taxonomy of validity threats from mainstream research design to single case experiments in applied behavior analysis. *Behavior Analysis in Practice*, *11*, 228–240.

Portney, L., & Watkins, M. P. (2000). *Foundations of clinical research: Applications to practice*. Upper Saddle River, NJ: Prentice Hall.

Pyrczak, F. (2016). *Making sense of statistics: A conceptual overview*. London: Routledge.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*, 487–510.

Sidman, M. (1960). *Tactics of scientific research—Evaluating experimental data in psychology*. New York: Basic Books.

Tincani, M., & Travers, J. (2019). Replication research, publication bias, and applied behavior analysis. *Perspectives on Behavior Science*, *42*, 59–75.

Wolery, M., Dunlap, G., & Ledford, J. R. (2011). Single case experimental methods: Suggestions for reporting. *Journal of Early Intervention*, 33, 103–109.