THIRD EDITION

Single Case Research Methodology

Applications in Special Education and Behavioral Sciences

Edited by
Jennifer R. Ledford
and David L. Gast

ROUTLEDGE

Single Case Research Methodology

Single Case Research Methodology, 3rd Edition presents a thorough, technically sound, user-friendly, and comprehensive discussion of single case research methodology. This book can serve as a detailed and complex reference tool for students, researchers, and practitioners who intend to conduct single case research design studies; interpret findings of single case design studies; or write proposals, manuscripts, or reviews of single case methodology research. The authors present a variety of single case research studies with a wide range of participants, including preschoolers, K-12 students, university students, and adults in a variety of childcare, school, clinical, and community settings, making the book relevant across multiple 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.

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

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

Single Case Research Methodology

Applications in Special Education and Behavioral Sciences

Third Edition

Edited by Jennifer R. Ledford and David L. Gast

Routledge
Taylor & Francis Group
NEW YORK AND LONDON

Third edition published 2018
by Routledge
711 Third Avenue, New York, NY 10017

and by Routledge
2 Park Square, Milton Park, Abingdon, Oxon, OX14 4RN

Routledge is an imprint of the Taylor & Francis Group, an informa business

First edition published by Routledge, 2009

Second edition published by Routledge, 2014

Library of Congress Cataloging-in-Publication Data

Names: Ledford, Jennifer R., editor. | Gast, David L., editor.
Title: Single case research methodology / [edited by] Jennifer R. Ledford, David L. Gast.
Description: Third Edition. | New York : Routledge, 2018. | Revised edition of Single case research methodology, 2014.
Identifiers: LCCN 2017040311 (print) | LCCN 2017043364 (ebook) | ISBN 9781315150666 (e-book) | ISBN 9781138557116 (hardback) | ISBN 9781138557130 (pbk.) | ISBN 978-1-315-15066-6 (ebk) | ISBN 9781315150666 (ebk)
Subjects: LCSH: Single subject research. | Psychology—Research. | Educational psychology—Research.
Classification: LCC BF76.6.S56 (ebook) | LCC BF76.6.S56 G37 2018 (print) | DDC 300.72/1—dc23
LC record available at https://lccn.loc.gov/2017040311

ISBN: 978-1-138-55711-6 (hbk)
ISBN: 978-1-138-55713-0 (pbk)
ISBN: 978-1-315-15066-6 (ebk)

Typeset in Minion
by Apex CoVantage, LLC

Visit the eResources: www.routledge.com/9781138557130

We dedicate this edition to Dr. Mark Wolery, who has had immeasurable influence in special education, single case research methodology, and our lives. He is the best colleague, mentor, and friend.

Contents

Preface

This third edition of *Single Case Research Methodology* was edited to include information regarding contemporary developments in single case experimental design, while retaining an emphasis on both historical precedent and 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 under the tutelage of 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 share their work with you via this text.

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 the book to serve as a detailed and complex reference tool for students, researchers, and practitioners who intend to conduct single case research design studies; interpret findings of single case design studies; or write proposals, manuscripts, or reviews of single case methodology research. We expect that these students, researchers, and practitioners 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. Throughout the book, we present a variety of single case research studies with a wide range of participants, including preschoolers, K-12 students, university students and adults in a variety of childcare, school, clinical, and community settings. Many studies have included young children with disabilities or individuals with significant behavioral or cognitive challenges; a large proportion of high-quality single case research has been conducted in these areas. However, we continue to encourage work in related areas, which is becoming increasingly common.

The organization of this edition is largely in keeping with previous editions, with information divided into 14 chapters for ease of using the text in a semester-long course in single case design. Early chapters focus on general information about research (Chapter 1), ethics (Chapter 2), writing tasks (Chapter 3), and replication logic (Chapter 4). Next, we focus on measurement, with a specific emphasis on measurement of dependent variables (Chapter 5) and independent variables and social validity (Chapter 6) in the context of single case design. Then, we discuss presentation (Chapter 7) and analysis of graphed data (Chapter 8). In Chapters 9–12, we present specific information on different design types, including A-B-A-B withdrawal and reversal designs (Chapter 9), multiple baseline and multiple probe designs (Chapter 10), comparison designs (Chapter 11), and other designs (Chapter 12). Finally, in two new chapters, we discuss evaluating quality and rigor of single case designs (Chapter 13) and systematic synthesis of findings across studies (Chapter 14).

Whether you are a student, practitioner, researcher, or have some other role in relation to conducting or interpreting single case research, the information presented in this book is intended to assist you to understand the logic behind single case research design, controlling for alternative explanations, conditions for the use of SCD, and how to conduct reliable and valid measurement in the context of single case research. The guidelines presented in this text are intended to assist you in the design, analysis, implementation, and dissemination of single case research. Given a thorough understanding of the workings of single case research, you can conduct well-designed studies to assist in accumulating meaningful evidence regarding effective interventions and play a consequential role in moving your knowledge and your field forward. Good luck!

JRL

Acknowledgements

We thank all of the hard-working students and professionals who contributed their time to chapters in this text, our current and former students who have asked good questions and made us better researchers, and especially Kate Severini and Katie Zimmerman, who read numerous chapters in this edition and provided valuable feedback about content and clarity.

1
Research Approaches in Applied Settings

David L. Gast and Jennifer R. Ledford

Important Terms

applied research, independent variables, dependent variables, internal validity, experimental control, functional relation, evidence-based practice, reliability, threats to internal validity, nomothetic, baseline logic, ideographic, validity, history, maturation, testing, instrumentation, procedural infidelity, attrition, attrition bias, sampling bias, data instability, cyclical variability, multitreatment interference, regression to the mean, adaptation, Hawthorne effect

The goal of science is to advance knowledge. The process by which we advance knowledge is generally via research—the systematic investigation and manipulation of variables to identify associations and understand processes that occur in typical (non-research) contexts. Of course, research processes are limited; for example, outcomes of research studies have been reported to be non-replicable (Open Science Collaboration, 2015); to be dependent on counterfactual conditions (Lemons, Fuchs, Gilbert, & Fuchs, 2014); to fail to generalize to outside of research contexts, in applied or authentic settings (Spriggs, Gast, & Knight, 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 levels of evidence based on research type, and explain three primary research approaches and their corresponding rationales and assumptions. We conclude the chapter by describing similarities and differences between research and practice.

Applied Research

If research is a set of processes by which we produce information about associations and processes of interest, what then is 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, basic research might inform science related to the association of running and behavioral abnormalities in a mouse model of Down syndrome (Kida, Rabe, Walus, Albertini, & Golabek, 2013). *Applied* research might seek to identify interventions that result in improved physical activity for young children with Down syndrome (Adamo et al., 2015). Researchers and practitioners often seek to engage in applied research to not only add to the knowledge base for a specific topic, but also to improve outcomes of specific participants (researchers) or clients (practitioners). We refer to practitioners who engage in research as scientist-practitioners (a label coined by Barlow, Hayes & Nelson in 1984 to describe interventionists who make data-based decisions an integral part of their practice). In applied research, we are most interested in determining the relation between **independent variables**—the variables manipulated by researchers (i.e., interventions) and **dependent variables**—the variables we expect to change given the manipulation (i.e., target behaviors), to solve problems of clinical and educational practice.

Integrating Science Into Educational and Clinical Practice

Is it possible to incorporate scientific methodology into the daily routine of practitioners in schools, clinics, and the community? It is, but it's not an easy task. Conducting applied research in authentic settings has the potential to advance science, to document changes in behavior, and to establish responsibility for the change. Before moving on to the research task itself, we would like to elaborate on the importance of these goals.

Advancement of Science

Through the work of Skinner and Bijou, a system of behavior analysis has been developed that includes a philosophy of behavior development, a general theory, methods for translating theory into practice, and a specific research methodology. This system was new in the scope of human evolution and the advancement of science. It has gained acceptance and verification through the successful application of concepts and principles. One general "test" of the system has been the demonstration of effectiveness in a variety of settings, in basic and applied applications. Applied behavioral analysis has been adopted and made an integral part of special and general education, speech language therapy, clinical and school psychology, neuropsychology, recreation therapy, adaptive physical education, and many other disciplines. Applied research, focused on specific problems of learning and reinforcement in schools, clinics, and communities, supports the advancement of science and knowledge in a given field while also making a direct impact on clients and consumers.

Not all practitioners may choose to be applied researchers, especially given the complexities of conducting applied research in authentic settings; however, most practitioners can contribute to the advancement of science and their discipline, by collaborating with those who do. Likewise, researchers and scientists can contribute to practice and enhance the applicability of their research by collaborating with practitioners. Eiserman and Behl (1992) addressed researcher- practitioner collaboration in their article describing how special educators could influence current best practice by opening their classrooms to researchers for the purpose of systematic research efforts. They pointed out the potential benefits of such collaborations, not the least of which was teachers becoming interested in conducting their own research and bridging the gap between research and practice (p. 12). More recently, Snow (2014) suggested educational research should include more collaboration with practitioners, to address applied problems. This position is not new, and that single case designs (SCDs) 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). Encouragement of practitioner involvement in applied research efforts, as defined by Baer, Wolf, and Risley

(1968, 1987), acknowledges their potential contribution by addressing "real" problems, which need to be addressed under "real" conditions, with available resources. It cannot be overstated that practitioners are often confronted with issues or problems overlooked by researchers. Thus, if practitioners collaborate with researchers, or acquire the skills to conduct their own research, they can generate answers to questions that will advance science for issues that are relevant to practice.

Advancement of Practice

Applied researchers in education, psychology, speech pathology, occupational therapy, and related fields have conducted experiments in controlled environments (lab schools, research institutes, private clinics, medical centers) by highly educated research professionals who have access to resources beyond those typically available. Research generated in such centers is important to advancing our understanding of human behavior and how to positively effect change, however, the extent to which effective interventions generalize to settings outside these "resource rich" and controlled environments needs to be shown. Thus, there are many research possibilities that the teacher/therapist-researcher can conduct in their classroom or community- based clinic that will add to our understanding on how to better serve those under their care.

Baer et al. (1987) addressed the need for applied researchers to determine the context with which interventions succeed and fail. When research is conducted under highly controlled conditions, as is often the case in studies using SCDs, the ability of those working in "typical" or "authentic" community settings to replicate conditions may be difficult, if not impossible. 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. It is important for applied researchers to identify the versatility and latitude of a particular intervention prior to advocating its use. In fact, through "failures to replicate" we seek out answers to "why?", and with perseverance, identify modifications to the original intervention that result in the desired behavior change. Such discoveries are important to the advancement of practice in that our goal is for changes in behavior to generalize and maintain in natural environments. Through collaboration with applied researchers, the contribution made by teachers and therapists will increase the probability that instructional strategies and interventions under study will improve practice as delivered by other teachers and therapists working in community schools and clinics. Moreover, the cross-discipline emphasis on implementation science (Cook & Odom, 2013; Forman et al., 2013) has clearly established that the likely implementation of an intervention, given typical contexts and supports, is a critical component of studying evidence-based practices. The applied researcher who demonstrates positive changes in participants' academic, adaptive, or social behavior, produces evidence of a benefit of the instructional process.

Empirical Verification of Behavior Change

Successful teachers and therapists must demonstrate that they can bring about positive behavior change in their students or clients. Practitioners should expect that increasingly informed parents and clients will ask for data on behavior change for meaningful outcomes, and then will ask for some verification that your efforts were responsible for that change. Advances in technology have made collecting, organizing, presenting, and sharing data increasingly accessible. Practitioners who use practices and collect data on client or student behavior can show behavior change that occurs over time; however, sometimes behavior change may be the result of other factors (e.g., additional services of which the practitioner was unaware). The utilization of experimental research designs, such as SCDs, allows the practitioner to go one step further—to show a causal link between his or her practices and the child's behavior change. A study with adequate mechanisms for ensuring that outcomes are related to your intervention procedures rather than extraneous factors is said to have adequate **internal validity**. Studies with high levels of internal validity allow researchers to demonstrate **experimental control**—to show that the experimental procedures (intervention) and *only* the experimental procedures are responsible for behavior change. A researcher does this by carefully eliminating other potential explanations for behavior change; this concept will be discussed at length in later chapters. When experimental control is demonstrated, we have verified that there is a **functional relation** between the independent and dependent variables—that is, that the change in the dependent variable (behavior) is causally (functionally) related to the implementation of the independent variable.

Evidence-Based Practice

At no time in history has accountability in education, psychology, behavior sciences, and related fields been more important. Recent guidelines in the Individuals with Disabilities Education Improvement Act (IDEIA) and the Every Student Succeeds Act (ESSA) mandate the use of *evidence-based practice* (alternately, "scientific, research-based intervention"; IDEIA; or "empirically supported practice"; Ayres, Lowrey, Douglas, & Sievers, 2011). Similarly, the American Psychological Association and the Behavior Analysis Certification Board have standards requiring the use of evidence-based interventions. **Evidence-based practice** 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. Though the term is relatively new, the idea that research should guide practice is not, particularly in the field of applied behavior analysis. Baer et al. (1968) defined applied behavior analysis and emphasized the importance of quantitative research-based decisions for guiding practice. Their emphasis on a low-inference decision model, based on repeated measurement of behavior within the context of an SCD, set a standard for practitioners determining intervention effectiveness 50 years ago. At the time of their article, published in the inaugural issue of the *Journal of Applied Behavior Analysis*, there was no shortage of critics who questioned the viability and desirability of an empirical scientific approach for studying and understanding human behavior, a response in part due to the controversial position articulated by B.F. Skinner in his classic book, *Science and Human Behavior* (1953). Having passed the test of time, as evidenced by the numerous SCD studies that have influenced practice across many disciplines, it has been shown that a behavioral approach can and does provide a scientific framework for understanding and modifying behavior in positive ways. Few would question that Baer et al. (1968) established evidence-based practice as a core value for applied behavior analysts, a value that has yielded best and promising practices across numerous disciplines within the behavioral sciences. Current zeitgeist and standards continue this long-standing tradition for researchers and practitioners in a variety of fields.

What constitutes a "practice"? Horner et al. (2005) defined *practice* as it relates to education as "a curriculum, behavioral intervention, systems change, or educational approach designed to be used by families, educators, or students with the express expectation that implementation will result in measurable educational, social, behavioral, or physical benefit" (p. 175). This definition applies to specific interventions and broader approaches used by professionals who provide educational and clinical services. It should not go unnoticed that the definition includes mention of a "measurable" benefit to those who are the focus of the practice.

What constitutes *evidence* that supports implementation of a particular practice? Must evidence be quantitative? Is clinical or professional judgment a consideration? Answers

to these questions are important since different research methods and designs yield different types of data. *The research question should determine the research method (group, single case, or qualitative) and design chosen.* 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 (see Chapters 4–6). As a scientist you will be expected to see things as they are, not as you wish them to be; this will necessitate ensuring **reliability** (i.e., consistency) by defining the target behavior (or event) clearly and concisely so that two independent observers consistently agree on scoring what they observe. Finally, you will need to be patient to see if your research findings stand up to the scrutiny of other researchers when they attempt to replicate your results. This latter criterion is critical, as *replication is at the heart of the scientific method*, without which you cannot have confidence in study findings.

Behavioral scientists have numerous scientific research designs from which to choose in their quest for answers to research hypotheses and questions. There is general agreement among researchers that different research questions or objectives require different research approaches—no one research method or design is appropriate for answering all research questions. However, for behavioral scientists, certain research methods and designs are deemed superior to others when generalizing findings to individuals or groups. This judgment is based on the degree to which data collection procedures, data analyses, and data reporting are viewed as objective, reliable and valid, and the extent to which the study can be replicated while yielding similar findings. Studies that are based on investigator perceptions and descriptions, that fail to objectively define and evaluate the reliability of investigator observations, and that lack detailed descriptions of conditions under which data are collected (thus making replication difficult if not impossible), are judged as lacking scientific rigor and "trustworthiness" of findings. Judging the rigor of the scientific method of a study that supports a particular practice is at the heart of determining whether a practice is evidence-based.

To that end—determining the rigor of the science supporting a particular policy, procedure, or practice—most professional organizations have recommendations and guidelines on their websites for evaluating research study adequacy (e.g., American Psychological Association, www.apa.org; American Speech-Language-Hearing Association, www.asha.org; Association for Behavior Analysis International, www.abainternational.org; Council for Exceptional Children, www.cec.org; etc.). Odom et al. (2005) point out that interest in and guidelines for the evaluation of research supporting clinical and educational practices has been addressed by medical, social science, and educational professional organizations for many years. As a result of ESSA and its predecessor, the No Child Left Behind (NCLB) legislation, families are increasingly holding professionals accountable for their choice of practices. Parents and other stakeholders expect to see positive changes in behavior, an expectation that is both

reasonable and consistent with ethical standards of educational and clinical professional organizations. Take for example an excerpt from a Policy Statement on "The Right to an Effective Behavioral Treatment" passed by the Association for Behavior Analysis International (ABAI) membership in 1989, which reads:

> An individual is entitled to effective and scientifically validated treatment; in turn, the behavior analyst has an obligation to *use only those procedures demonstrated by research to be effective*. Decisions on the use of potentially restrictive treatment are based on consideration of its absolute and relative level of restrictiveness, the amount of time required to produce a clinically significant outcome, and the consequences that would result from delayed intervention [italics added]
>
> (Van Houton et al., 1989, para. 8).

Applied behavior analysts have historically held themselves accountable for designing and employing curricula, interventions, systems for change, and educational/therapeutic approaches that bring about positive behavior change. As will be discussed throughout this book, SCDs will permit researchers and scientist-practitioners to repeatedly evaluate practices, suggesting continued use when data support their effectiveness; informing modifications when progress is slow or plateaus; and suggesting replacement when behavior change does not occur. These research decisions can be made while retaining the experimental integrity of a study if you are familiar with measurement and design guidelines presented in later chapters. To determine whether a given intervention is an evidence-based practice, multiple agencies have suggested guidelines, including the Institute of Education Sciences and Council for Exceptional Children; we will discuss those further in Chapter 13.

Dissemination of Evidence-Based Practices in Education

It is important that practices supported by research be disseminated to practitioners. To that end, the Education Science Reform Act of 2002 was established within the U.S. Department of Education's Institute of Education Sciences (IES; www.ed.gov/about/offices/list/ies); its mission, to "provide rigorous evidence on which to ground education practice and policy" (Institute of Education Sciences, n.d., para. 1) by government funded research projects. IES's oversight responsibilities were a direct response to concerns regarding the quality of educational research and the requirement put forth in NCLB that teachers use scientifically proven practices (Odom et al., 2005). To disseminate its findings, IES established the What Works Clearinghouse (WWC; http://ies.ed.gov/ncee/wwc) to inform stakeholders (teachers, researchers, community members, policymakers) by providing a source of information regarding scientific evidence of effectiveness for education practices that could be used to encourage making informed and data-based decisions and in turn improve child outcomes.

Prior to 2006 the WWC only "certified" and disseminated practices that were shown to be effective by a randomized experimental group design or random clinical trial. However, in September 2006, in one of its technical working papers, it revised its guidelines to include three additional research designs (provided they met certain basic standards regarding rigor): quasi- experimental, regression discontinuity, and SCDs. This policy revision showed an understanding by IES and WWC that applied research studies, particularly studies conducted with low-incidence populations and conducted in clinical and classroom settings, may require research designs other than those that require random assignment of participants to experimental conditions. Standards for evaluating SCDs were published in 2010, and include systematic manipulation of an independent variable (intervention) with evidence of adequate implementation, and reliable and repeated measurement of a dependent variable (e.g., participant behavior) in multiple conditions. These recommendations, and additional recommendations related to the analysis of data from single and multiple studies, are discussed in detail in Chapters 13 and 14. WWC has designated one evidence-based practice based solely on evidence from studies using SCD research (functional behavior assessment; WWC, 2016).

Characterizing Designs Based on Attributions of Causality

Experimental design studies are defined by an investigator's manipulation of an independent variable to verify what effect it has on a dependent variable. The act of intentionally manipulating some variable to see if there is a measurable change in a behavior *while controlling for probable other reasons for behavior change* differentiates experimental research from other research approaches. Appropriately utilized SCDs can be categorized as experimental (Horner et al., 2005). Experimental studies include (a) descriptions of the target behavior(s), (b) predictions regarding what impact the independent variable will have on the dependent variable(s), and (c) appropriate tests to see if the prediction is correct. In doing this, the research design must control for alternative explanations for the observed behavior change(s).

What differentiates an experimental design study from a *quasi-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. Within the context of the group research design approach, this differentiation is based on how research participants are assigned to study conditions. In experimental group design studies participants are randomly assigned to a study condition (e.g., experimental group or control group; intervention A or intervention B), while quasi-experimental group design studies do not use random assignment of participants but other strategies to control for differences in study group composition (e.g., counterbalancing techniques, participant matching; Fraenkel & Wallen, 2006). In SCD, studies are considered experimental, rather than quasi-experimental, if there are adequate potential demonstrations of effect—this concept will be elaborated on in the remaining chapters.

In experimental designs, if the prediction "proves" true, it is said there is a functional relation (i.e., cause-effect relation) between independent and dependent variables. The demonstration of a functional relation adds evidence in support of the independent variable being a promising and possibly "best practice" if findings are independently replicated. Greater support is attributed to results of an experimental group design study, compared to a quasi-experimental group design study, because of the random assignment of participants. Within SCD, which can also be experimental, randomization of participants is generally neither feasible nor helpful; randomization only functions to control for differences between groups when the number of participants is very large (e.g., N=50 or greater; see Chapter 13 for more information regarding randomization in SCD studies).

Correlational design studies, like experimental and quasi-experimental design studies, predict and describe the relation between independent and dependent variables; however, *in correlational studies there is no manipulation of the independent variable by the investigator*. Such studies represent a quantitative-descriptive research approach in

which the relation between variables is established by using a correlation coefficient (Fraenkel & Wallen, 2006). When independent and dependent variables co-vary there is said to be a correlational relation between variables. Practices supported by correlational evidence are deemed less trustworthy or convincing than those supported by experimental and quasi-experimental evidence since correlational design studies do not rule out alternative explanations because there is no manipulation of the independent variable. In a correlational study, for example, you might find that the number of hours a child spends with other children is correlated with his antisocial behaviors (e.g., more hours with children is related to higher levels of anti-social behavior). But, other causes of antisocial behavior are not ruled out in this example (for instance, children who spend many hours with other children might be in low-quality child care—the lack of access to appropriate services may be the reason for anti-social behavior). Some SCD studies (e.g., A-B designs, see Chapter 9) can be considered correlational (rather than causal or experimental) in nature.

Characterizing Designs Based on Research Approach

As the book title connotes, the focus of this text is on SCD research methodology and its use by applied researchers in behavioral sciences. In spite of this focus on a single type of research design, it is important for you 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. Through your analysis and understanding of research approaches you will be better able to choose the appropriate research design for answering your research question(s). As previously noted, no single research approach or design is appropriate for answering all research questions. Thus it is your responsibility, both as a consumer of and contributor to research, to be familiar with the various research approaches. 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 deMarrais and Lapan (2004), Fraenkel and Wallen (2006), Portney and Watkins (2000), and Schlosser (2003).

Before describing the individual approaches, it might be helpful to introduce concepts of nomothetic and idiographic research. **Nomothetic** research approaches are generally based in the natural sciences and are characterized by attempting to explain associations that can be generalized to a group given certain characteristics. **Idiographic** approaches to research, 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) although some have argued that an idiographic approach is most appropriate for practice, at least in the field of special education (Deno, 1990).

Group Research Approach

Gersten, Fuchs, Coyne, Greenwood, and Innocenti (2005) provide an excellent discussion of indicators for evaluating scientific rigor of group experimental and quasi-experimental research reports and proposals. Much of what is presented in this section is a summary of key points they present in determining the level of support assigned to group studies investigating the efficacy of a practice. They point out that there was not complete agreement among authors on all issues discussed. Nevertheless their presentation provides a framework from which to judge the level of support for an evidence-based practice with group designs. Table 1.1 summarizes the "Essential and Desirable Quality Indicators for Group Experimental and Quasi-Experimental Research Articles".

Table 1.1 Essential and Desirable Quality Indicators for Group Experimental and Quasi-Experimental Research Articles and Reports.

Essential Quality Indicators

Quality Indicators for Describing Participants

1. Was sufficient information provided to determine/confirm whether the participants demonstrated the disability(ies) or difficulties presented?
2. Were appropriate procedures used to increase the likelihood that relevant characteristics of participants in the sample were comparable across conditions?
3. Was sufficient information given characterizing the interventionists or teachers provided? Did it indicate whether they were comparable across conditions?

Quality Indicators for Implementation of the Intervention and Description of Comparison Conditions

1. Was the intervention clearly described and specified?
2. Was the fidelity of implementation described and assessed?
3. Was the nature of services provided in comparison conditions described?

Quality Indicators for Outcome Measures

1. Were multiple measures used to provide an appropriate balance between measures closely aligned with the intervention and measures of generalized performance?
2. Were outcomes for capturing the interventions effect measured at the appropriate times?

Quality Indicators for Data Analysis

1. Were the data analysis techniques appropriately linked to key research questions and hypotheses? Were they appropriately linked to the limit of analysis in the study?
2. 2. Did the research report include not only inferential statistics but also affect size calculations?

Desirable Quality Indicators

1. Was data available on attrition rates among intervention samples? Was severe overall attrition documented? If so, is attrition comparable across samples? Is overall attrition less than 30%?
2. Did the study provide not only internal consistency reliability but also test-retest reliability and interrater reliability (when appropriate) for outcome measures? Were data collectors and/or scorers blind to study conditions and equally (un)familiar to examinees across study conditions?
3. Were outcomes for capturing the intervention's effect measured beyond an immediate posttest?
4. Was evidence of the criterion-related validity and construct validity of the measures provided?
5. Did the research team assess not only surface features of fidelity implementation (e.g., number of minutes allocated to the intervention or teacher/interventionist following procedures specified), but also examine quality of implementation?
6. Was any documentation of the nature of instruction or series provided in comparison conditions?
7. Did the research report include actual audio or videotape excerpts that

capture the nature of the intervention?
8. Were results presented in a clear, coherent fashion?

*A study would be acceptable if it included only measures of generalized performance. It would not be acceptable if it only included measures that are tightly aligned.

Source: Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*, 149–164.

Characteristics of Group Design

The basic logic underlying all group research studies 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 and may depend on the research objective or population being studied. When random assignment of participants is not feasible, it is recommended that interventionists be randomly assigned to conditions. Gersten et al., 2005 point out that random assignment of participants does not guarantee study group equivalence, an important consideration when analyzing group research findings. It is the fundamental logic of group design, experimental and quasi-experimental, that groups of participants assigned to each study condition are equivalent on "key" characteristics or status variables (e.g., chronological age, gender, ethnicity, test scores etc.) at the start of a group study (Rosenberg et al., 1992). By starting with equivalent groups across conditions, it is possible to attribute later differences between groups to the independent variable rather than group composition. Because group equivalence is critical, some investigators have chosen to match participants on key characteristics prior to the start of their study and then randomly assign one matched member to each study condition. Implied in this process is the importance of the researcher providing a detailed description of group members, thereby convincing study evaluators that groups were equivalent at the start of the study.

Other participant and interventionist variables should also be addressed when evaluating or reporting results from group studies, including participant attrition and interventionist characteristics. Specifically, it is important to note the number of

participants who have withdrawn from a study and the condition to which they were assigned. If attrition is comparable across conditions there isn't a problem, however, if one condition has a substantially higher attrition rate than another condition, problems arise when analyzing the data, since groups will no longer be comparable. In such cases it is always important to document and report the reasons for participant withdrawal, noting whether it was in some way due to the condition to which they were assigned. For studies in which one or more interventionists are participating, it is important to describe each interventionist in detail so that there are no critical differences between them (e.g., education, certification, experience etc.), as some differences could influence the consistency and fidelity with which the independent variable is implemented. To avoid this potential problem researchers randomly assign or counterbalance interventionists across conditions. When neither option is possible for logistical reasons (e.g., clinical group or teacher classroom assignment), the degree to which condition procedures were followed as specified in the research proposal (procedural fidelity, see Chapter 6) is critical.

The group research approach is the most common research methodology used in some areas of behavioral science. 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 with a specific population, on average, will be effective. With such research questions a group design methodology is recommended. Numerous group 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. Don't be surprised if results are not replicated if your participant or client differs substantially from the average group study participant.

Qualitative Research Approaches

The term *qualitative research* is an "umbrella" term that refers to a number of descriptive research approaches "that investigate the quality of relationships, activities, situations, or materials" (Fraenkel & Wallen, 2006, p. 430). Brantlinger, Jimenez, Klingner, Pugach, and Richardson (2005) define qualitative research as "a systematic approach to understanding qualities, or the essential nature, of a phenomenon within a particular context" (p. 195). A quantitative analysis of outcome measures is typically not of interest to qualitative researchers. The qualitative paradigm is discussed here in spite of its descriptive rather than experimental analysis of behavior due to what appears to be an increase in interest among some researchers who believe it is "ideal for phenomena

that are patently complex and about which little is known with certainty" (Lancy, 1993, p. 9). Table 1.2 identifies and briefly describes 16 different qualitative research approaches that Brantlinger et al. place under the qualitative research paradigm. Of the 16 approaches, 3 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 (individuals, events, activities, or processes), while *ethnography* refers to the study of culture, defined as "the customary beliefs, social forms, and material traits of a racial, religious, or social group" (Merriam-Webster Online Dictionary, 2008), in which the investigator unobtrusively observes people in their natural setting without an attempt to influence their behavior or the event. Sometimes confused with ethnography, *phenomenology* is the study of people's reactions and perceptions of a particular event or situation. For a more in-depth discussion of these and other qualitative research approaches see Glasser and Strauss (1967), Lincoln and Guba (1985), or Lancy (1993).

Characteristics of Qualitative Research

Qualitative research approaches share a number of common characteristics not the least of which is a desire to provide a detailed, in depth description of the case or phenomena under study. Data are collected using several methods, including direct observation in which the investigator's role is that of a "participant-observer" in the natural environment, with neither an interest nor attempt to influence the person or event being observed. As a participant-observer the researcher takes field notes, sometimes referred to as "reflective notes", that are intended to capture the "essence" or "themes" of the observations. Other data collection techniques include audio and video recordings that are summarized and presented in written narratives. Interviews and questionnaires are important data collection instruments used in qualitative research. In terms of these two data collection tools and their use in phenomenology, Fraenkel and Wallen (2006) describe the role of the researcher as one who "extracts what he or she considers to be relevant statements from each participant's description of the phenomenon and then clusters these statements into themes. He or she then integrates these themes into a narrative description of the phenomenon" (p. 437). Unlike the group study approach in which hypotheses are formulated prior to conducting a study to test a theory, known as a *deductive analysis* approach (i.e., general to specific), researchers who use a qualitative study approach collect data and describe themes or trends in the data without offering a theory, an approach known as *inductive analysis* (i.e., specific to general). In this regard, studies using qualitative and SCDs are similar. A critical difference between qualitative and quantitative research approaches is, as Brantlinger et al. (2005) states, "Qualitative research is not done for the purposes of generalization but rather to produce *evidence* based on the exploration of specific contexts and particular individuals" (p. 203). If this is in fact how qualitative researchers view their approach, we as consumers of research must ask the question, "How can qualitative research findings support evidence-based

practice if they can not be generalized beyond the case studied?"

Table 1.2 Types and Descriptions of Qualitative Research.

Case study—exploration of a bounded system (group, individual, setting, event, phenomenon, process); can include autobiography and biography.

Collective case study—a study that takes place in multiple sites or includes personalized stories of several similar (or distinctive) individuals.

Ethnography—description/interpretation of a cultural or social group or system; typically includes observations, interviews, and document analysis.

Action research—researcher brings ideas for practice to fieldwork to have an impact on the setting/participants while collecting data.

Collaborative action research—researcher and practitioner share ideas about how to change practice and work together to modify a situation as well as collect information for a study.

Grounded theory—research done to generate or discover a general theory or abstract analytical hunch based on study of phenomena in a particular situation(s).

Phenomenology—studies the meanings people make of their lived experiences.

Symbolic interactionism—studies interpretive processes used by persons dealing with material and social situations.

Narrative research—collection of personal narratives; based on recognition that people are

Life (oral) *history*—extensive interviews with individuals to collect first person narratives about their lives or events in which they participated.

Quasi-life-history research—encouraging participants to recall and reflect on earlier as well as current meaningful occurrences in their lives.

Interpretive research—used synonymously with "qualitative work" and/or to refer to research framed within certain (critical, feminist, disability study, critical race) theories.

Content analysis—close inspection of text(s) to understand themes or perspectives (also refers to the analysis stage of qualitative studies).

Conversational analysis—studying interactional situations, structure of talk, and communicative exchanges; includes recording facial expressions, gestures, speed or hesitancy of speech, and tone of voice.

Discourse analysis—deconstructs common sense textual meanings; identifies meanings that undergird normative ways of conceptualizing and discussing phenomena.

Ideological critique—discourse analysis that assumes political meanings (power disparities) or ideologies are embedded in, and infused through, all discourses, institutions, and social practices.

storytellers who lead storied lives.

Source: Brantlinger, E., Jimenez, R., Klingner, J., Pugach, M., & Richardson, V. (2005). Qualitative studies in special education. *Exceptional Children*, *71*, 195–207.

Data analysis, Reliability, and Validity Issues

The issue of credibility and trustworthiness of research findings is central to practitioners using promising, if not best practices in their service to students and clients and their families. Guidelines for evaluating the credibility of qualitative research studies have been developed by Brantlinger et al. (2005) and are presented in Table 1.3. These measures are how qualitative researchers address the **validity** (i.e., accuracy) and reliability (i.e., consistency) of information in their research reports, but the authors caution against "using credibility measures as a checklist in a rigid and unreflective way", and although they "encourage" researchers to use credibility measures "they believe that authors who succinctly clarify the methods used and the rationale for them can convey that their reports are reliable and worthy of attention without alluding to credibility measures" (p. 200–201). As you may have deduced from the quotes cited (e.g., "extracts what he or she considers relevant"), the primary criticism of qualitative research approaches are their lack of objectivity.

A common characteristic of qualitative studies is the position of the researcher as an "insider" who has close personal contact with participants and who is both the data collector and data analyst. Brantlinger et al. acknowledge that they (qualitative researchers) are "the instrument" in their research and that, "To do qualitative work well (be *valid* instruments), we must have experience related to our research focus, be well read, knowledgeable, analytical, reflective, and introspective" (p. 197). If true, the position of the qualitative researcher raises concerns because of the subjectivity of the data collected and reported, which in turn influences the validity and reliability of findings since observational safeguards (e.g., independent observations) are rare. This lack of reliability of measurement alone is a major threat to the internal validity of findings, a confounding known as instrumentation. The use of field notes, narrative descriptions, and the freedom of investigators to "consider what is relevant" all signal a method that is prone to subjectivity and findings that would be difficult, if not impossible to replicate. Replication, as previously noted, is at the heart of the scientific method. If replication of a study's findings has not been attempted or not been achieved those findings cannot be considered trustworthy or valid. So, what does qualitative research offer to the science of human behavior? In spite of concerns over subjectivity and lack of replication, qualitative studies can and do provide detailed descriptions of behavior under natural conditions that could subsequently lead to asking research questions, or testing research hypotheses, that employ more objective, quantitative research approaches.

Table 1.3 Credibility Measures for Qualitative Research.

1. *Triangulation*—search for convergence of, or consistency among, evidence from multiple and varied daily sources (observations/interviews; one participant and another; interviews/documents).
    a. *Data triangulation*—use of varied data sources in a study.
    b. *Investigator triangulation*—use of several researchers, evaluators, peer debriefers.
    c. *Theory triangulation*—use of multiple perspectives to interpret a single set of data.
    d. *Methodological triangulation*—use of multiple methods to study a single problem.
2. *Disconfirming evidence*—after establishing preliminary themes/categories, the researcher looks for evidence inconsistent with these themes (outliers); also known as negative or discrepant case analysis.
3. *Researcher reflexivity*—researchers attempt to understand and self-disclose their assumptions, beliefs, values, and biases (i.e., being forthright about position/perspective).
4. *Member checks*—having participants review and confirm the accuracy (or inaccuracy) of interview transcriptions or observational field notes.
    a. *First level*—taking transcriptions to participants prior to analyses and interpretations of results.
    b. *Second level*—taking analyses and interpretations of data to participants (prior to publication) for validation of (or support for) researchers' conclusions.
5. *Collaborative work*—-involving multiple researchers in designing a study or concurring about conclusions to ensure that analyses and interpretations are not idiosyncratic and/or biased; could involve interrater reliability checks on the observations made or the coding of data. (The notion that persons working together will get reliable results is dependent on the "truth claim" assumption that one can get accurate descriptions of situational realities.)
6. *External auditors*—using outsiders (to the research) to examine if, and confirm that, a researcher's inferences are logical and grounded in findings.
7. *Peer debriefing*—having a colleague or someone familiar with phenomena being studied review and provide critical feedback on descriptions, analyses, and interpretations or a study's results.
8. *Audit trail*—keeping track of interviews conducted and/or specific times and dates spent observing as well as who was observed on each occasion; used to document and substantiate that sufficient time was spent in the field to claim dependable and confirmable results.
9. *Prolonged field engagement*—repeated, substantive observations; multiple, in-depth interviews; inspection of a range of relevant documents; thick description validates the study's soundness.
10. *Thick, detailed description*—reporting sufficient quotes and field note descriptions to provide evidence for researchers' interpretations and conclusions.
11. *Particularizabilitv*—documenting cases with thick description so that readers

can determine the degree of transferability to their own situations.

Source: Brantlinger, E., Jimenez, R., Klingner, J., Pugach, M., & Richardson, V. (2005). Qualitative studies in special education. *Exceptional Children*, *71*, 195–207.

Single Case Research Approach

SCD methodology has a long tradition in the behavioral sciences, and has become increasingly common in special education and other fields over time (see Figure 1.1). Historically, studies using SCDs 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) first described the SCD research approach in his seminal book, *Tactics of Scientific Research: Evaluating Experimental Data in Psychology*, which exemplified its application within the context of basic experimental psychology research. In 1968, Baer et al. elaborated on SCD 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 SCD methodology and its use in a number of disciplines, including psychology (Bailey & Burch, 2002; Barlow & Hersen, 1984; Johnson & Pennypacker, 1993, 2009; Kazdin, 1998; Kratochwill & Levin, 1992, Skinner, 2004), special education (Gast, 2005; Kennedy, 2005; Richards, Taylor, Ramasamy, & Richards, 1999; Tawney & Gast, 1984), "helping professions" (Bloom & Fischer, 1982; Lane, Ledford, & Gast, 2017), literacy education (Neuman & McCormick, 1995), communication sciences (McReynolds & Kearns, 1983; Schlosser, 2003), and therapeutic recreation (Dattilo, Gast, Loy, & Malley, 2000).

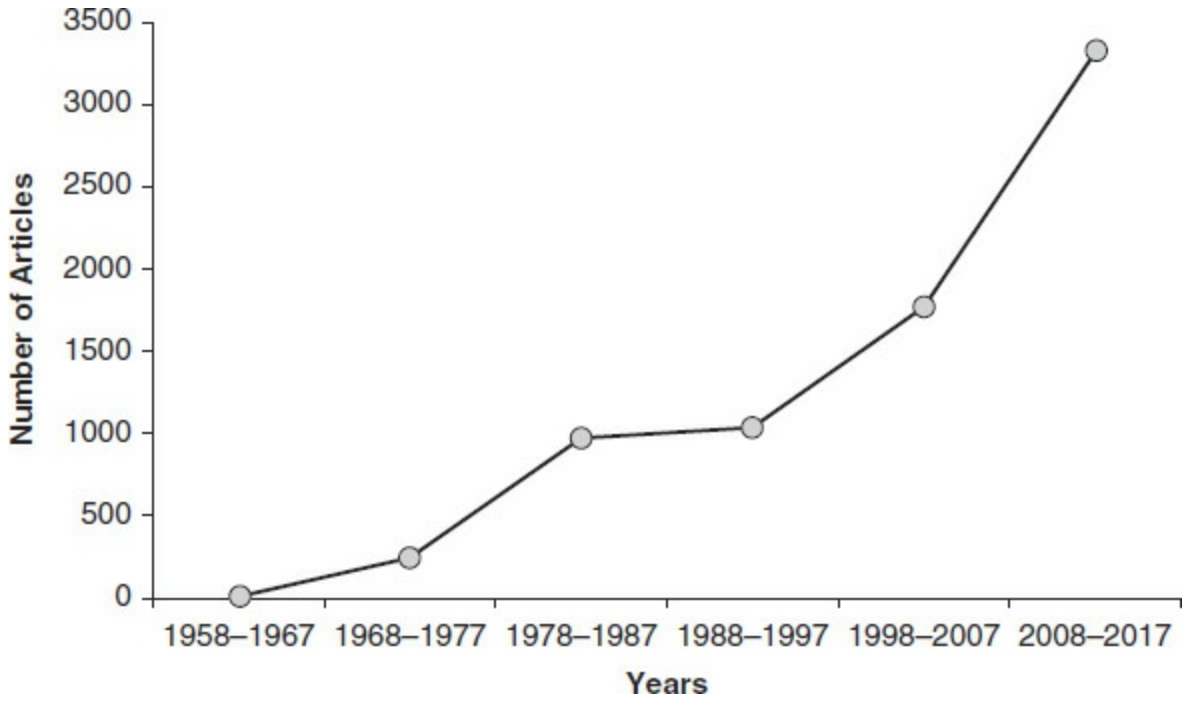


Figure 1.1 The number of citations retrieved by PsycINFO over time, using a string of search terms related to single case design studies ("single subject design" OR "single case design" OR "multiple baseline" OR "multitreatment" OR "withdrawal design" OR "reversal design" OR "multiple probe" OR "alternating treatments design").

As Horner et al. (2005) pointed out, over 45 professional journals publish SCD studies. A common misnomer about SCD research methodology is that it is appropriate only if you ascribe to a behavioral psychology model, which is incorrect. Although it is based in operant conditioning, applied behavior analysis, and social learning theory, interventions based in other theoretical models may be evaluated within the context of an SCD. In this section the basic parameters of SCD research methodology are overviewed as a means of comparison with previously described research approaches. Quality indicators for evaluating studies using SCDs have been developed by Horner et al. and are presented in Table 1.4. The topics introduced in this section, including criteria for evaluating supportive evidence of a practice, are discussed in detail in the chapters that follow.

Table 1.4 Quality Indicators for Single-Case Research.

Description of Participants and Setting

1. Participants are described with sufficient detail to allow others to select individuals with similar characteristics (e.g., age, gender, disability, diagnosis).
2. The process for selecting participants is described with replicable precision.
3. Critical features of the physical setting are described with sufficient precision to allow replication.

Dependent Variable

1. Dependent variables are described with operational precision.
2. Each dependent variable is measured with a procedure that generates a quantifiable index.

3. Measurement of the dependent variable is valid and described with replicable precision.
4. Dependent variables are measured repeatedly over time.
5. Data are collected on the reliability or interobserver agreement associated with each dependent variable, and IOA levels meet minimal standards (e.g., lOA = 80%; Kappa = 60%).

Independent Variable

1. Independent variable is described with replicable precision.
2. Independent variable is systematically manipulated and under the control of the experimenter.
3. Overt measurement of the fidelity of implementation for the independent variable is highly desirable.

Baseline

1. The majority of single-case research studies will include a baseline phase that provides repeated measurement of a dependent variable and establishes a pattern of responding that can be used to predict the pattern of future performance, if introduction or manipulation of the independent variable did not occur.
2. Baseline conditions are described with replicable precision.

Experimental Control/Internal Validity

1. The design provides at least three demonstrations of experimental effect at three different points in time.
2. The design controls for common threats to internal validity (e.g., permits elimination of rival hypotheses).
3. The results document a pattern that demonstrates experimental control.

External Validity

1. Experimental effects are replicated across participants, settings, or materials to establish external validity.

Social Validity

1. The dependent variable is socially important.
2. The magnitude of change in the dependent variable resulting from the intervention is socially important.
3. Implementation of the independent variable is practical and cost effective.
4. Social validity is enhanced by implementation of the independent variable over extended time periods, by typical intervention agents, in typical physical and social contexts.

Source: 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.

Characteristics of Single Case Research Design

In spite of its name, it is important to understand that this research approach is *not* a case study approach in which there is only one participant whose behavior is described, in detail, in written narrative, based on primary data collected using qualitative research

techniques (e.g., field notes, interviews etc.). SCD is a quantitative experimental research approach in which study participants serve as their own control, a principle known as **baseline logic** (Sidman, 1960). In the simplest SCD study, each participant is exposed to both a "control" condition, known as baseline, and an intervention condition. As with group design studies, it is possible to compare two treatments; in this case, each participant is exposed to both intervention conditions. The target behavior is repeatedly measured within the context of one of several research designs that evaluate and control for threats to internal validity. Depending on the research design used, baseline (A) and intervention (B) conditions are slowly alternated across time (e.g., A-B-A-B or withdrawal design; Chapter 9), rapidly alternated (e.g., ATD and AATD; Chapter 11), or the intervention condition is introduced in a time-lagged fashion across several behaviors, conditions, or participants (Chapter 10). Return to a previously introduced condition or introduction of a new condition to a new behavior, condition, or participant occurs only after data stability is evident. Data for individual participants are presented on a line graph for each participant and decisions to maintain or change the current condition are made in accordance with visual analysis guidelines (see Chapter 8 for information on visual analysis). Baseline logic is very 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 SCDs, each participant participates in both conditions of interest (e.g., baseline or control *and* intervention). In group design, posttest data are collected at an a priori specified time point (e.g., after 3 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 SCD, intervention conditions are generally continued until a performance criterion is met or until progress is apparent via visual analysis of graphed data. The use of visual analysis of graphic data for individual participants make SCD 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.

Controlling Threats to Internal Validity

As with experimental group design approaches, experimental SCD research must adequately control for or detect threats to internal validity. In SCD, as in group design, there are multiple procedures for controlling for these threats, including ensuring reliability of measurement and fidelity of procedures. In addition, rather than randomizing participants to reduce the likelihood of threats, SCD researchers use systematic ordering of conditions to do so. Controlling for threats to internal validity for specific SCDs will be discussed in detail in Chapters 9–12.

Applied Research, Practice, and Single Case Design

Evidence-based practices, supported by rigorous and internally valid research, may be preferred by scientists, but another term, *practice-based evidence* (PBE), is also important. PBE can be identified through research that occurs in applied settings, with typical resources; SCD may be particularly well-suited to conducting this type of research (Smith, Schmidt, Edelen-Smith, & Cook, 2013). Although some might argue that the basic purposes of research and practice are not aligned, we would like to draw some parallels between the behaviors we consider to be fundamental to both science and educational/clinical practice, while acknowledging that some differences exist.

Similarities Between Research and Practice

Practitioners must:

1. Analyze an individual's performance to identify the initial performance level (a form of hypothesis testing).
2. Specify instructional/therapy objectives including criterion performance levels.
3. Operationally define instructional/therapy procedures so that another informed adult is able to implement procedures with fidelity.
4. Conduct concept and/or task analyses as a means of sequencing intervention programs for individual learners.
5. Implement procedures consistently.
6. Collect repeated measures on each individual's performance.
7. Analyze data and make program decisions based on the data.
8. Maintain data records.
9. Share an individual's performance regularly with significant others.
10. Follow professional/ethical guidelines.

Applied researchers must:

1. Identify a behavior challenge.
2. Generate a research question ("If I do this, will the behavior improve?")
3. State specific research program objectives.
4. Define the elements of the research procedure: stimuli, arrangement, materials and equipment, target response topography, consequent events.
5. Write specific, replicable research procedures and implement with fidelity.
6. Collect direct, repeated, and reliable measures of performance.
7. Analyze graphically displayed data and make research decisions based on data.
8. Maintain data records.

9. Share research progress with research team members and significant others.
10. Conduct research in an ethical manner.

The similarities in these sets of behaviors are apparent. They can be synthesized by noting that both the teacher/therapist and applied researcher must (a) be able to identify and analyze problems, (b) generate creative solutions, (c) implement an intervention in a systematic manner, (d) document the effect of the intervention, and (e) act on the data in an ethical and responsible way. Barlow et al. (1984) referred to those teachers and therapists who engage in applied research as "scientist-practitioners", a reference we believe aptly describes those who conduct applied research as part of their daily service delivery activities.

Some Differences Between Research and Practice

Schools and community-based programs and clinics seldom have the same level of resources as those used in typical research studies. The fact that teacher/therapist-researchers who work in community settings often must utilize existing resources can add to the generality of their research findings. In recent years there has been concern that some "applied" research being disseminated may not be so "applied" after all, in that it requires special resources that are out of reach of most teachers and therapists working in typical community service and educational settings. Many organizations have attempted to respond to this problem by disseminating practitioner-friendly journals in addition to typical journals including peer-reviewed research studies (e.g., ABAI, *Behavior Analysis in Practice;* Council for Exceptional Children [CEC], *Teaching Exceptional Children*; Division for Early Childhood of the CEC, *Young Exceptional Children*).

The point we want to make here is that the typical classroom is obviously not a Skinner box; instead it is a complex social environment that includes an almost immeasurable number of potential extraneous variables. In special education, speech/language therapy, and child psychology, however, the trend is for most instruction to occur within the context of natural activities and routines implemented across the day. If this is the context in which you plan to conduct your research, it is important for you to know that you may need to create detailed plans for data collection and environmental control. This may not be an easy task, but the more familiar you are with measurement and design alternatives the easier it will be. What follows are suggestions on how to proceed and questions you should ask if you are: (a) planning to conduct your own research project in your own classroom or clinic setting as both the primary researcher and service provider; (b) a collaborating teacher or therapist opening your work environment to someone else who will serve as the primary researcher; or (c) a visiting researcher who needs to be sensitive to the demands placed on the collaborating teacher or therapist. These questions are framed from the teacher/therapist perspective (i.e., the person who has primary responsibility for ensuring that teaching or

therapy is not disrupted by the research process; Eiserman & Behl, 1992).

1. Does the research question address an educational or therapy objective? Will participants benefit from their participation?
2. Is there a research base that leads you to believe that your participation is likely to improve practice?
3. Are the research objectives and procedures consistent with current agency policies?
4. Do you have an interest in the answer to the research question?
5. How will participation affect your daily schedule and the schedule of participants? Will the current daily schedule have to be altered?
6. How does the intervention under study affect continuation of interventions currently in use? Are you willing to modify or abandon current interventions and replace with the new intervention for a period of time?
7. Will participation disrupt other activities or events typically attended by participants?
8. How much of your time, and that of each student, will be required each day? How many days, weeks, or months are you willing to commit to this project? Is this commitment reasonable and justifiable?
9. How will participants, in your judgment and experience, respond to their participation?
10. Will significant others (parents, guardians, agency administrators etc.) support the research objective and participation?
11. Are the necessary resources available (e.g., data collectors, reliability observers, computers, software programs, cameras, assistive or adaptive equipment) for conducting the research? If a piece of equipment breaks downs is there back-up equipment available?
12. Do you have any ethical concerns?

Answers to these questions, which only sample the range of questions you must ask, are important prior to committing yourself and others to a research project. In that SCD studies typically occur over several weeks, if not months, you must understand the practical implications of your commitment from the outset. We encourage you to enter any research project with a thorough understanding of its research base, potential contributions, logistical challenges, procedural requirements, and ethical implications. All studies are not equal in their research requirements (data collection procedures, intervention procedures, research designs etc.), and the more you understand measurement and design alternatives, the more likely you are to design a study that will be practical for your setting while advancing both science and practice.

Threats to Internal Validity

The internal validity of a study depends on attempts by the researcher to ensure that plausible reasons for behavior change, other than planned experimental changes, are controlled for. Two concepts are important for understanding the pragmatics of experimental control and internal validity. First, it is impossible to control for every possible threat to internal validity. Second, a possible threat may not be an actual threat. Each possible threat 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 direct replications, will determine the level of confidence you should 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 course of 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 SCD; many are also applicable for other experimental studies (e.g., group comparison studies).

History

History refers to events that occur *during* an experiment, but that are not related to planned procedural changes, that may influence the outcome. Generally speaking, the longer the study the greater the threat due to history. Potential sources of history threats, when a study is conducted in community settings, are the actions of others (parents, siblings, peers, childcare providers) or by study participant themselves (independent online research, observational learning, serendipitous exposure through the 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 in an attempt 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 believable. 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; careful research notes may assist in explaining this variability due to transient history effects.

Maturation

Maturation refers to changes in behavior due to the passage of time. In a "short" duration study (4–6 weeks) 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, however, the study is carried out over several months (4–6 months) with the same young child or if a weak intervention is used, there is a greater likelihood that maturation may play a role in observed behavioral changes. Some researchers have referred to "session fatigue" as a maturation threat to validity. Session fatigue refers to a participant's performance decreasing over the course of a session (e.g., 80% accuracy over the first 20 trials and 20% accuracy over the last 20 trials of a 40-trial session). We may debate whether session fatigue is a maturation threat but we would certainly agree it is a threat to the validity of the findings. To avoid session fatigue it is important to be sensitive to a participant's age and attention span, scheduling shorter sessions with fewer trials for younger children and individuals who have a history of inattentive behavior. It may also be helpful in restoring attention to task and responding to take a short break (3–5 minutes) midway during a lengthy session.

Testing

Testing is a threat in any study that requires participants to respond to the same test repeatedly, especially during a baseline or probe condition; it is the likelihood that the repeated assessment task will result in participant behavior change. 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 effect. Test sessions of long duration, requiring substantial participant effort, with minimal or no reinforcement for attention and active participation may result in an inhibitive effect. It is important to design your baseline and probe 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 randomizing stimulus presentation order across sessions; not reinforcing correct responses, particularly on receptive tasks;

not correcting incorrect responses; and not prompting (intentionally or unintentionally) correct responses. Procedural reliability checks will help with detecting these 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; they are of particular concern in SCD studies because of repeated measurement by human observers who may make errors. In studies using SCD, 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. You can avoid common problems by carefully defining behaviors of interest, using appropriate recording procedures, and frequently checking for reliability by using a secondary observer. Historically, percentage agreement at or above 90% is preferred in applied research, while percentage agreement below 80% is considered unacceptable. Unfortunately, determining what percentage IOA is acceptable, or unacceptable, is not as easy as it may seem since some behaviors are easier to record (permanent products, behaviors of long duration, gross motor responses) than others (high rate behaviors, behaviors of short duration, vocal responses). In addition, the conditions under which data are collected will influence what percentage agreement you find acceptable. Assuming behavioral definitions are clearly written and observers are properly trained, you would expect measurement errors to be lower when data are collected from permanent products (audio or video recordings, written assignments, assemblies, computer printouts), compared to live observations in "real time". Issues related to reliability of measurement are discussed in Chapter 5. Suffice it to say here you must attend to the details of your measurement system to avoid instrumentation threats to internal validity.

Procedural Infidelity

Procedural infidelity refers to the lack of adherence to condition protocols by study implementers. If the procedures of an experimental condition (baseline, probe, intervention, maintenance, generalization) are *not* consistently implemented across behavior episodes, time, interventionists etc., as described in the Methods section of the research proposal or report, confidence that outcomes are related to the intervention is considerably reduced. Control for procedural infidelity threats to internal validity is discussed in Chapter 6.

Selection Bias

Selection bias involves choosing participants 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, Cook, & Campbell, 2002). In SCD, the "population" would be individuals who meet the inclusion criteria for the study and have similar functional characteristics to the participants (Lane, Wolery, Reichow, & Rogers, 2007; Wolery, Dunlap, & Ledford, 2011). **Attrition** refers to 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). A *minimum* of three participants is typically recommended for inclusion in any one SCD investigation. However, since it is unlikely that you will have much control over participants who choose to withdraw from your study, or who are required to withdraw due to the family moving, incarceration, hospital admission, or school expulsion, it is recommended that you start with four or more participants when available and if practical. With four participants the loss of one participant will have less of an impact on your analysis of independent variable generality. **Attrition bias** refers to the likelihood that participant loss (attrition) impacts the outcome of the study. When attrition occurs, you should always (a) explicitly report it, along with relevant information about why it occurred, and (b) include any data collected for that participant in your research report. This ensures that data from "nonresponders" are not systematically excluded from published research, resulting in bias regarding evidence of intervention effectiveness.

Another type of selection bias, sampling bias, occurs in group designs when non-random samples of the population are recruited (i.e., some members of a population are more likely to be included than others). **Sampling bias** occurs in SCD studies when researchers use additional, non-explicated, reasons for including or excluding potential participants. For example, Ledford, Chazin, Harbin, and Ward (2017) included 12 children in a study to assess preference for massed versus embedded instruction, and named the following inclusion criteria: (a) ability to play age- or developmentally-appropriate games with turn-taking, (b) ability to make choices given line drawings, and (c) 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 being noncompliant during teacher-led activities (e.g., massed instruction) to reduce the risk of attrition. This decision leads to the potential for overestimating differences between conditions because of the purposeful exclusion of participants unlikely to perform well in one of the two conditions. This risk could be mitigated by randomly choosing participants when the pool of participants who meet inclusion criteria is larger than the total number who can participate. 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 cognition or language skills because students with more impaired skills received more therapy and were thus available less frequently).

Multiple-Treatment Interference

Multiple-treatment interference can occur when a study participant's behavior is influenced by more than one planned "treatments" or interventions during the course of a study. An interactive effect may be identified due to *sequential confounding* (the order in which experimental conditions are introduced to participants may influence their behavior) or a *carryover effect* (the effect when a procedure used in one experimental condition influences behavior in an adjacent condition). To avoid sequential confounding, the order in which experimental conditions are introduced to participants is counterbalanced (e.g., participant 1, A-B-C-B-C; participant 2, A-C-B-C-B). Carryover effects are detected via visual analysis; they can be minimized by continuing the condition until data are stable (see Chapters 9–11).

Data Instability

Instability refers to the amount of variability in the data (dependent variable) over time. As Kratochwill (1978, p. 15) noted, "Experiments involving repeated measurement of a single participant or group over time typically evidence some degree of variability. If this 'instability' is large, investigators could attribute an effect to the intervention when, in fact, the effectiveness was no larger than the natural variation in the data series." Your attention to the amount of variability in a data series is important in deciding if and when it is appropriate to move to the next experimental condition. As will be discussed in Chapter 8, during a visual analysis of graphic data, both level and trend stability must be considered before changing conditions if there is to be a clear demonstration of experimental control. The premature introduction of the independent variable into a data series may preclude such a demonstration. As a consumer of research, you should determine if there is high percentage overlap between data points of two adjacent conditions, and, if there is, you should be skeptical of any statements a researcher might make regarding the effectiveness of the independent variable. In your own research, when data variability is observed, it is best to a) maintain the condition until the data stabilize, or b) attempt to isolate the source of the variability. Threats to internal validity due to data instability are preventable if you are patient and analytical in your research decisions, rather than following some predetermined schedule that dictates when to move to the next experimental condition (e.g., every 7 days the experimental conditions will change).

Cyclical Variability

Cyclical variability is a specific type of data instability that refers to a repeated and predictable pattern in the data series over time. When experimental conditions are of equal length (e.g., 5 days in each condition of an A_1-B_1-A_2-B_2 withdrawal design) it is possible that your observations coincide with some unidentified natural source that may account for the variability. For example, if your experimental condition schedule coincides with a parent's work schedule (away from home for 5 days, at home for 5 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 condition lengths across time.

Regression to the Mean

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 back 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 the cause. Instead, continue collecting data until stability is established.

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. It is recommended that study participants be exposed to unfamiliar adults, settings, formats, data collection procedures (e.g., video recording) etc. prior to the start of a study, through what is sometimes referred to as *history training*, to increase the likelihood that data collected on the first day of a baseline condition is representative of participants' "true" behavior. A "*reactive effect*" to being observed has been reported and discussed in the applied research literature for quite some time (Kazdin, 1979), leading to recommendations to be as unobtrusive as possible during data collection (Cooper, Heron, & Heward, 2007; Kazdin, 2001).

Hawthorne Effect

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. Self-management studies, in which participants record their own behavior, are particularly susceptible to a Hawthorne Effect. As Cooper et al. (2007) state, "When the person observing and recording the target behavior is the participant of the behavior change program, maximum obtrusiveness exists, and reactivity is very likely"(p. 591). Like adaptation, familiarizing participants with experimental conditions, specifically data recording conditions, prior to the start of a study may decrease the likelihood of a Hawthorne Effect.

Summary

There are a number of research approaches available to the scientist-practitioner who chooses to add evidence in support of a particular practice he or she is currently using or is considering for use. As a contributor to research evidence, it is important to choose the appropriate research methodology that best answers the research question. Group research methodology is appropriate and best suited for testing hypotheses when your interest is in the average performance of a group of individuals, but it will have limited generality to individuals who differ from those for whom the intervention was effective. Unfortunately for practitioners who are consumers and evaluators of group design research, sufficient details are seldom provided on individual participants that would allow them to make an informed decision as to the likelihood of their student or client responding positively to the intervention studied. Qualitative research approaches (e.g., case study, ethnography, phenomenology etc.) may be appropriate if your interest is in an in depth descriptive report of an individual, activity or event. Studies using this research approach make no attempt to intervene, control for common threats to internal validity, or generalize findings beyond the case studied. The SCD research approach focuses on individual performance and permits practitioners and researchers to independently evaluate the merits of a study or a series of studies since all primary data are presented on all participants in graphic displays and tables. In accordance with scientific method principles, sufficient detail is typically presented in SCD research reports to permit replication by independent researchers. It is through such replication efforts that the generality of findings of a single study is established and evidence generated in support of an intervention. In the chapters that follow we have attempted to provide sufficient detail on the parameters of SCD research methodology to allow you to objectively evaluate and conduct studies using SCDs. Through your efforts and the efforts of other applied researchers it is possible to advance our understanding of human behavior and add evidence in support of effective practices. To this end, scientist-practitioners must disseminate their research findings in professional journals, at professional conferences, and during clinic or school in-services.

References

Adamo, E. K., Wu, J., Wolery, M., Hemmeter, M. L., Ledford, J. R., & Barton, E. E. (2015). Using video modeling, prompting, and behavior-specific praise to increase moderate-to-vigorous physical activity for young children with Down syndrome. *Journal of Early Intervention*, *37*, 270–285.

Association for Behavior Analysis International. (1989). *The right to an effective behavioral treatment.* Retrieved January 22, 2017, from www.abainternational.org/about-us/policies-and-positions/right-to-effective-behavioral-treatment,-1989.asp

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.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1987). Some still-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.

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

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

Bloom, M., & Fischer, J. (1982). *Evaluating practice: Guidelines for the accountable professional.* Englewood Cliffs, NJ: Prentice-Hall.

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

Brantlinger, E., Jimenez, R., Klingner, J., Pugach, M., & Richardson, V. (2005). Qualitative studies in special education. *Exceptional Children*, *71*, 195–207.

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

Cook, B. G., & Odom, S. L. (2013). Evidence-based practices and implementation science in special education. *Exceptional Children*, *79*, 135–144.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

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,

Inc., Publishers.
Deno, S. L. (1990). Individual differences and individual difference: The essential difference of special education. *The Journal of Special Education*, *24*(2), 160–173.
Eiserman, W. D., & Behl, D. (1992). Research participation: Benefits and considerations for the special educator. *Teaching Exceptional Children*, *24*, 12–15.
Forman, S. G., Shapiro, E. S., Codding, R. S., Gonzales, J. E., Reddy, L. A., Rosenfield, S. A.,. . Stoiber, K. C. (2013). Implementation science and school psychology. *School Psychology Quarterly*, *28*, 77.
Fraenkel, J. R., & Wallen, N. E. (2006). *How to design and evaluate research in education* (6th ed.). New York, NY: McGraw-Hill.
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.
Gersten, R., Fuchs, L. S., Compton, D., Coyne, M., Greenwood, C., & Innocenti, M. (2005). Quality indicators for group experimental and quasi-experimental research in special education. *Exceptional Children*, *71*, 149–164.
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.
Institute of Education Sciences. (no date). Retrieved March 17, 2017, from https://ies.ed.gov/aboutus/
Johnson, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). New York, NY: Psychology Press.
Johnson, J. M., & Pennypacker, H. S. (2009). *Strategies and tactics of behavioral research* (3rd ed.). New York, NY: Routledge.
Kazdin, A. E. (1979). Unobtrusive measurement in behavioral assessment. *Journal of Applied Behavior Analysis*, *12*, 713–724.
Kazdin, A. E. (1998). *Methodological issues and strategies in clinical research*. Washington, DC: American Psychological Association.
Kazdin, A. E. (2001). *Behavior modification in applied settings* (6th ed.). Belmont, CA: Wadsworth.
Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Pearson/Allyn and Bacon.
Kida, E., Rabe, A., Walus, M., Albertini, G., & Golabek, A. A. (2013). Long-term running alleviates some behavioral and molecular abnormalities in Down syndrome mouse model Ts65Dn. *Experimental Neurology*, *240*, 178–189.
Kratochwill, T. R. (Ed.) (1978). *Single subject research—Strategies for evaluating change*. New York, NY: Academic Press.
Kratochwill, T. R., & Levin, J. R. (1992). *Single-case research design and analysis: New direction for psychology and education*. Hillsdale, NJ: Lawrence Erlbaum.

Lancy, D. F. (1993). *Qualitative research in education*. New York, NY: Longman.

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.

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.

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.

Merriam-Webster Online Dictionary (2008). Ethnography. Retrieved May 1, 2008, from www.merriam-webster.com/dictionary/ethnography.

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.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harms, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children*, *71*, 137–148.

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.

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.

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.

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

Schlosser, R. W. (2003). *The efficacy of augmentative and alternative communication*. Boston, MA: Academic Press.

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

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

Skinner, B. F. (1953). *Science and human behavior*. New York, NY: Palgrave Macmillan.

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

Smith, G. J., Schmidt, M. M., Edelen-Smith, P. J., & Cook, B. G. (2013). Pasteur's quadrant as the bridge linking rigor with relevance. *Exceptional Children*, *79*, 147–161.

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

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.

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

Van Houton, R., Axelrod, S., Bailey, J. S., Favell, J. E., Foxx, R. M., Iwata, B. A., & Lovaas, O. I. (1989). *Statement on the right to effective behavioral treatment*. Association for Behavior Analysis Task Force on the Right to Effective Behavioral Treatment. Retrieved from www.abainternational.org/about-us/policies-and-positions/right-to-effective-behavioral-treatment-1989.aspx

What Works Clearinghouse. (2016). *WWC intervention report: Functional behavioral assessment-based interventions*. Washington, DC: U.S. Department of Education, Institute of Education Sciences. Retrieved March 1, 2017, from https://ies.ed.gov/ncee/wwc/Docs/InterventionReports/wwc_fba_011017.pdf

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

2

Ethical Principles and Practices in Research

Linda Mechling, David L. Gast, and Justin D. Lane

Important Terms

institutional review board, respect for persons, beneficence, justice, undue influence, consent, assent, minimal risk, confidentiality, anonymity

You may be surprised to learn that a research project, independently conceived and carefully designed, must undergo formal scrutiny before it is carried out with human participants. Later, in the midst of the institutional review process, it may come as an even greater surprise to hear members of a human subjects review committee raise serious questions about potential harmful effects as they consider what you perceive to be a most benign intervention program. Or, review team members may question whether the benefits of the proposed study outweigh the risks, as *they* perceive them. It may seem that some interventions are primarily educational or therapeutic and thus need not be presented for human subjects to review. Yet, any intervention that presumes to alter the social or academic behavior of research participants, and that presumes to have scientific merit (i.e., to contribute to a knowledge base) raises fundamental ethical and specific procedural questions. Under present federal regulations (The Public Health Service Act as amended by the National Institutes of Health Revitalization Act of 1993,

P.L. 103–143), sponsoring institutions must ensure the rights of research participants are protected. These assurances are made only after the proposed study has been brought under public scrutiny through examination by a human subjects review committee and the investigator has undergone completion of a training program for conducting research with human participants. Note that we will sometimes refer to participants as subjects in this chapter, consistent with the terminology associated with ensuring ethical treatment of "human subjects".

History of Ethics in Applied Research

The goal of science is the advancement of knowledge. Well-designed applied research studies allow for (a) systematic study of behaviors in the typical environments (baseline measures), (b) evaluation of a new intervention or innovation, and (c) replication of findings from other studies under similar and novel conditions (Sidman, 1960). All forms of scientific inquiry are presumed to be important, whether they seem to offer benefits that are immediate and practical or long range and esoteric. Thus, the scientist pursues knowledge, along whatever path that may take her; at least that is the common view. Historically, some have misrepresented the pursuits of science by violating the basic human rights to which all individuals are entitled, committing crimes against humanity and attempting to hide unspeakable atrocities under the veil of science (National Institutes of Health [NIH], 2008). Examples include such instances as the Nuremberg War Crime Trials (Nazi medical war crimes) and the Tuskegee Study (untreated syphilis in African-American males; Breault, 2006). Such atrocities led to the development of a number of regulations, all designed to ensure the highest levels of protection for human participants in research studies (e.g., Surgeon General, 1966), including the development of committees responsible for reviewing proposed research studies, known today as the **Institutional Review Board** (IRB).

A key historical moment in applied research is the passing of the National Research Act (Pub. L. 93–348), which led to the development of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research in 1974. In 1978 the commission put forth the Belmont Report, which provided guidelines, and corresponding action points, for those conducting applied research. The Belmont Report focused on three overarching principles to improve protection of human participants in applied research studies: (1) respect for persons, (2) beneficence, and (3) justice. The principle of **respect for persons** highlighted the importance of voluntary involvement in research and explaining the purpose of a study and corresponding procedures (informed consent), as well as protection of vulnerable populations (e.g., children with intellectual disability). The principle of **beneficence** focused on the rules of "do no harm" and "maximize possible benefits and minimize possible harms" (ratio of cost to benefits; p. 23194). Finally, the principle of **justice** highlighted the importance of fairness, especially as it relates to recruitment of participants and treatment of those from vulnerable or underrepresented populations. To further assure the protection of human participants, the Federal Policy for the Protection of Human Subjects, also known as the Common Rule, was published in 1991, which further specified the application of the principles of the Belmont Report in applied research. These examples highlight some of the work, and continued efforts, to ensure protection of human participants in research studies.

The following sections present ethical issues that must be considered in applied research, describes the steps the you must go through to obtain institutional approval to

conduct thesis, dissertation, or independent research, and, then, describes the ethical guidelines researchers must follow when using SCDs to answer research questions related to the effectiveness of their instructional and treatment programs.

Conducting Research in Applied Settings

Prior to official approvals discussed later in the chapter, you must obtain initial permission to conduct research at a particular site. Many school systems have specific and delineated procedures for requesting this permission; some smaller agencies may simply require permission from a director or board of directors. Some agencies have separate processes for practitioners (employees) and outside researchers. It is prudent to attempt to learn as much as possible about the agency and their procedures prior to requesting permission to conduct research. Researchers should take care to maintain good relations with all stakeholders (e.g., participants, parents, practitioners, administrators). After the study is completed, parents, practitioners, and participants should be debriefed. Copies of the final report may be requested by the school or agency as a condition for permission to conduct the research; carefully document how much information you will provide to stakeholders prior to beginning the project. At every stage, the behavior of the researcher is critical to successful completion of the project. Further, personal interactions between the researcher and others may determine, in large part, whether other researchers are subsequently permitted to work in the system. We assume that basic courtesy is a firmly established part of the researcher's repertoire and thus we will refrain from sermonizing on how to behave in schools, clinics, and community agencies.

Recruiting Support and Participation

Each institution has a set procedure to follow to request permission to work in schools and clinics. In some cases, in a university, a request may be transmitted to a coordinator of field experiences in their college (College of Education, College of Arts and Sciences, College of Public Health) and then transmitted to a specific individual in the proposed school district or agency. In smaller school systems or agencies, that person may forward the request to the school board or board of directors for approval, then, if approved, to the principal or on-site administrator. In a clinical setting permission may be required from the clinic director or even the chair of an academic department that houses the clinic. Undoubtedly, the process varies from place to place but in each system there is a clearly defined channel. In an active College of Education, for example, there are many channels of communication. Many practitioners are program graduates, maintain close working relations with former professors, and offer an open invitation to work in their classrooms. On some occasions, practitioners will request help directly from the university for students who have behavioral challenges. In other situations, when faculty have close working relationship with principals and faculty, it will seem logical to informally discuss a project which then becomes tacit approval to carry out the research.

In other situations, you may be encouraged to proceed along formal and informal channels at the same time. Little advice can be offered that applies across settings and situations except that (a) a request to conduct research should be submitted through formal channels as soon as human subjects approval is granted; (b) when invited into the schools informally, specifically ask the person issuing the invitation who should be apprised of the visit; and (c) when discussing a potential study during informal contacts (e.g., when you are in the school on community business) stress the fact that the project is only a possibility, describe what stage of development the study has reached, state when it is likely to be formally submitted, and clearly indicate that potential interest, rather than formal approval, is being sought.

Emphasis on real-life, community-based experiences require that research no longer be conducted within the confines of clinics and classrooms. Projects may require collaboration with public or private community agencies and companies. Students should be aware that permission to conduct research in community settings, such as a local grocery store chain, might take weeks and even months to obtain. Local managers will often be required to obtain approval from district-level managers and in some instances from corporate headquarters. The level of approval will depend on the type of research involved and the format proposed (i.e., video recording, involvement of the company's employees, interaction with the business' customers). Moreover, the goals and values of a particular organization (i.e., school, private service provider) may not align with your research interests (e.g., focusing on grade-level standards for individuals with significant disabilities; cf. Ayres, Lowery, Douglas, & Sievers, 2011, 2012; Courtade, Spooner, Browder, & Jimenez, 2012).

Common Courtesies

Researchers should be sensitive to the disruption a research project may cause; however, they may be less familiar with the concerns that will be presented by community employers and managers. For example, questions may be raised concerning: (a) time of day proposed to be on-site for conducting research (e.g., peak business hours and maximum demands on employees); (b) days of the week the proposed research will be conducted (e.g., the day during the week when several extra adults are already present to volunteer or provide therapy); and (c) liability risk to the company. If the research project holds promise of making life more comfortable for the school faculty member, clinician, or business manager (e.g., an educational intervention provides tutorial instruction for a child and thus frees the teacher for other work, or includes job training for future company employees) any reasonable disruption is likely to be tolerated. If the benefit is most direct for the researcher and holds only potential benefits for the achievement of science, disruption is less likely to be tolerated. Researchers may consider how to contribute time, technical assistance, or other needed resources to research sites. For example, in our previous work, we have (a) provided consultation outside of research participation for non-participant individual or untargeted behaviors, (b) provided

materials used in research following study completion (e.g., left toys for classroom use), and (c) provided classroom assistance before and after study sessions (e.g., setting up for an upcoming activity).

Divulging information during the study may create a dilemma. Practitioners, parents and participants may want to know how things are going. Full, open, and honest disclosure of information is fundamental to ethical practice and the protection of participants' rights. However, specific feedback given frequently may constitute another independent variable in some studies. Suppose that a novel intervention is employed to shape a desired social behavior, that parents request and receive daily progress (Ralph said "please" and "thank you" three more times than yesterday), and that the parents naturally increase the opportunities and reinforcement at home. The change in behavior may occur for two reasons, thus confounding the study. One solution to the problem is to decide ahead of time what type of honest but neutral response will be given ("things seem to be going as expected"). If the project goes badly and parents request that you terminate the project, remember that one element of informed consent is the participant's or parents' right to withdraw participation at any time without repercussions.

Recognition and Reinforcement of Participation

Researchers should realize that they have an obligation to others who may follow them. One way to meet this obligation and leave the research site on a positive note is to spontaneously provide information about the outcome of the study and to recognize those who participated. Having plotted data daily, graphs will be available for discussion as soon as the research project is finished. The written narrative from the human subjects review protocol or the written thesis proposal provides a frame of reference to discuss what was done, and the graphs provide a referent for the outcome. Immediate feedback to parents, practitioners, and others is likely to be positively reinforcing. When the researcher has made commitments to provide written reports, it would seem advisable to provide them before they are due. Hand delivered, with an additional word of thanks for the research opportunity, they should leave a favorable impression and increase the probability that the next researcher will be well received. You should always tell participants (and/or their parents or guardians) with whom you will share data, during the informed consenting process (see below). For example, if you intend to assess an intervention for improving a teacher's positive interactions with children, and you plan to share these data with their principal, this could be viewed as potentially negative (e.g., a principal might view behavior changes as insufficient and determine that the teacher should not receive a renewed contract). Thus, always explicitly define with whom data will be shared.

You should further consider recognition to participating practitioners, or public recognition to the community agency or business. Within a published manuscript some authors will recognize the school or business with an "acknowledgement statement" as a

footnote to the article. Participating practitioners and agencies can also be nominated for various awards and certificates (i.e., "Making a Difference") within the community or through state and national agencies (Autism Society of America). Public acknowledgements are generally well-received; however, always take care to ensure that participants are only identified as such with their informed consent.

Finally, some researchers choose to recognize participation by including monetary support for participating professionals. This may be accomplished through an "honorarium" line in the budget of a research grant or by providing a "small" gift at the completion of the study (i.e., a gift card from a local grocery for participation in a study to teach purchasing skills; end-of-the-year pizza party for the class; or contribution to a school fundraising effort). You should always take care that these awards are not large enough that it could be considered coercive, or to have **undue influence**. This is generally defined as convincing participants to enroll in a study when they would not otherwise do so (Williams & Walter, 2015).

Securing Institutional and Agency Approval

The process outlined here begins after a researcher has selected a research topic and (if a student) obtained approval from an academic advisor or thesis research sponsor to proceed with the project. The process of institutional review varies among agencies and universities. Two typical procedures are outlined; one requires researchers to defend a written proposal before an Institutional Review Board (IRB; full review), the other does not (expedited review). The process starts with a very mundane act: locating the proper forms. The forms will be accompanied, in most cases, by guidelines for preparation of the narrative portion of the protocol and deserve intensive study by the researcher (see Figure 2.1 for a sample form). Preparation of a clearly written proposal, submitted early, should increase the probability that approval is obtained in sufficient time to conduct the research. A second requisite of early attention is completion of "training for human subject researchers" which is required prior to IRB reviewing the research application and fulfills the NIHs human subjects training requirement (www.citiprogram.org). This mandatory training is generally provided online and available through university research foundations, sponsored programs or Offices of the Vice President for Research, as well as agencies like the National Cancer Institute. Free, web-based tutorials provide information about the rights and welfare of human participants in research, and are based on the all-important Belmont Report (National Commission for the Protection of Human Subjects, 1978).

FORM A

F2.0050

GENERAL INFORMATION SHEET: NONMEDICAL IRB

IRB # _____

THIS FORM MUST BE TYPED

Note: For best results in opening links contained within this document, it is recommended that you first sace this document to the location of your choice. Open the document from that location, then right-mouse click on a link and select "open hyperlink".

This application is described by (check one):

☐ A. New IRB Research Protocol (Not previously reviewed)

☐ B. Previously Approved Study for which IRB Approval has Lapsed : Previous IRB # _____
Please include with your submission either a written statement that verifies no research activities (recruitment or enrollment of new subjects; interaction, intervention, or data collection from currently enrolled subjects; or data analysis) have occurred since tha lapse in approval, or a summary of events that occurred in the interim.

☐ C. Modification to Currently Approved Protocol

1. Check type of review: Expedited _____ Full: _____ Check IRB: Medical _____ Nonmedical _____

2. Name and Address of Principal Investigator (PI) (where mail can most easily reach PI): If research is being submitted to or supported by an extramural funding agency such as NIH, or a private foundation, the PI listed on the grant application must be the same person listed below. If the PI is completing this project to meet the requirements of a University of Kentucky academic progam, also list name and campus address of faculty advisor.

   PI Name: _____ ☐ PI is R.N.

   Department: _____

   *Room # & Bldg.: _____

   Speed Sort #: _____

   *Students should list preferred mailing address (i.e., an address where mail will most quickly reach them).

3. PI's Link Blue: _____ Degree and Rank: _____
   ("username" to log in to your UK network account, i.e., jdoe)

   PI's Telephone #: _____ Dept. Code: _____

   PI's e-mail address: _____ PI's FAX Number: _____

4. Title of Project: (If applicable, use the exact title listed in the grant/contract application. **When applicable to your research, it is important that yoy add to the beginning of your title the following: "UK/P" if your research involves prisoners; "UK/D" if your research is supported by the Department of Defense".**

   _____

   _____

   _____

   _____

FORM A F2.0050

GENERAL INFORMATION SHEET: NONMEDICAL IRB

5. indicate which of the categories listed below accurately describes this protocol:

☐	Not greater than minimal risk
☐	Greater than minimal risk, but presenting the prospect of direct benefit to individual subjects
☐	Greater than minimal risk, no prospect of direct benefit to individual subjects, but likely to yield generalizable knowledge about the subject's disorder of condition
☐	Research not otherwise approvable which presents an opportunity to understand, prevent, or alleviate a serious problem affecting the health or welfare of subjects

6. Anticipated Beginning and Ending Date of Research Project: _____ (Month/Day/Year) / _____ (Month/Day/Year)

7. Number and age level of human subjects: _____ (Number) / _____ (Age Range)

Indicate the categories of subjects and controls to be included in the study. You may be required to completed additional forms depending on the subject category applicable to your research. Check All that apply:

- ☐ Children (17 yrs or less) [attach Form W]
- ☐ Wards of the State [attach Form W]
- ☐ Emancipated Minors
- ☐ Impaired Consent Capacity [attach Form T]
- ☐ Impaired Consent Capacity (Institutionalized) [attach Form T]
- ☐ Neonates [attach Form U]
- ☐ Pregnant Women [attach Form U]
- ☐ Military Personnel or Dod Civilian Employees [DoD SOP may apply]
- ☐ Prisoners [attach Form V]
- ☐ Non-English Speaking [see Form H info (HTML)]
- ☐ International Citizens [DoD SOP may apply]
- ☐ Students
- ☐ Normal Volunteers
- ☐ Patients
- ☐ Appalachian Population

8. Does this study focus on subjects with any of the clinical conditions listed below that present a high likelihood of impaired consent capacity or fluctuations in consent capacity?

☐ No

☐ Yes

If yes, does the research involve interaction or intervention with subjects?

☐ No, direct intervention/interaction is not involved (e.g., record-review research, secondary data analysis)

☐ Yes - direct intervention/interaction is involved - complete and attach Form T to your IRB application.

Examples of such conditions include:

- *Traumatic brain injury or acquired brain injury*
- *Severe depressive disorders or Bipolar disorders*
- *Schizophrenia or other mental disorders that involve serious cognitive disturbances*
- *Stroke*
- *Developmental disabilities*
- *Degenerative dementias*
- *CNS cancers and other cancers with possible CNS involvement*
- *Late stage Parkinson's Disease*
- *Late stage persistent substance dependence*
- *Ischemic heart disease*
- *HIV/AIDS*
- *COPD*
- *Renal insufficiency*
- *Diabetes*
- *Autoimmune or inflammatory disorders*
- *Chronic non-malignant pain disorders*
- *Drug effects*
- *Other acute medical crises*

Figure 2.1 Sample forms from an expedited IRB application at the University of Kentucky. Retrieved from www.research.uky.edu/ori/human/HumanResearchForms.htm.

Increasing the Probability of Approval

The best recommendation to increase the probability of approval on the first submission of the application is to write clearly and succinctly elaborating on those points that are

likely to be viewed critically, easily misunderstood, or sensitive. In other words, the skill required for preparing an article submission to a peer-reviewed journal will serve to prepare a research application for human subjects review. The application should also use technical (jargon) free language that can be understood by a review board, which may be comprised of people from an array of disciplines. Investigators new to the field of research may also find it helpful to consult other investigators who have recently had successful submissions. You should take special care to describe your procedures in a way such that they can be understood by someone without specialized knowledge in your field; reviewers of your application should be able to understand exactly what will happen to your participants, who will be doing it, for how long they will be doing it, and under what circumstances they will stop.

Special Populations

Researchers should note that special populations receive specific attention in the application and may require full review by the IRB, depending on the nature of the intervention and vulnerability of proposed participants. For example, if you plan to conduct a study with young children with autism, you might need to justify the need to include these participants in the research; this often involves describing and explaining characteristics of participants that make them likely to benefit from the potential research. Selecting vulnerable populations, such as individuals with disabilities, those who are institutionalized, or people who are imprisoned, due to convenience or availability is not an acceptable justification. Guidelines also require a description of safeguards for protection of vulnerable populations. For example, you might need to describe what special forms of dissenting behaviors will be accepted for individuals with limited communication repertoires and to explain whether your intervention sessions will interfere with ongoing therapy regimens.

Potential Risk

Researchers will be required to indicate the level of potential risk to participants and whether the level constitutes "minimal" or "more than minimal" risk. **Minimal risk** is considered to be the same risk that a person would encounter in daily life or while performing routine physical or psychological examinations (United States Department of Health and Human Services: Code of Ethics, 2005). At first glance, it might seem that the issue of potential risk is easily dismissed in a classroom-based academic intervention or in a social behavior change project employing positive consequences. However, risk may be interpreted broadly. Suppose that the researcher engages a student in an intervention during a time when the student would otherwise be receiving academic instruction; then suppose that the intervention does not succeed. The student's behavior is unchanged and he has lost instructional time. If the intervention proceeds for an extended period

without positive results, is the researcher responsible for the student's falling behind schedule? Suppose that a traditional and an experimental instructional program are presented alternately to students who perform better under the traditional program. Has the experimental program interfered with more effective instruction and thus disadvantaged the students? Suppose that the intervention involves physical manipulation of the research participants (e.g., using physical prompting of motor responses such as switch activation) with children with cerebral palsy. What is the potential risk of physical injury to a child who resists or responds defensively? How will the researcher decide if the participant is being "harmed" and what alternate plan will be employed to assure that the element of risk is removed? What are the risks involved with community-based instruction where research participants will be required to cross streets, ride public transportation, or learn to seek adult assistance when "lost" in the community?

We assume that researchers generally will not be permitted to conduct research that involves the presentation of aversive or noxious stimuli and, thus, that is not a topic of concern here. However, suppose that an intervention involves positive consequences for a correct response and extinction for an incorrect response. What level of risk is present due to this intervention or the distress exhibited by a student? In response to these issues, the researcher might consider that the length of an intervention may constitute only a fraction of a school day and that there may be ways to make up potentially lost instructional time. Further, it is possible to describe how one "feels" physical resistance during prompting, or to list the obvious signs of behavioral responses that signal distress. Thus, observable behaviors serve as a proxy for concepts that might, at first glance, seem difficult to define. For example, suppose in a small project conducted as a course requirement, a professor and a graduate student attempted to shape *drinking from a glass* by a student with severe intellectual disabilities and minor physical limitations. Having observed and determined that the child could complete all movements in the response chain, they developed a program to shape a consistent and durable response using physical guidance, extinction (looking away), and withholding a preferred drink until a correct approximation response was emitted. In the midst of the program, the child began to whine, cry, and then tantrum when the drink was withheld. At this point, the classroom teacher intervened to terminate the intervention since it was obviously so distressing to the student (in the teacher's opinion). Suppose that this intervention had been a student research project and had been challenged by a human subjects review committee member; the protocol might have been developed in this way:

1. The target is drinking from a glass, without assistance, at every meal.
2. The benefits—the child will acquire a new skill and adults will be freed from the necessity of helping the child with the task.
3. The teaching sequence will be:
    a. Set glass in front of student
    b. Bring the student's hand and arm to proper position

c. Physically prompt drinking by placing hand over the child's and initiating the grasp, lift, tilt, drink response chain
d. Gradually withdraw physical assistance (operationally defined)
e. Praise at each step
f. When resistance occurs
    i. Remove glass from the child's hand,
    ii. Turn away for 5 seconds,
    iii. If tantrum behaviors occur, continue with intervention for three sessions or until tantrum behaviors are not emitted,
    iv. If the previous step is unsuccessful after 3 sessions, withdraw the student from the study or (preferably) move to modified procedures

This brief outline of a strategy contains two important elements. It acknowledges that a negative response may occur and sets a limit on the length of time that the behavior will occur before the intervention is removed or modified. This strategy sets the stage for the researcher to account for tantrum behaviors as a typical response in an extinction procedure and then allows for the development of an alternate strategy.

The questions raised here are intended to sensitize the researcher to different perspectives on the issue of risk, to raise issues that cannot be answered definitely, and to suggest in one instance a strategy to account for the possibility of duress. Researchers may find it helpful to share their protocols with fellow students to identify potential sources of risk and challenge the rationale for engaging in the project. Even when a risk is somewhat unlikely, it is prudent to identify potential problems and specify solutions a priori.

Defining the Methods and Procedures

The human subjects review application requires an abbreviated version of the written thesis research proposal (see Chapter 3 for information regarding writing a proposal). We recommend that the human subjects review prospectus be drawn from a fully developed proposal to ensure you begin the task of technical writing early in the research process. The human subjects review process focuses on specific elements of the procedures. The protocol requires a complete but abstracted description of the procedures. Then, special attention is directed to two questions, "What will happen to the participants?" and "What will happen to the data?"

Researchers with a background in education, psychology, and the various therapies (speech, occupational, physical) should be skilled in task analysis and should be adept at writing an explicit description of the steps or sequence of events in the research procedure. You will find it helpful to "walk through" the procedure as you write it and "talk through" the procedures with colleagues. A review committee is less likely to take special interest in antecedent events that are common or easily defined, than in esoteric or potentially noxious stimuli. If academic behavior change research projects use

common materials (e.g., a well-known basal reading series), it should be sufficient to identify the materials by publisher, content area, and daily "units" of instruction. If academic stimuli are experimental and designed specifically for the study, you may be questioned to determine why they are expected to produce positive results. If assistive technology is used, the dimensions or features of the device should be presented. If a procedure involves a series of statements and actions by the researcher, these should be written exactly as they will occur. If the procedure involves physical prompting, the nature and degree of effort should be described (e.g., the researcher will say, "Ralph, throw me the ball", and if no spontaneous response occurs, she will gently grasp him at the wrist and lift the hand/arm so that it rests on the ball). Physical assistance will be terminated if the child pulls away, cries, or shows other signs of distress. If consequences involve preferred edibles or liquids, the review committee is likely to request justification and assurances by the researcher that nutritional and allergy factors will be considered. That challenge should be satisfied by describing what is now common practice—assessing a child's preferences; consulting with teachers, parents, professionals, and significant others; identifying a menu of reinforcers; using a schedule of reinforcement; and so on.

We suggest that researchers include, in their original IRB application, a "Plan B" for all research studies. Because single case designs (SCDs) are dynamic in nature, you can (and should) modify or change interventions in the case of non-response. If you need to make substantive changes (e.g., use a different intervention), and you have not specified this in your original IRB, you will need to go through a potentially lengthy amendment process later. Thus, it is prudent for researchers to assume that the planned intervention may not work optimally for all participants, and to explicate a priori conditions under which a modified or different intervention will be used.

Data Storage and Confidentiality

The concern for "What happens to the data?" is based on three factors: (a) Is the information sensitive? (b) Can individual participants be identified? (c) Is there a plan to control access to the data and then to destroy it when it is no longer needed?

Data Storage

Researchers should ensure careful storage of all data collected for research purposes. Data should be stored separately from identifying information (e.g., consent forms with participant names stored in a different location than participant data—which can be identified with an identification number rather than a name). If sensitive tests are conducted as part of research (e.g., tests identifying a child's IQ or achievement levels, which may be considered "high stakes" in schools), researchers should take care to report to participants (or guardians) the manners in which they will and will not be used (i.e.,

whether you will share individual results with non-researchers). It is most conservative to treat all data collected as potentially sensitive, and to always use identification numbers rather than participant names or other identifiers. Note that participation in SCD studies is generally not **anonymous**—that is, researchers will be generally able to connect data with a specific participant. Non-anonymous data collection always results in the *possibility* that participant data could be matched with the corresponding participant by a non-researcher. Examples of anonymous data collection include asking a large group of teachers to fill out questionnaires without asking them to report their names or any identifying information. Generally, SCD data are not anonymous; thus, we must take appropriate steps to ensure confidentiality of participants.

Confidentiality

Protecting the **confidentiality** of participants (i.e., ensuring that *only* researchers can tie individual responses with a particular participant) is a potential problem in SCD research studies. To minimize the potential for loss of confidentiality, you should (a) describe how participants will be coded (e.g., by fictitious names or initials), (b) verify that the researcher will be the sole holder of the code (or the researcher and academic advisor), and (c) state where the code will be stored (e.g., in a locked file in the advisor's office). When a study is prepared for publication, you may use a fictitious name and so label it, fictitious initials, or the real initials of the participant. The location where the study was conducted may be described in ambiguous terms (e.g., a resource room in an elementary school in a medium-sized city in the Northeast). Participants should be informed about how all data including photographs, audio, and video recordings will be used and stored. You should also be cognizant of the vulnerability of information exchanged electronically through the Internet. Expert advice may be necessary to learn how to protect data and confidential information and participants should be informed of the risks to privacy and limits of confidentiality of information exchanged electronically (Smith, 2003).

A special problem arises when a study is conducted in a small school or community, when a participant is unique (e.g., the only child with cerebral palsy in the school). Under such circumstances, the review committee may question the disposition of the final report and the number of individuals who have access to it. Academic review committee members will be well aware of the disposition of the research and who will have access to it. If a study is prepared for publication, the committee may question whether confidentiality can be sufficiently guaranteed, arguing that any person reading the article would recognize the participant. Professional journals, however, are specialized and have a relatively limited circulation. Thus, the probability is low that someone from the local community would have access to the information. Presumably one who did have access would treat the information in a professional manner, but that is outside the scope of concern for the researcher.

Confidentiality is an extremely complex issue that requires considerable attention by

applied researchers. Accepted confidentiality procedures have been delineated in the Belmont Report (National Commission for the Protection of Human Subjects, 1978) and by the American Psychological Association (APA, 2002, section 5), Behavior Analyst Certification Board (BACB, 2014, section 10), and incorporated into IDEA. At the end of the research project, the researcher should have a file of raw data sheets, coded for anonymity, and in a separate place, the key to the code. Whether this information should be destroyed is a matter of judgment. If the study is publishable, good scientific practice dictates that the raw data should be kept intact so that other researchers have access to it if they challenge the findings or otherwise wish to examine the data. The APA (2009) position, as stated in the *Publication Manual of the American Psychological Association*, is that raw data be retained for a minimum of five years after publication of the research. Researchers take great care in protecting the confidentiality of study participants, however, in rare cases the legal system may require divulging information. Procedures for storing and destroying data may vary across researchers and studies; the critical point is that you need to delineate specific procedures before your study begins and then closely follow them during and after the study.

Informed Consent and Assent

Written **consent** must be obtained from the participant or the participant's parent or legal guardian. The critical elements of informed consent are:

1. The procedures must be described fully, including purpose and expected duration.
2. Potential risks, as well as benefits should be discussed.
3. Consent can be revoked at any time, and the participant is free to withdraw from participation (or withdraw his or her child from participation).
4. The consent form and the description of the study must be communicated in simple language, at approximately an 8th-grade reading level.
5. Information as to who to contact if questions or concerns arise during the study should be shared.

Assent *and* informed consent must be obtained if working with minors or participants who cannot legally provide informed consent, including some individuals with disabilities. Assent is non-legal permission provided by this participant. For example, you might ask a high-school aged student to sign a simple form stating they understand the research and want to participate in it. Or, you might read a script to a very young child (e.g., "We're going to do some work every day after circle time. You can say no if you don't want to come with me. Are you ready to do the work now?"), and the *researcher* may sign the script, attesting that they read the script and the child agreed to participate. Researchers should allow participants to *dissent* as well—to decide at any time that they do not want to participate, separately from their guardian's rights to withdraw consent.

Sharing of Information

Parents, as well as other persons involved in the research project, are likely to be interested in the outcome of the study. Human subjects review committees require that such information be provided. In order to conduct a study, at a minimum, you may be expected to share information with participants' teachers and therapists, and sometimes the school principal or clinic administrator. Other professionals (e.g., speech, occupational, and physical therapists) may benefit from knowing the results of the study as well. You should learn what is typical, or expected, in the school system, clinic, or community business where the study is to be conducted, and should list those who will be informed of study outcomes in your IRB protocol and your consent documents. You should decide in advance how detailed an explanation will be given to those who have limited direct involvement with the participant. Parents, teachers, and therapists may request a step-by-step review of the study, focusing on daily sessions where performance was well above or below other data points. Others will be satisfied with a general description of the procedure and the extent to which it was successful. Refer to Table 2.1 for examples of ethical scenarios and appropriate responses related to data storage, anonymity, confidentiality, and informed consent and assent.

Expertise of the Researcher

Human subjects review committees require assurance that the researcher knows what she is doing and, if a student, that she is going to be supervised by a knowledgeable faculty member. Researchers who have teaching or clinical experience should list and describe the length and type of the experience. Certification, licensures, and endorsements should be shared. The committee may wish to be assured that practitioners have worked with children, have worked in and/or understand the protocol of working in schools, and have experience with the procedure under study. It is helpful to explicate (and carry out) specific training procedures for implementers—for example, you may set a training criterion of 90% correct and accurate implementation of procedures prior to study onset.

Table 2.1 Ethical Scenarios Related to Data Storage During Study and Additional Compromise of Anonymity and Confidentiality.

Scenario 1	*Response 1*
Jonathan is a third-year doctoral student in special education who is implementing a study with preschool age children with Down syndrome at a local public school. Per guidelines laid out by the university's IRB, Jonathan is to keep data in a locked file cabinet in a locked office on campus. Due to	Jonathan should adhere to guidelines in place per the IRB and store data in a locked cabinet in a locked room on campus. Data should only be in his car during his commute from the school to campus. In addition, Jonathan is required to report to the IRB that potentially confidential

Jonathan's schedule he does not always adhere to the IRBs guidelines for storing data and keeps data at home and in his car for purposes of convenience. One day, Jonathan's home is burglarized and his backpack, which contained his computer and data, were stolen.

Scenario 2

Cora is a first-year professor in psychology and her research interests include training teachers to conduct and implement functional behavior assessments for students with intellectual disability who display aggressive behavior during academic tasks. Cora received a grant to conduct a study on teacher training for decreasing aggressive behavior. One evening, when Cora is purchasing groceries, she meets by chance a paraprofessional who works in a classroom where she is conducting her study. The paraprofessional asks questions about study participants and proceeds to provide personal information, as well as things she has heard others say about the participants. While hesitant to discuss participants, Cora does not want to offend the paraprofessional and discusses current classroom issues.

Scenario 3

Matthew is a professor who specializes in increasing social interactions for high-school age students with social delays. He is conducting a study, training students to practice appropriate social interactions with same age peers. During training, an adult provides prompts for participants to engage in appropriate social interactions. During the third week of instruction Jon decides to drop out of the study because

information was stolen by an unknown party and participants' confidentiality and related information is compromised. Jonathan would also likely need to resubmit his IRB application and conduct the study again, pending approval of the IRB after consideration of stolen data.

Response 2

Cora should not discuss research participants in a public setting due to a possible breach of confidentiality and potential exposure of personal information to persons who may know participants and/or their families. Cora should also be aware that she should not discuss specific information about participants with someone not directly involved in the study. Cora should have indicated to the paraprofessional that she ethically cannot discuss participants due to confidentiality issues.

Response 3

Matthew is attempting to coerce Jon to remain in the study, even though it is Jon's right to leave the study at any time. Jon reported feeling uncomfortable and Matthew responded by providing multiple statements about Jon's social skills. Prior to implementing the study, Matthew should provide clear guidelines for responding to participant requests to leave a study in the IRB

intervention sessions make him "uncomfortable" and "anxious". Matthew tells Jon he needs to remain in the study because it will "help him interact with peers". Jon does not agree, but Matthew tells him that he needs social support and this study can assist in improving his social skills. Matthew also informs Jon that the peer involved in the study will want to be his friend once the study is complete.	application. Coercion is never an option for persons conducting research studies. In addition, Matthew reported false claims related to the effects of intervention on friendship, which Matthew was not directly measuring. It is the responsibility of researchers to only provide known information to participants and not do so in an attempt to coerce participants to start or continue participation.

Publication Ethics and Reporting of Results

You will face additional challenges regarding the preparation and submission of a written manuscript for publication consideration (also refer to Chapter 3) after the completion of the formal research procedures.

Publication Credit: Authorship

The *Ethical Principles of Psychologists and Code of Conduct* (American Psychological Association, 2010) recommends under Standard 8.12c that faculty advisers clearly discuss publication credit with students from the onset of their research relationship and that discussion continue throughout the research process. Some institutions may have formal procedures for establishing credit and obtaining an authorship agreement among students and faculty. Others may rely on a verbal agreement or "understanding" between contributors to the research. Standards 8.12b and 8.12c specifically address provision of credit for students who substantially contribute to the conceptualization, design, and implementation of the research and who analyze or interpret results of the study. For masters and doctoral students who are conducting their capstone projects (i.e., thesis, dissertation), this implies first authorship, unless there are "exceptional circumstances" (8.12c). It is important that early agreements contain information on the tasks to be completed, the level of credit that should be given (order of authorship), and that students, new to research and publishing be made aware of the guidelines set forth by the APA (2010) and the BACB (2014, section 10) on publication credit.

Reporting of Results

Researchers will likely be familiar with ethical procedures for preparing manuscripts and professional documents which avoid the issue of plagiarism or the use of others' ideas and work without proper credit being given to the author or originator of the work. They may be unaware, however, that these procedures apply to their work as researchers, even if the work is completed under the direction of faculty advisors. They may be even less familiar with provision of intellectual credit for non-published material including information shared at meetings, conferences, and through informal conversations with advisors, other students, and professionals. You should be given appropriate acknowledgement for your original ideas, and your work, whether published or unpublished should not be used by others for personal gain (Sales & Folkman, 2000). You should also be aware that your unpublished work is "copyrighted from the moment it is fixed in tangible form—for example, typed on a page" and that this copyright protection is in effect until the author transfers the copyright on a manuscript accepted

for publication (Publication Manual of the American Psychological Association, 2009, pp. 19–20). Finally, you will be required to present a statement with a manuscript submitted for publication consideration that the manuscript is not being simultaneously submitted to any other journal.

Researchers should take care to report *all applicable results*, including data for all dependent variables and participants. Although it is not uncommon for journal editors to request removal of specific participants (based on our experiences, generally because a participant withdrew or did not respond to the intervention), this increases the likelihood of biased results and is not an ethical practice.

Ethical Practice

A practitioner is likely to use SCDs under two conditions—as part of graduate training or as part of an evidence-based practice. In the first instance, you will follow the processes described in the first part of this chapter. In the second instance, you have a somewhat different set of responsibilities. When SCD is used as an integral part of the instructional or therapeutic process, you will seldom need to seek approval from the school or clinic administration. However, to the extent that such applied research represents an innovation, you are advised to make public the strategies (data collection, experimental design, baseline and intervention procedures etc.) that will be employed. When SCD investigations address *social behavior change*, ethical considerations for the use of positive behavior supports must be employed.

Table 2.2 Ethical Scenarios Related to Methodology, Results, and Publication of Data.

Scenario 1	*Response 1*
William is a second-year master's student with interests in reading instruction for middle-school students with dyslexia. William is implementing a study for increasing fluency of reading known passages for students with dyslexia who spend at least 50% of their day in a resource classroom. Two of four participants in William's study display challenges related to attending to materials and require multiple prompts to begin reading. William decides to add a specific attending cue for participants who require multiple prompts to begin the reading intervention. He decides to video record sessions to show colleagues for purposes of obtaining their feedback on changes. William did not include the specific attending cue or video recording permission in his IRB application	It is not necessary for William to report the addition of an attending cue to IRB prior to implementing the methodological change since it does not alter the primary intervention procedures or add any additional risk for participants, but he would report such changes in any final reports or publication of information. It is necessary for William to submit an amendment to his IRB application for purposes of requesting to video record sessions. If

and consent forms.	the amendment were approved, William would then need to obtain consent for video recording from participants and their legal guardians.
Scenario 2 Sheila is an associate professor of communication sciences at a research university and has been employed by the university for the past 10 years. Sheila has focused her work on increasing novel words students with autism use at home and school. She has published multiple articles replicating positive effects of a language intervention for students with autism and has decided to extend her work to students with aphasia. Following completion of her study with students with aphasia, using the language intervention, the results are highly variable with some students making no progress following 12 weeks of intervention. Sheila feels strongly the intervention was successful, even though the data indicates otherwise. She decides to submit an article based on her perceptions of the data and omit or limit information related to participants who made no progress.	*Response 2* It is the responsibility of persons involved in research to be honest when reporting results of a research study. While results may violate expectations of outcomes, personal biases related to expectations of results and related areas cannot impede clear, concise, and honest reports of results. While there are multiple issues with dishonest claims, some key issues to consider are future misuse of monies for persons who attempt to replicate this study and a waste of time and resources for persons who choose to use this intervention in practice with persons with aphasia.

The major element of ethical practice that applies to empirically verified intervention is the principle of full and open disclosure of information. Critical elements of IDEA require that individual students' programs be planned in conjunction with parents and in collaboration with other specialists and school administrators (i.e., Individual Education Program team). The scientist-practitioner's major tasks are to set up data systems, explain the logic for the specific research design, and describe how the design permits certain conclusions (i.e., evaluates threats to internal validity). Since these events go beyond typical practice, little disagreement should be encountered. Refer to Table 2.2 for examples of ethical scenarios and appropriate responses related to methodology, results, and publication of data.

Summary

In this chapter we have provided a context for conducting applied research within a set of ethical principles. We have stated our assumptions about the prerequisite behaviors necessary to conduct academic and social behavior change programs within the framework of SCD research methodology. Specific procedures have been listed, designed to help you obtain approval to conduct research in a manner that protects the rights of participants. We close by stating again that there are no clear answers to the problems we have raised, and finally, that these guidelines are useful only to those applying mature judgment to the problems they confront.

References

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct*. Washington, DC: Author. Retrieved August 13, 2007, from www.apa.org/science/research/regcodes.html

American Psychological Association. (2009). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

American Psychological Association. (2010). *Ethical principles of psychologists and code of conduct*. Retrieved from www.apa.org/ethics/code/

Ayres, K. M., Lowery, A., Douglas, K., & Sievers, C. (2011). I can identify Saturn but I can't brush my teeth: What happens when curricular focus for students with severe disabilities shifts. *Education and Training in Autism and Developmental Disabilities*, *45*, 11–21.

Ayres, K. M., Lowery, A., Douglas, K., & Sievers, C. (2012). The question remains: What happens when curricular focus for students with severe disabilities shifts? Rejoinder to Courtade et al. (2012). *Education and Training in Autism and Developmental Disabilities*, *47*, 14–22.

Behavior Analyst Certification Board. (2014). *Professional and ethical compliance code for behavior analysts*. Retrieved from https://bacb.com/wp-content/uploads/2016/03/160321-compliance-code-english.pdf

Breault, J. L. (2006). Protecting human research subjects: The past defines the future. *The Ochsner Journal*, *6*(1), 15–20.

CITI. (n.d.). *Course in the protection of human research subjects*. Retrieved February 22, 2013, from www.citiprogram.org

Courtade, G., Spooner, F., Browder, D., & Jimenez, B. (2012). Seven reasons to promote standards-based instruction for students with severe disabilities: A reply to Ayres, Lowery, Douglas, & Sievers (2011). *Education and Training in Autism and Developmental Disabilities*, *47*, 3–13.

General, S. (1966). *Surgeon General's directives on human experimentation*. Retrieved from https://history.nih.gov/research/downloads/surgeongeneraldirective1966.pdf

National Cancer Institute: U.S. National Institute of Health. (n.d.). *Human participants protections education for research teams*. Retrieved August 13, 2006, from www.cancer.gov/clinicaltrials/learning/humanparticipant-protections

National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research. (1978). *The Belmont Report: Ethical principles and guidelines for the protection of human subjects of research*. (DHEW Publication No. OS 78–0012). Washington, DC: Government Printing Office. Retrieved August 27, 2007, from http://ohsr.od.nih.gov/ guidelines/belmont.html.

National Institutes of Health (2008). Protecting Human Research Participants. NIH Office

of Extramural Research. Retrieved from: https://phrp.nihtraining.com/users/PHRP.pdf

Sales, D. B., & Folkman, S. (Eds.) (2000). *Ethics in research with human participants*. Washington, DC: American Psychological Association.

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

Smith, D. (2003). What you need to know about the new code. *Monitor*, *34*, 62. Retrieved August 13, 2007, from www.apa.org/monitor.jan03/newcode.html

US Department of Health and Human Services. (2005). Code of federal regulations (45 CFR 46). *Sub-part D: Additional Protections for Children Involved as Subjects in Research*.

Williams, E. P., & Walter, J. K. (2015). When does the amount we pay research participants become "undue influence"? *AMA Journal of Ethics*, *17*, 1116–1121.

3
Writing Tasks

Literature Reviews, Research Proposals, and Final Reports

Mark Wolery, Kathleen Lynne Lane, and Eric Alan Common

Important Terms

literature review, peer review, exhaustive search, demonstration question, parametric question, component analysis question, comparison question, research proposal, introduction section, methods section, manuscript, results section, discussion section, dissemination, PRISMA guidelines

Written language is a major way scientists, including single case design (SCD) researchers, establish a record of their work and communicate with one another and with practitioners. As with other writing forms, technical scientific writing has its own organization, style, and standards. Across disciplines those styles and standards vary, but value is placed on organization, factual reporting requiring minimal inference, precise

and concise detail, and brevity. Learning to write technically is similar to many other skills; it requires attempting the skill, purposeful attention to the act, feedback from others, and an ongoing commitment to improve. Scientific writing is so important that Baer, Wolf, and Risley (1968) in their seminal article on applied behavior analysis included it as a major dimension (i.e., technological). Their guideline for judging the quality of technological writing is: "The best [test] for evaluating a procedure description as technological is probably to ask whether a typically trained reader could replicate that procedure well enough to produce the same results, given only a reading of the description" (p. 95, information in brackets added). Technological descriptions are especially critical in applied research, because the procedures are used in non-standard settings (not in laboratories). Fortunately, a number of helpful resources exist such as the *Publication Manual of the American Psychological Association* (6th edition). This manual provides guidance on preparing written documents and is used by most journals in psychology, education, and other fields. Additional information can be found at www.apastyle.org. It is important to stay current regarding the most recent APA standards or similar standards used in other fields.

The purposes of this chapter are to (a) describe some of the standards for preparing technical documents and (b) present suggestions for helping readers acquire and use a scientific writing style. The chapter presents information on reviewing the literature, conducting systematic reviews and meta analyses using PRISMA procedures, stating research questions, writing research proposals for evaluation by others, and describing completed studies. The chapter concludes with information on disseminating information from studies.

Reviewing the Literature

Research is conducted for many reasons such as satisfying your curiosity about how nature works, solving problems presented by individuals to whom you provide services, fulfilling the requirements of a degree or grant, challenging or supporting a policy or practice, convincing others about the effectiveness of a technique, and many others. Regardless of the reason for conducting research, it starts with learning what has already been studied, which means reviewing the literature. Reviewing literature is an extremely beneficial use of time. A careful review allows you to discover what is known about a topic; how it has been studied (e.g., what measures and designs were used); what implications exist for practice or policy; what factors qualify or limit the findings; and what questions are answered, partially answered, or unasked. Reviewing the literature also allows you to develop a rationale for conducting specific studies. For example, studies may support using a procedure in a particular way to produce desired results, but if no one has studied a variation of the procedure that may make it easier to apply, then it is logical to focus on that variation in future studies. It also allows you to benefit from the successes and problems encountered by other researchers. How other researchers measured the behaviors of interest, controlled or failed to control for certain threats to internal validity, or applied an independent variable may yield important information to guide you in developing the purpose and procedures for your study. Thus, a **literature review** has three main functions: (a) articulating what is known and not known about a topic, (b) building a foundation and rationale for a study or series of studies, and (c) improving plans for future studies by identifying successful procedures, measures, and designs used by other investigators and detecting issues and problems they encountered.

Process of Conducting Literature Reviews

Several general steps are involved when reviewing the literature. These include (a) selecting a topic, (b) narrowing that topic, (c) finding the relevant sources, (d) reading and coding relevant reports, (e) sorting the sources with sound information from those with less trustworthy information, and (f) organizing the findings and writing the review. Literature reviews are used as introductions for study reports (i.e., proposals and articles), and stand-alone products such as review articles or chapters in books, theses, or dissertations. The processes for both types of products are similar, but they are different in how the literature is described and in how current published reviews on the topic are used. Introductions to study reports contain less detailed information about the reviewed studies than do stand-alone reviews. When writing an article introduction, recent reviews are very useful; however, when writing a stand-alone literature review, the presence of a recently published review of the topic suggests another may not be needed.

The following paragraphs place emphasis on conducting stand-alone literature reviews (often referred to as systematic reviews of the literature), but also provides information for writing introduction sections to reports on studies. In addition to these general strategies for conducting systematic reviews, we also provide a detailed description of the *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses—the PRISMA Statement (Moher, Liberati, Tetzlaff, Altman, & The PRISMA Group, 2009).

Selecting the Topic of the Review

Selecting and subsequently narrowing a topic for the literature review can be a challenging task. A primary reason for selecting a topic should be you are interested in it! Searching for, summarizing, evaluating, and describing the literature can be tedious and time consuming; it is painful if one is not interested in the topic. If you are not sure what is interesting, you can establish an interest by reading about the field you are studying. Find several issues of relevant journals, and read the abstracts of all the studies published in those issues. Identify the ones that you find intriguing and read those articles. Another complementary suggestion is to talk with professionals who are practicing in your field. Identify what issues they face and what dilemmas they encounter. Finally, no substitute exists for being involved in the profession you are studying. This can be done through volunteer work, field experiences, interactions with others in your professional organizations, and so forth.

Narrowing the Topic

Often, interesting topics are quite broad. Examples might be how to teach reading, how to make schools more effective, how to help families of children with Down syndrome, inclusion of children with disabilities in general education classes, the causes of and treatments for aggression, how to deal with the problem behaviors of adolescents, how to support students with internalizing behaviors, and the effects of neurotransmitters on daily functioning. These are interesting and important topics, but they could, and often have, filled entire books and then only in a cursory manner. Sometimes by reading the literature on the broader topic, a more refined or narrower topic will emerge. Of course there is no pure definition of what is too broad or too narrow. Some have as few as 10 while others have as many as 90 or more (cf. Doyle, Wolery, Ault, & Gast, 1988). Having fewer than 10 sources usually is not sufficient for a review.

A useful technique for narrowing a broad topic is to generate a list of questions about it. For example, if one was interested in the inclusion of children with disabilities in regular education classrooms, a number of questions could be asked:

1. What are the legal and social reasons for inclusive practices?
2. What do teachers find helpful in implementing inclusive practices?
3. What strategies are successful in training teachers for inclusion?

4. What teaching practices are effective in inclusive classes?
5. How prevalent is the use of inclusion?
6. What types of children with disabilities tend to be included?
7. What barriers exist to inclusive practices?
8. What effect does inclusion have on children with disabilities?
9. What effect does inclusion have on children without disabilities?
10. What are the roles of special educators in inclusion?

Once a list of questions is generated, one or two of them are selected as being of special interest for the focus of the review.

Another useful strategy for narrowing a broad topic is to conceptualize it by various study components. Common elements would be the independent variable (e.g., intervention, strategy, practice, or treatment), how the independent variable is implemented, the participants, the context, and the behaviors involved. Focusing on the independent variable often is highly useful. Examples of reviews using this strategy include reviews of:

1. Instructive feedback (Werts, Wolery, Holcombe, & Gast, 1995)
2. Alternative treatments (Green et al., 2006)
3. Functional assessment-based interventions (Common, Lane, Pustejovsky, Johnson, & Johl, 2017).

Further, these reviews can be narrowed by participant type, target behaviors, or contexts. Or, they can be narrowed via use of specific measurement systems or methodologies. For example:

4. Lane (2004) focused on academic and tutoring interventions for students with emotional and behavioral disorders.
5. Hitchcock, Dowrick, and Prater (2003) reviewed the literature on video self-modeling but restricted the topic to studies conducted in schools.
6. Ledford, Lane, Elam, and Wolery (2012) reviewed the studies using response prompting procedures (independent variables) in small group arrangements.
7. Schuster et al. (1998) reviewed the literature on constant time delay, but restricted the review to a specific type of behaviors, chained behaviors.
8. Ledford and Gast (2006) reviewed the literature on feeding problems in children with autism.
9. Munson and Odom (1996) focused on the use of ratings scales for measuring parent- infant interactions.
10. Logan and Gast (2001) were interested in identifying reinforcers for students with significant disabilities, and they narrowed their review to preference assessments.
11. Lane, Robertson, and Graham-Bailey (2006) focused on methodological considerations in school-wide interventions in secondary schools.

12. Alexander, Smith, Mataras, Shepley, and Ayres (2015) reviewed data regarding baseline conditions in studies designed to assess acquisition of chained tasks.

In three special cases, a method for narrowing a broad topic is to limit the review to recently published research. These special cases are (a) the past publication of a comprehensive review of a topic, (b) the publication of a major conceptual paper or seminal article on the topic, and (c) the literature published after a specific event that should impact the research field in a given way. In the first case, if a review on the topic was published 10 years earlier, then limiting that review to research published since the first review is appropriate. When this is done, you should account for the publication lag. For example, a review published in 2000 was probably completed and submitted to the journal in 1998 or 1999; thus, papers published in 1998 and 1999 may not be included in the original review, and those dates should be searched in a new review of the topic. In the second case, a major conceptual paper may have impacted subsequent research. For example, Stokes and Baer's (1977) review of procedures for promoting generalization was a major milestone in the applied behavior analysis literature. Reviewing the research on generalization since publication of that paper would be relevant. Finally, in the third case, examples of events that may impact the research field would be federal laws, such as No Child Left Behind or the Individuals with Disabilities Education Improvement Act of 2004. In isolated situations, authors have restricted their review to a specific journal. For example, Gresham, Gansle, and Noell (1993) conducted a review of studies reporting treatment integrity measures in the *Journal of Applied Behavior Analysis*.

Finding Relevant Sources

Finding relevant literature requires persistence and use of systematic search strategies. The goal of literature reviews often is to describe what is known about a topic; as such, you are obligated to find all the relevant sources. Before starting a search, however, identify what you are trying to find. This is determined, in part, by narrowing your topic (discussed above). It is also necessary to specify inclusion and exclusion criteria. These criteria are used to sort relevant from irrelevant sources. Inclusion criteria are the specific factors or characteristics a study or article must have to be included in your review; and exclusion criteria are the specific factors or characteristics of a study or article that exclude it from being reviewed. These criteria vary greatly across reviews and often are related to how the topic was defined and narrowed.

The "data" for most literature reviews are research articles rather than chapters, discussion articles, and other reviews of the topic. A common inclusion criterion is to include only research articles, and often only research of a given type (e.g., experimental rather than descriptive or causal-comparative). Chapters, discussion papers, and review articles on the topic are relevant. They may provide useful background information on the topic, but they are not the primary sources for understanding what is known from research.

Many reviewers include only journal articles, because such documents have undergone **peer review**, meaning impartial judges have read and evaluated the study and concluded it was worthy of publication. Dissertations, conference presentations, final reports, and similar documents may be peer reviewed, but the level of evaluation may be less rigorous. These sources tend to be less accessible and sometimes they are eventually published as journal articles. However, using unpublished sources may be a way to minimize publication biases if only published literature—which often does not include studies with null effects—is used (Cook, 2016).

A third strategy is to include only sources published in the English language and exclude sources published in other languages. This practice is often done for expedience, because of the time, difficulty, and cost of translating the source into English can be excessive. Nonetheless, limiting the review to English language publications may limit the knowledge about the topic, and this should be acknowledged in the discussion section of the review

Finally, many individuals include characteristics of their population of interest as an inclusion/exclusion criterion. These may be the age of the population (e.g., preschoolers, adolescents), their diagnoses, or other characteristics (e.g., incarcerated, homeless). While this is useful, some practices of interest are used across demographic characteristics. For example, some practices (e.g., providing choices) have been studied with preschool-, elementary-school, middle-school-, and high-school-aged participants (e.g., Kern et al., 1998; Shogren, Fagella-Luby, Bae, & Wehmeyer, 2004). If a search were limited by age, then the reviewer would get a partial picture of what is known about this intervention. Similarly, many practices used with a given diagnostic group may have been studied with participants with other diagnoses as well. There is a balance between finding all research reports relevant to the topic *versus* the specific focus of the review. Including and excluding studies based on functional characteristics of participants may be more defensible. For example, including studies that only included participants whose problem behavior was maintained by attention (or by escape) is a viable division. Once the inclusion and exclusion criteria are identified, the reviewer should write a description of each criterion. This information will become part of the method section (described below) of the literature review.

Five search strategies are suggested to ensure an **exhaustive search** (e.g., to make sure you have included all studies that meet your criteria): electronic searches, ancestral searches, hand searches, author searches, and expert nomination. Perhaps the most widely known search strategy is the use of electronic search engines. This does not mean typing a term into general search engines; rather it means taking advantage of specific databases of scientific publications available in most university libraries or online through a library connection. Common examples of such search sources are PsycINFO, Medline, and ERIC or the publicly available Google Scholar. A disadvantage of Google Scholar is the lack of replicability of the search process, as the internet and what is accessible to its search algorithms are always changing (e.g., paywalls). Databases affiliated with university library subscriptions, such as PsycINFO, may be more likely to

remain constant. Other electronic databases exist and others will appear in the future, but these are established and widely recognized in behavioral sciences. Other disciplines, such as sociology, will have their own databases; select the databases most relevant to your topic.

An important issue in using electronic searches is the terms entered for the search. Most databases have instructions or suggestions for selecting and entering terms to identify the largest number of relevant sources (e.g., Boolean operators). They often allow for combining terms to make the focus of the search more precise. Reading and following the instructions for the various databases is highly recommended. Most of these databases have an option for displaying an abstract of the study, which is useful in making an initial decision about whether a report potentially meets the inclusion criteria. It is often wise to use multiple electronic databases, because some studies might be found using one but not another.

Keep a record of (a) what search terms were used, (b) how many sources were found on each search, and (c) how many of the found sources met the inclusion criteria. This information is often reported in the literature review. After each search, make a list of the reports that appear to meet the inclusion criteria. This list should contain the complete reference and should be written as an APA style reference list. After conducting electronic searches and finding the full body of the selected reports, a second search strategy should be implemented. This strategy, called an ancestral or bibliographic search, involves examining the reference list of each report as well as the reference lists of reviews of your topic. This strategy often results in finding additional reports meeting the inclusion criteria. These reports may have been published in journals not included in the electronic database or may be relevant but for one reason or another did not appear during the electronic search. This is also a useful way to identify non-research reports (e.g., reviews or chapters) that can help in describing the topic.

Another useful search strategy is to conduct a "hand" search of selected journals. Usually, research reports on given topics are published in a few journals. These journals can be selected based on their reputation or on the frequency with which they appear in the list of reports generated from other search strategies (e.g., searching journals in which two or more articles identified in the previous searches were published). When doing a hand search, it is useful to read the abstracts of each article; this is time consuming but important as it often yields additional relevant reports.

From the reference list of reports meeting the inclusion criteria or appearing to meet the inclusion criteria, scan the list for authors who have multiple entries. If there are authors who have published multiple times on a topic, which is common, then it is wise to do an electronic search of those authors' names. Many of the electronic databases allow for searching by author name. This strategy may well identify additional relevant reports.

After completing the above search strategies and finding all reports that appear to meet the inclusion criteria, read the reports and make a final determination of whether they will be included in the review. Once a list of reports is established, it is useful to

identify authors who appear to be actively conducting research in the topic. Two strategies are useful; you can search at their academic institution and see if their curriculum vita is posted. If it is up to date, you can scan through the list of publications to identify ones relevant to your search. You can also contact them through email or regular mail. Send a message saying you are conducting a literature review on the topic and include the reference list of identified sources. In the message, ask whether they are aware of any other reports, whether they have any relevant reports accepted but not yet published (i.e., in press papers), and if they would send you a copy of in press papers. Not everyone will respond, but many will. Keep track of to whom such messages (e.g., those with three published and included articles) are sent as well as their replies. This information will be included in the method section of the literature review. Similarly, you might also contact journal editors for journals publishing each article included in the review to determine if there are any "in press" for the topic at hand.

In most cases, using the above search strategies will result in finding all relevant sources. Often after completing two or three of these strategies, no new reports will appear. When this occurs, it often indicates the relevant sources have been found. Report in your review how many sources were found with each search strategy. We also encourage you to assess the reliability of the search procedures as one safeguard for ensuring all relevant articles are detected (see Kettler & Lane, 2017 for additional direction).

Reading and Coding Relevant Reports

After finding the relevant reports, they must be read and information gleaned from them. A coding sheet (or electronic database) should be constructed to enter the information from the reports. Often a set of rules and definitions is written to guide the coding of reports. Although coding sheets will vary by topic, some general information is needed: the reference and study purpose or research questions. In addition, information is needed about participants (e.g., age, gender, diagnosis, race, ethnicity), the setting, materials, response definitions, measurement procedures, independent variable, and findings. Often, information is recorded about the methodological rigor of the study, including the design, number of replications, and presence of specific threats to internal validity. The function of the coding sheet is to summarize information from each study which in turn will allow more efficient and accurate description of the studies. When deciding what components to code, one option is to review and code the quality features of SCDs posed by Horner et al. (2005) and more recently CEC (2014). It is important to consider the rigor of the studies included in a literature review; for more on this topic, see Chapter 13.

Organizing Findings and Writing the Review

Literature reviews and articles often have parallel sections: an abstract, introduction, method, results, and discussion. The abstract is often a single paragraph. The *introduction* also usually is relatively short (2–8 pages), identifying the topic, describing the rationale for reviewing the literature on the topic, and ending with a purpose statement. Although sometimes not included, the purpose statement can contain a series of questions about the literature. Generating such a list is useful for organizing the results section.

The second section of the literature review is the *method*. This section includes a description of the search strategies used to find the literature and what the results of each strategy were. If electronic searches were conducted, then the terms entered into the search should be listed. The method section also should specify the inclusion and exclusion criteria for selecting individual study reports. Finally, this section includes a description of the coding manual and coding procedures used to analyze the study reports. Ideally, two or more individuals would review a subset of the study reports using the same coding definitions to calculate inter-coder agreement (*see* Chapter 5). The proportion of study reports coded by two or more persons, the method for calculating interobserver agreement (IOA), and results of those calculations should be described in the method section. Ideally, two or more individuals use the same search procedures to see if they identify the same articles for review; if this is done, the agreement in finding articles also should be reported here. Random selection of included and excluded studies can be used to calculate inter-rater agreement of the inclusion and exclusion criteria.

The third section of the paper is the *results* section. This is the most idiosyncratic section of a literature review; it varies greatly by the content being studied as well as the purpose of the review. A logical organization, of course, also is helpful to the reader. Some reviews describe each study separately and sequentially—one study after another. This organizational structure should be avoided, because it does not allow for easy synthesis of the findings across studies for the writer or reader. Instead, we recommend the research questions posed at the conclusion of the introduction be used to organize the structure of the results section.

To assist in the synthesis, a useful method is to put the coded information from each study in a table. Such tables are often constructed with each study occupying a row and the variables of each study occupying the columns. The tables are often long (many columns), but allow quick examination of various variables across studies. For example, this would allow you to scan the table and get a picture of who the participants were, what locations were studied, what measures were used, variations on the use of the independent variables, and many other important elements of the studies. Such tables make major differences across studies obvious; for example, you can readily scan the table and identify the studies in which a generalization measure was included, or the studies in which persons indigenous to the participants' service systems implemented the independent variables, which designs were used, and so forth. Such tables in their totality rarely are included in the final written document. Rather they are used to get an overview and identify elements for description and discussion. Sometimes parts of such

tables (e.g., a table of the participants' characteristics and settings; a table of results) may be included in the final review. Additional information on synthesizing outcomes in a review is discussed in Chapter 14.

The *discussion* section is the final section of the literature review. This section should restate the purpose of the review and it should summarize the major findings from the review. Often, a helpful way to discuss the individual findings is to report the limitations or qualifications of each finding as well as directions for future inquiry. It is more useful to know when, with whom, and under what conditions a finding (i.e., functional relation) appears to exist than to simply state a functional relation exists. It also is important to note when, with whom, and under what conditions a given functional relation does not appear to exist. From the reporting of the findings, implications can be drawn for future research and for practice. Finally, the literature review should contain an articulation of the limitations of the review. For example, if only reports published in peer-reviewed journals were included, then it is useful to qualify this restriction to that fact. Another part of the discussion section is to identify issues needing more research. This can be in the context of discussing each major finding or can be a separate section of the discussion. Similarly, drawing implications for practice is often a subsection of the discussion section.

Using the Literature Review

As noted above, the function of literature reviews are (a) to describe what is known about a topic, (b) to build a rationale for a study or series of studies, and (c) to identify research procedures to strengthen a study. For the latter two functions (building a rationale, and identifying procedures), the review may be primarily useful to the person who conducted the review. However, when the first function (describing what is known) is met, then the review may be useful to other individuals. It can be used as a chapter in a book, thesis, or dissertation, but it also may be suitable for submission for review and possible publication in a professional journal.

PRISMA Guidelines for Reviews

As discussed more thoroughly in Chapters 13 and 14, formal procedures for conducting systematic literature reviews and meta-analyses have been suggested and widely adopted. Systematic reviews and meta-analyses are the reference standard for synthesizing evidence in health, education, and other related fields for their methodological rigor. The systematic review and meta-analysis derives strength from their articulation of a clear, transparent methodology, including the search process and analyses which are to be described with replicable precision (Maggin, Talbott, Van Acker, & Kumm, 2017). A systematic review is defined as the attempt to make the research summarizing process explicit and systematic to ensure the author's

assumptions, procedures, evidence, and conclusions are transparent (Lipsey & Wilson, 2001). The meta-analysis refers to the explicit use of statistical methods to synthesize results from a series of independent research studies and derive a pooled estimate across studies (e.g., weighted average effect; Borenstein, Hedges, Higgins, & Rothstein, 2009; see Chapter 14). All meta-analyses should include systematic review, but not all systematic reviews use meta-analytic procedures to describe outcomes. The systematic review has become critical for identifying and summarizing the evidence base of educational practices (Maggin et al., 2017; Wendt & Miller, 2012).

In response to suboptimal reporting of meta-analyses in leading health journals (Mulrow, 1987; Sacks, Reitman, Pagano, & Kupelnick, 1996), an international group developed guidelines for reporting meta-analyses called the QUOROM Statement (Quality Of Reporting Of Meta-analyses; Moher et al., 1999). In 2009, the guidelines were updated to address recent advances in the science of systematic reviews more broadly, and was renamed PRISMA (Preferred Reporting Items of Systematic reviews and Meta-Analyses; Moher et al., 2009). The goal of PRISMA is to help authors improve reporting across systematic reviews and meta-analyses, with a focus on evaluations of interventions (Moher et al., 2009).

The PRISMA statement consists of a 27-item checklist and a four-phase flow diagram (see Figure 3.1) to document each step of the systematic process, including (a) *identifying* potential articles, (b) *screening* articles for possible inclusion, (c) assessing *eligibility* of potential articles, and (d) *including* articles for further analysis. In addition to the PRISMA statement, the PRISMA Checklist specifies what items should be included when reporting a systematic review or meta-analysis specific to each section of the manuscript (i.e., title, abstract, introduction, method, results, discussion, funding). Downloadable documents for researchers to reuse, including the PRISMA flow diagram (e.g., found at: doi:10.1371/journal.pmed.1000097.s001) and checklist for reporting (e.g., found at: doi:10.1371/journal. pmed.1000097.s002) are under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction, provided the original author and source are credited.

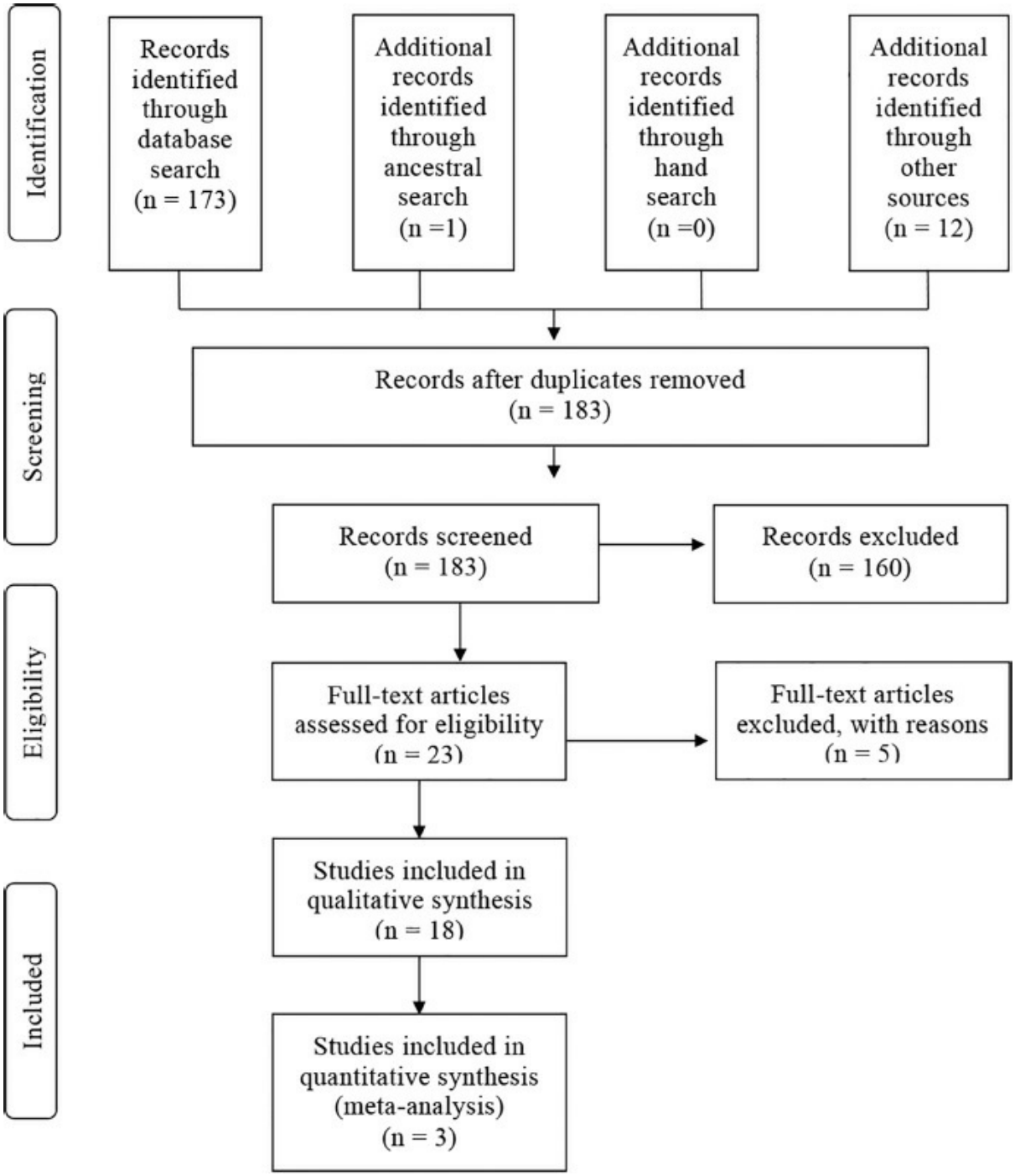


Figure 3.1 PRISMA diagram.

Adapted from: Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Med*, *6*(7), e1000097. doi:10.1371/journal.pmed1000097

PRISMA for Flow Diagram

The flow of information in each phase of a systematic review or meta-analysis should be presented in a flow diagram specifying the process for selecting studies (i.e., screening,

eligibility, inclusion in systematic review, and if applicable, inclusion in meta-analysis). Provision of information on the reliability of the screening process (e.g., interrater agreement) should be reported. When reporting individual participant data, additional information should be reported to specify availability and analyses of both individual and group level data (Stewart et al., 2015). We encourage interested readers to read the PRISMA statement and related articles prior to engaging in a rigorous systematic review or meta-analysis. In addition, we encourage you to read several examples of recent reviews incorporating these PRISMA features (e.g., Royer, Lane, Cantwell, & Messenger, 2017).

Research Questions

This section focuses on issues related to starting new studies; specifically, how you can find topics to study and translate them into research questions. The purpose, elements, and forms of research questions are discussed. Finally, various types of research questions are described.

Finding Research Topics and Research Questions

Using the Literature

As noted, research is often reviewed to build a rationale for a study and to identify a new study or series of studies to conduct. From a literature review, specific study ideas or researchable questions may arise from the limitations of the existing research. For example, if a procedure for reducing challenging behaviors has been studied only at one level of intensity (e.g., only for parts of the school day), then a gap exists in our knowledge about how the procedure would work if it was used at other intensity levels (e.g., all day). Conducting studies on those other intensity levels would help reduce the limitations in the existing research.

Sometimes research questions come from a single study rather than a review of several studies. The discussion sections of studies often note avenues for future research, which are timely and important areas of future research. The method sections of completed research also identify the conditions under which procedures have been studied. Systematic replication of completed studies (see Chapter 4) is an ideal way to find new studies. In systematic replications, the investigator attempts to use participants, settings, procedures, measures, and designs similar to those used in the original study. Research questions for the replication focus on a variation of one aspect of the study while attempting to keep other elements the same as the original study. For example, Filla, Wolery, and Anthony (1999) studied the effects of an environmental arrangement and adult prompts to increase the frequency of preschoolers' conversations with classmates during play. Cuneo (2007) replicated the study by using similar procedures but measured the content and frequency of what children said to one another.

Other Sources

Not all research questions come from the literature; in fact, "too 'slavish' a reliance on archival literatures as a source of experimental questions carries its risks" (Johnston & Pennypacker, 1993, p. 40). Those risks include not looking afresh at phenomena and potential relations in nature, not considering alternative explanations for behavior, and minimizing the chances of discovering interesting relations.

Several alternatives exist to relying solely on the literature, and one is observing individuals similar to your potential participants in their usual environments. There is no substitute for knowing one's organism and its behavior in context; and there is no better way to acquire such knowledge than spending attentive time watching and interacting with them in their usual settings. If your potential participants are students with disabilities, then spending time watching them at home, in their classrooms, on the playground, and in their communities is critical for planning studies of their behavior. Careful observation often generates questions about potential relations between environmental events and structures and behavior. Those questions often can be converted into good research questions.

Other alternatives for research questions arise from day-to-day practice, and the problems and issues encountered. For example, when constant time delay was initially studied with chained behaviors, the studies involved delivering the instruction in a 1:1 arrangement (one teacher, one child). A teacher, Ann Griffen, who participated in earlier studies, suggested the 1:1 arrangement was impractical in classrooms. Based on her suggestion, a study was conducted evaluating an arrangement where one child was taught while two others observed. Fortunately, the two observing children learned by observation what was being taught directly to one child (Griffen, Wolery, & Schuster, 1992). Thus, a practical problem was solved, and a more efficient instructional arrangement was found effective. In some cases, a procedure appears to be working well, and a study is initiated to evaluate its effectiveness. Jolivette, Wehby, Canale, and Massey (2001) conducted a classroom study where the teacher was providing students with choices about the order with which three required assignments were completed. This appeared to be resulting in high levels of task engagement and low levels of disruptive behavior. Systematic data collection and use of choice *versus* no choice condition across children confirmed the choice of the order in which assignments were completed was indeed related to higher levels of task engagement and lower disruptive behavior.

Some research questions come from principles derived from research. For example, the Premack principle states that if high probability behaviors (frequently occurring behaviors) follow low probability behaviors (behaviors that do not occur often), then the frequency of the low probability behaviors will increase (Premack, 1959). Many children with disabilities frequently emit stereotypic behaviors that are repetitive behaviors (rocking, hand flapping, flipping things in front of the eyes) that appear to have few social consequences. These stereotypic behaviors often are high probability behaviors. Thus, for children for whom other reinforcers are not readily available or deliverable, investigators asked: If stereotypic behaviors are high probability behaviors, will their contingent use result in increases in correct responding (Wolery, Kirk, & Gast, 1985) and decreases for challenging behaviors when used in a differential reinforcement of other behaviors contingency (Charlop, Kurtz, & Casey, 1990; Charlop-Christy & Haymes, 1996). In these cases, as predicted, the stereotypic behaviors (high probability behaviors) functioned as positive reinforcers for low probability behaviors (Charlop et al., 1990;

Wolery et al., 1985).

Other Factors in Selecting Research Questions

The literature, practice issues or problems, and principles of behavior may serve as sources of research questions. Two additional factors should be considered; these are the investigator's interest in the question, and the feasibility or practicality of conducting the proposed study. Being interested in the research questions and study are not prerequisites for conducting studies with rigor, but it is much better when investigators are interested in what they are studying.

Unlike interest, assessing the feasibility of doing a study is a critical step in deciding whether to attempt to answer specific research questions. Little is as disheartening as planning and starting a well conceptualized and important study only to realize it is not feasible. Evaluating feasibility varies by study, but some common issues can be proposed. You should consider whether an adequate number of appropriate participants are available. Also, you must ensure you have access to sites where potential participants are educated, treated, or reside. Other critical feasibility issues focus on data collection. Most SCD studies require repeated direct observation over many days, weeks, and months; having adequate observer times which can match the schedule of the study are critical issues. Also interobserver agreement assessment should occur often (e.g., a minimum of 20% of the sessions for each participant and condition), thus, additional observer time is needed for this data collection. When data are collected using video cameras, personal assistance devices, or hand-held computers, additional expense and resources are required. Of course, when data are not collected from live observations, then time is needed for coding the data. Because decisions are made as the study progresses (e.g., are the data stable, when should the conditions be changed), data must be summarized and graphed on a daily or regular basis, and evaluating whether adequate time exists for coding and analysis must be considered. You must also consider whether the implementers of the experimental conditions can carry out the procedures with sufficient precision and integrity, what training they need, how much time it would take, and do demands exist which preclude their implementation of your procedures. You also must judge whether adequate time exists to complete the study before inalterable deadlines such as the close of school or student graduation occur. Using the past literature is one way to judge this issue, but certain events outside the investigator's control are likely to occur, such as, participant illness, scheduled holidays and breaks, snow days, and so forth. Thus, it is wise to start studies when you have 1.5 or 2 times the amount you think you will need.

Stating Research Questions

By tradition, single-case experimental investigators ask experimental questions rather

than state and test hypotheses (Kennedy, 2005). This is done in part, because research is viewed as a way to build an explanation of why behavior occurs as it does in nature rather than evaluating a theory. The function of research questions is to focus the investigator on the purpose or goal of the study. The research question provides broad boundaries for making decisions about which measures, procedures, participants, and designs are relevant. For example, if the research question focuses on whether a given environmental manipulation (teaching procedure) will result in particular participants learning a specific skill (behavior), then the measures selected must address that skill directly. Similarly, it would eliminate participants who did not meet the inclusion criteria. Finally, certain research designs would be more appropriate for evaluating teaching procedures than other research designs. Thus, the research question focuses the investigator on key elements of the study plan.

Stating the research question in the written proposal for a study, or in a report of a study after it is completed, functions to orient readers' attention to the purpose and nature of the study. The introduction of such reports should build the rationale for the study, and the last paragraph should state the purpose of the study and include the research questions. This allows readers to know what the investigator wants to learn, and thus, allows readers to evaluate independently whether the study method actually will allow a defensible answer to the questions.

Research questions have three elements: participants, independent variable, and dependent measures (Kennedy, 2005). All research questions should have these three elements, but the order in which they are included can vary. For example, the same research question could take the following forms: (a) Does "X" independent variable influence "Y" dependent variable for "Z" participants? (b) For "Z" participants, does "X" independent variable influence "Y" dependent variable? Or (c) Is "Y" dependent variable of "Z" participants influenced by using "X" independent variable? The three elements are present in each question, but the sentence structure is different.

Research questions should be stated in a directional (falsifiable) form. For example, rather than asking if a self-monitoring intervention results in changes in challenging behavior for young children with autism, researchers should ask whether a self-monitoring intervention results in *decreased* challenging behavior. This is consistent with guidelines for writing research questions using other research approaches (e.g., group design), allows the reader to easily identify author expectations, and provides a means for determining direction of predicted effect for the purposes of evaluating outcomes. However, it is by no means universally accepted. For example, Johnston and Pennypacker specified that asking neutral questions about nature may introduce fewer biases into the research enterprise than stating the likely outcome (a hypothesis) before the study is initiated (1993). To guard against this, think of your task as *answering a question* rather than proving the effectiveness of some intervention.

Studies may have more than one research question. For example, when conducting treatment- outcome studies, often you will see (a) one question regarding the degree to which the introduction of the intervention results in increases or decreases in student

performance on a specific variable, (b) one question regarding treatment integrity (to examine the extent to which the intervention was able to be implemented as designed), and (c) one question related to social validity (to examine stakeholders' views on the social significance of the intervention goals, social acceptability of the intervention procedures, and social importance of the effects; Wolf, 1978). When multiple questions are used, their number should be limited (e.g., to three or four) to ensure the study is manageable. Often multiple dependent measures are involved, and in such cases, a separate question can be asked for each dependent measure. To assist the reader, it is wise to use a parallel sentence structure for each question. For example, a teacher of fifth-grade students may be interested in the effects of assigned seating or students' choosing where they sit on their problem behavior, accuracy of their assigned work, and on-task behavior. Rather than grouping these different measures into one question, parallel questions could be asked for each behavior of interest:

- Does teacher assignment of seating increase the number of disruptive classroom behaviors for fifth-grade students in comparison to student-choice seating?
- Does teacher assignment of seating increase the percentage of correctly completed mathematics problems for fifth-grade students in comparison to student-choice seating?
- Does teacher assignment of seating increase the percentage of intervals in on-task behaviors for fifth-grade students in comparison to student-choice seating?

By asking these three questions, the investigator and reader can focus on whether the method section will allow each to be answered. The questions provide a focus for planning the study, but do not prescribe how each element of the study will occur. For example, the above questions do not specify the duration of the observations, the definitions of the behaviors, which recording systems will be used, how students will be told which condition is in effect, what content will be taught, what students' tasks will be, how the teacher will make the seating assignments, or how seats will be arranged in the classroom.

In practice, we think about the teaching procedures, interventions, and treatments (independent variables) we use to help individuals learn and function better in their lives. These procedures and interventions are what practitioners can do to help them change. Our questions about those independent variables are: Does this procedure work? Does more or less of this procedure work better? Does it work better with some or all of its parts? Does it work better than some other procedure? Implicit in each of these questions are the behaviors we want to change. Four types of research questions are useful (Kennedy, 2005), and they parallel the questions we have about procedures and treatments. Examples of these four types of questions (demonstration, parametric, component, and comparative) are shown in Table 3.1.

A common type of question is a **demonstration question** (Does it work?); they follow the form: "What relations exist between an independent variable and a behavior for a

given set of participants?" Demonstration questions are straightforward; they ask about whether and how use of an independent variable changes our participants' behaviors. For example: Does use of a teaching procedure increase children's percentage of correct responses on mathematics worksheets? Or, does use of a given reinforcement schedule for other behaviors reduce the rate of aggressive behavior on the playground? These are important questions because they ask whether particular environmental arrangements or events are related to patterns in participants' behaviors.

Another type is a **parametric question** (Does more or less of this procedure work better?). Parametric questions focus on the amount of the independent variable and the effects of those various amounts on behavior. They follow the form of: "What relations exist between one level of the independent variable and another level of the independent variable on a given behavior for specific participants?" Examples might be: Do 30 minute speech therapy sessions each week result in fewer articulation errors by 8-year-old children with articulation disorders when compared with 50-minute sessions? Or: Does self-monitoring and self-reinforcement every 10 minutes result in faster reduction of disruptive behavior compared to every 20 minutes? These are important questions because they help us understand how much of a procedure or treatment must be used to get the desired effects for given participants.

Table 3.1 Sample Research Questions.

Question Type Form *Examples*
Demonstration Questions
Does independent variable result in increases/decreases in levels of behavior for given p
Does point-of-view video modeling result in increase in the percentage of intervals of pre preschoolers with autism?
Does contingent observation during free-play time reduce the number of aggressive acts in inclusive classrooms?
Parametric Questions
Does one level of the independent variable result in increase/decrease in levels of a beha level of that independent variable, for given participants?
Does conducting 8-trial-per-stimulus constant time delay sessions result in reduced trial language delays, when compared with 3-trial sessions?
Does daily feedback increase the number of attempted social initiations during recess by internalizing behavior problems, when compared with weekly feedback?
Component Analysis Questions
Does the independent variable with a given component result in increased/decreased lev participants?
Does contingent observation with reinforcement for appropriate behavior result in a mor acts by young children with behavior disorders than contingent observation alone?
Does self-monitoring with reinforcement for work completion result in higher rates of w

students with learning disabilities, compared with self-monitoring alone?

Comparative Questions

Does one independent variable result in increased/decreased levels of behavior for given different independent variable?

Does the system of least prompts result in faster acquisition of numeral naming (i.e., ste with intellectual impairments, when compared with video modeling?

Do peer-based strategies result in higher levels of self-determination skills for high schoo disorders, when compared with teacher-led strategies?

The third type of research question is a **component analysis question** (Does it work better with some or all of its parts?). Component analysis questions are an acknowledgement that many of our procedures and interventions have many different parts (components); in short, many interventions are treatment packages. A study can demonstrate (using a demonstration question) that a given package is related to consistent shifts in participants' behaviors. The question then becomes are all of the components of the package necessary; these questions follow the form of: "What relations exist when a teaching procedure is used with or without a given component on a specific behavior of given participants?" An alternative is to study whether a given package becomes more or less effective when another component is added. For example, we might ask, Does participation in a check-in-check-out procedure implemented in conjunction with a school-wide primary prevention positive behavior support plan influence work completion of middle-school students initially identified as being non-responsive to the primary positive behavior support plan? These types of questions are important, because they allow us to take apart treatment packages as well as build treatment packages.

The final type of research question is a **comparative question** (Does one procedure work better than another procedure?). Although there are many pitfalls in such research questions (Johnston, 1988), there are many times when these questions are highly useful (*see* Chapter 11). The questions take the form of: "Does one teaching procedure result in more rapid learning of a specific behavior for given participants versus another teaching procedure?" These types of questions are useful for making recommendations about which practices should be used, and about which of a couple procedures is more efficient under given situations.

Summary

Research questions function to help investigators articulate the focus of a study and to set broad boundaries for making study decisions. Research questions have three elements (participants, independent variable, and dependent variable), and the syntax used in writing them is flexible. Four common types of research questions are: demonstration questions, parametric questions, component analysis questions, and comparative questions.

Writing a Research Proposal

Research proposals are written to communicate to others your plans about conducting a study. Research proposals are often submitted to students' research committees to evaluate whether the studies are worthy of being a master's degree thesis or a doctoral dissertation, and to determine what changes are needed in the plans to increase the chances of successful completion of quality studies. Grant proposals submitted to funding agencies also often contain research proposals. Emphasis is placed below on writing research proposals for thesis and dissertation studies. Research proposals have the following sections: an abstract, introduction, and method.

Abstract

The abstract is a challenging writing task, because a lot of information is summarized into a few sentences. Different journals have varying length requirements, but it ranges from 120–250 words; check in the journal to which you wish to submit your paper. It is wise to read the abstracts of several other studies to see how they are structured. The information needed is (a) a sentence about the general topic or purpose of the study; (b) an overview of the method, particularly the participants, measured behaviors, setting, and type of design; and (c) a statement about the potential implications of the study. In general, abstracts should not replace full articles in informing and guiding readers, but for time-pressed readers and those with limited access to the full texts reports (e.g., because of pay wall, low internet download capacity), the abstract must stand alone in presenting a clear and concise account of the paper (Beller et al., 2013).

Introduction

The writer has three tasks in the **introduction section**. The first is to introduce the topic to the readers; this is usually done in the first paragraph, which starts with a general statement. The second task is to provide a summary of existing literature while building a rationale for the study. This task comprises the major portion of the introduction. Several models for organizing the literature summary and stating the rationale are presented in Table 3.2. Other models can be used, but these are common. Finally, the last paragraph of the introduction states the purpose and lists the research questions. The length of the introduction will vary greatly across universities and committees; generally, four to eight pages are sufficient.

Table 3.2 Various Models for Writing the Introduction to a Research Proposal.

Model 1—Accumulating Evidence Approach
A. General statement about the topic/issue

B. Series of referenced statements about the current knowledge of the topic
C. Statement about gap in knowledge—leading to the rationale for the proposed study
D. Purpose statement and research questions

Model 2—Contrasting Options Approach
A. General statement about the topic/issue
B. Description of one alternative or option
C. Description of the other alternative or option
D. Contrasts of the two alternatives or options
E. Rationale for current study
F. Purpose statement and research questions

Model 3—Historical Perspectives Approach
A. General statement about the topic/issue
B. Description of the emergence of a body of evidence about the topic
C. Identification of the next steps in extending the body of evidence
D. Rationale or justification for studying the particular step
E. Purpose statement and research questions

Model 4—Deficit in Knowledge About a Practice Approach
A. General statement about the practice
B. Description of the practice with discussion of supporting research
C. Identification of consistent weakness in the studies supporting the practice and statement about why this weakness is problematic
D. Purpose statement and research questions

Model 5—Discrepant Knowledge Approach
A. General statement about the topic/issue
B. Description of one body of knowledge about the topic/issue
C. Description of the second (discrepant) body of knowledge about the topic/issue
D. Potential resolution of the discrepant body of knowledge—leading to the rationale for the study
E. Purpose statement and research questions

Model 6—Expanded Application Approach
A. General statement about the topic/issue
B. Description of previous applications of the approach with the effects
C. Rationale for expanding the application (e.g., to a new population, new context, new skill, new implementers)
D. Purpose statement and research questions

Method

The **method section** is the main body of research proposals; it is a detailed plan of the study being proposed. Because the study has not yet been conducted, it should be written in the future tense. This section should have the same sections as commonly found in articles describing research studies in journals: participants, setting, materials, response definitions and measurement procedures, experimental design—with a description of each experimental condition, and data analytic plan. Other common sections include interobserver agreement assessment, procedural fidelity assessment procedures, and

social validity procedures. These sections are described below. Although the participants, setting, and materials sections usually appear in that order at the beginning of the method section, the order of other sections may vary across studies. The method section should include a detailed operational description of the procedures. As Baer et al. (1968) indicated, this description should be sufficiently detailed and precise so that a trained individual could read the proposal and then implement the study as you intended without additional guidance. In general, a research proposal should be more detailed than a method section in a published article.

Participants

Almost always when proposals are written, your participants are not known. You will know, however, how many you need. It is wise to start with more participants than the minimum required by your design. This section may include adult as well as student participants. Generally, three types of information are needed about participants (Wolery & Ezell, 1993). First, it should include the demographic characteristics such as gender, age, diagnoses, race, ethnicity, and socioeconomic status. This should also include the types and intensity of services they are receiving. If you do not know who the participants are, your proposal should state whether each of these will be used as inclusion criteria, and in some cases give ranges (e.g., ages) or options (e.g., children with behavior disorder or learning disabilities). These characteristics are not the basis for making generalizations from SCD studies (Birnbrauer, 1981), but they should be reported for archival purposes.

Second, you should identify the measures (e.g., tests) used to describe your participants' academic and functional performance. These measures may not be used to make decisions about including or excluding participants, but will be used to present a description of who participated. The full name and APA citation should be included when using published tests or scales. For young children and individuals with significant disabilities, specific tests and scales may be less relevant than detailed descriptions of each participant from repeated observations and from interviews with their parents and other adults who know them well.

Third, you need to identify the inclusion and exclusion criteria and how those criteria will be measured. While age, diagnosis, and occasionally other demographic categories will be used, the more important criteria are functionally based. Specifically, these include how participants behave under the baseline conditions of the study, what events are related to the behaviors of interest, and whether participants have specific behaviors needed to respond successfully to the independent variable or to acquire the behaviors in the dependent variable (Birnbrauer, 1981; Lane, Little, Redding-Rhodes, Phillips, & Welsh, 2007; Wolery & Ezell, 1993). If reinforcers are used in the independent variable, then you should describe how reinforcers will be identified. Similarly, if you use a functional behavioral assessment to select participants whose challenging behaviors are maintained by specific factors, then you should state this in this section. The procedures

used in the assessment should be described later in the method; but the criteria should be stated here.

Setting

The setting section should describe the location of all experimental procedures and conditions that are planned, including where primary, secondary, and generalization measures are collected; where assessment procedures were conducted; and where the independent variable was implemented. The description of these locations should include their physical dimensions—reported in metric units; and how those spaces are arranged. The description should identify in the larger space where the experimental activities will occur (e.g., on the floor, at a table). An important aspect of the arrangement describes the relative location of the implementer to the participants (e.g., seated across a table from one another). If quality measures (e.g., rating scales) of those settings are available, then these should be reported. For example, the *Assessment of Practices in Early Elementary Classrooms* (Hemmeter, Maxwell, Ault, & Schuster, 2001) is a measure of the quality of kindergarten through third-grade classrooms. If an investigator used such a measure and reported the results, then readers would have a summary of the quality of the classrooms in which a study occurred. Likewise, if a study was conducted in a school that was listed as failing by the state, then information is provided about the context in which the study occurred.

Procedures

This section should include a detailed description of study procedures, beginning with how university and district approvals (or other institutional approvals) are to be secured. This information should include a detailed step-by-step discussion of how the study will take place: How will practitioners or agencies be contacted? How will implementers and parents provide consent, and how will student assent occur? When writing this section it is important to note the organizational structure can vary, but the level of detail must be sufficient to ensure anyone could read the proposal and know exactly how the study procedures will take place. Some student committees will require a description of how the proposed design addresses threats to internal validity.

Each experimental condition should be described as a sub-section. This would include the baseline condition and treatment conditions. Sometimes, especially in comparative studies, a general procedures section is included that describes procedures to be used across all conditions. The description of the baseline condition, called probe condition in some designs, should include a description of the procedures used and the parameters (quantification) of those procedures (Lane, Wolery, Reichow, & Rogers, 2007). In Table 3.3, a number of dimensions of baseline procedures are listed. These dimensions might not be included in each study, but many are relevant for most studies. Some of these

dimensions may be addressed in other sections of the method. The intervention conditions should include a description of the independent variable. As with other procedures, the procedures used to implement the independent variable must be described, as should the parameters of those procedures. Ideally, the only difference between the baseline and intervention conditions is the independent variable or the level at which the independent variable is used (Messenger et al., 2017). However, if other factors are different across the conditions, these should be specified. The goals of this section are to describe in detail who will do what to whom (Wolery, Dunlap, & Ledford, 2011). When a condition is repeated, include a description of how the second and subsequent use is similar to or identical to the first time the condition is used.

Materials

This section should include a description of the materials, supplies, and equipment used. This includes the materials participants use during experimental sessions and observations. When published curricula are used, the citation should be included. Any equipment used to collect the data (e.g., video camera) should be reported by name, make, and model. When trademarked materials are used, include the trademark symbol (™). This section may include the criteria and rules used for selecting individualized materials.

Response Definitions and Measurement Procedures

This section is a complete description of the dependent variable. Each behavior being measured in the study should be defined in this section. The definitions should not be generic descriptions of the construct, but should be the definitions that will be used during the observations or the coding of video records. Also, any rules to be used in recording the data should be reported. The type of recording system to be used during the observations should be identified and described operationally, specific enough that someone else could replicate these exact procedures. This section also should identify how long each observation will be; how often observations will occur; and if more than one observation occurs within a day, when and how much time will occur between observations. It is acceptable and desirable to use definitions and measurement procedures used in other similar research. When appropriate, citations of the published literature should be included. When data collection forms are used, these should be included in an appendix of the proposal.

In this section and sometimes in a section labeled reliability, plans for training observers and documenting how they are trained should be described. Finally, this section should describe the interobserver agreement assessment procedures. You should indicate how often interobserver agreement will be assessed, what formula will be used to calculate the agreement estimates, and what levels of agreement will be considered

acceptable.

Procedural Fidelity

This section should describe how the implementation of the procedures will be measured (Billingsley, White, & Munson, 1980; see Chapter 6). This section also should report how often these data will be collected and how they will be calculated.

Social Validity

Wolf (1978) described the importance of assessing social validity of studies. This section should describe how the social validity of the study will be assessed, including what aspects of social validity (goals, procedures, and effects) will be assessed, when the assessment will occur, and procedures used in that assessment. Finally, the consumers who will judge the social validity of the study should be identified.

Experimental Design

This section usually includes a paragraph describing the experimental design you will use. Often a citation to the design is included. However, it is important to describe the manner in which the design will be implemented in your study. We encourage you to also include the rationale for selecting the proposed experimental design.

Data Analytic Plan

In research proposals, the data analysis plan should be presented at the end of the method. This section should have two sub-sections: formative evaluation, and summative evaluation. The formative evaluation section should describe (a) how often interobserver agreement will be assessed, (b) what levels will be considered acceptable, (c) what actions will be taken if the agreement estimates are unacceptable, (d) how often procedural fidelity assessments will occur, (e) what levels of procedural fidelity will be considered acceptable, (f) what actions will be taken if the procedural fidelity data are too low, (g) how the data on the primary dependent variables will be graphed, and (h) how the graphed data will be analyzed to make decisions about changing experimental conditions. The summative evaluation section should describe (a) how the interobserver agreement data will be summarized and presented in your final report; (b) how the procedural fidelity data will be summarized and presented in your final report; (c) how the dependent measure data will be summarized and presented, and (d) what rules you will use to make a judgment that a functional relation exists. Sample graphic displays and tables are recommended to ensure reviewers (e.g., committee members) are clear on the expected data to be presented and how they will be analyzed.

Table 3.3 Description of Factors to Be Described in Baseline Conditions.

Question
• Dimensions
• Who did what to whom?
• Individuals other than participants
• How many were present
• Role in study
• Preparation relative to those roles
• Relationship to the participant before the study
• Participant's familiarity with those individuals
• Any unique factors relative to their involvement
Procedures
• Behaviors of the person(s) conducting the experimental procedures/sessions
• Contingencies in effect for the studied behaviors
• The consistency with which the contingencies were delivered
• Activities, tasks, materials, or curriculum used
• Familiarity or history of the participant with the routines and procedures
Participants
• Familiarity with the baseline procedures
• Familiarity with the persons conducting the study
• Participant's peers also experienced the procedures of the baseline condition or whether it was applied only to the participant
Where were those actions taken?
• Size of the setting
• The arrangement of the equipment and materials of the settings
• Participant's location compared to the person(s) implementing the study
• The setting(s) in which ancillary measures were taken
• Qualitative ratings of the setting
• The participant's familiarity with the setting
When did those actions occur?
• Frequency and consistency with which observations occurred
• When within the day did observations occur
• Duration of each observation

Writing a Final Report

The final report of a study can take several forms, such as a master's thesis, doctoral dissertation, and a **manuscript** (report for submission to a journal for review and possible publication). These reports usually include five sections (abstract, introduction, method, results, and discussion), although some thesis and dissertation reports have university-unique sections.

Abstract

The final report abstract should be similar to the proposal abstract. However, you should add information about actual results, rather than expected results, and should identify any useful implications of the study, rather than potential implications.

Introduction

The introduction of a final report is similar to the introduction for research proposals. However, while you are conducting your study, you should be reading the literature to identify new articles bearing on your study. When new articles appear, these should be integrated into your introduction.

Method

The method section of a final report should include an exact, precise, detailed, operational description of what you did in your study. The sections are identical to those for the research proposal, but it should be written in the past as compared to future tense. Nonetheless, writing the method of a final report is more than simply changing the tense of the research proposal. It involves ensuring you provide an accurate, thorough description of what occurred and present enough information to allow another to replicate your study. The following additions or modifications to the methods section may be required:

1. **Participants**: A description of who actually participated. This would include the same content (demographic information, abilities, and inclusion criteria), but it is a report of those who were involved. Often including this information in a table will conserve space and make accessing the information easier for readers. The setting and material sections should also be changed to describe the actual locations and materials.
2. **Response Definitions**. Similar information is provided regarding response

definitions and data collection procedures. Sometimes in the process of conducting a study, the proposed definitions of the behaviors will change slightly as observers are trained and data collection is initiated. Thus, ensure the definitions in the final report represent the definitions used in the actual data collection. This is true of the measurement procedures as well.

3. **Procedures**: A description of what actually occurred, including any intervention modifications or unexpected changes. Additional content related to the details of securing approvals and the consenting and assenting activities should be added, including the number of children and adults from whom consent and assent was requested and obtained.
4. **Reliability**. When reporting procedural fidelity and IOA measurement and results, information should include a listing of the specific variables measured, the procedures for measuring them, the frequency of measurement by participant and condition, formula for calculating it, and the results of the measurement by participant and condition. When describing how treatment agents were taught to conduct the intervention activities, information regarding these training activities will be provided (e.g., How many sessions, and for what duration? Were there checks for understanding? Modeling? Coaching?).
5. **Experimental design**. As in other sections, this section should be revised to describe what actually occurred, including any revisions to design type.

Results

The **results section** is often framed using the research questions. Sometimes, this section starts with a description of the results of the interobserver and procedural fidelity assessments if that information is not presented in the method section. However, often authors describe the results of the interobserver agreement data in the method section, and others put it in the results section.

The purpose of the results section is to describe how the data paths changed, or did not change, with the experimental manipulations. This section is *not* the place to describe the meaning of the findings, to speculate about what influenced the behavior, to suggest other research, or to draw implications for practice. This is the place to describe the patterns in the data. Two major ways exist for organizing the results: by dependent measure, and by participant. The former is preferred particularly when the data are consistent across participants. If the participants required multiple modifications to the procedures before change reliably occurred, then organizing the results by participant may be used.

The results section almost always contains figures depicting participants' data across experimental conditions. The narrative should include a general description of the data, and comments about data stability, systematic changes in the data (e.g., changes in stability, level, trend), the extent to which changes co-occurred with the experimental conditions, and potentially the ranges of performance within and across conditions.

Although some authors report means and ranges for each condition, this should be avoided unless the data are neither accelerating nor decelerating within and across conditions and only changes in level are evident. The description should be about the patterns existing within the data and the shifts in those patterns that occurred or did not occur when the experimental conditions changed. Thus, describing the baseline data (e.g., high, low, accelerating, decelerating, variable, stable) and describing how the data changed when the intervention was introduced are critical elements. This style of describing the patterns in the data across time as conditions changed is the type of description appropriate for SCDs. Unusual events (e.g., extended absences, changes in implementer) should be described and may be represented on figures also.

Discussion

The purpose of the **discussion section** is to describe the relevance of the study's data; it can be a relatively brief section (e.g., three to five pages). The cardinal rule is: "Say no more than the data permit" (Tawney & Gast, 1984, p. 364). The first paragraph should restate the purpose of the study and note the major findings from it. After this paragraph, you should accomplish three things in the discussion section: (a) describe your findings by tying them to the existing research—much of which you will have cited in the introduction section; (b) identify areas for future research; and (c) note the limitations and qualifications of the study. In most cases, you may also want to draw implications for practice.

There are a couple ways to organize the Discussion section. First, if several findings exist, then each finding can be discussed sequentially. The similarities and differences with previous research should be addressed and referenced. It is important to connect your findings to the literature, citing how findings from your study converge or diverge with existing studies. During this discussion, you may point to future research, particularly modifications of your procedures. You also can describe how each finding is limited or qualified by the manner in which you conducted your study. The second way to organize this section is to have separate sections discussing these elements (findings and relevance to literature, future research, limitations, and implications for practice). You may be hesitant to note the limitations of your study; however, this is part of being skeptical about one's own work and it is a legitimate scientific behavior. Scientific knowledge is necessarily conditional—functional relations exist in the contexts of studies in which they are identified. All studies are limited in many ways; thus, acknowledging those limits is not a sign of weakness but evidence you are being objective in dealing with the realities of your study. Further, addressing your limitations and providing recommendations for future studies can help shape future investigations.

Disseminating Research

Once you have completed your study or literature review, the next step is to disseminate what you have learned. In the social sciences, **dissemination** generally refers to the publication of a research report in a peer-reviewed journal, or the presentation of results at a professional conference. Other avenues for dissemination include practitioner conferences or meetings (e.g., sharing your research with a local school system), practitioner publications, and (more recently), websites, podcasts, and other electronic media. Not every study or review should be disseminated, but if new information is learned about a functional relation, then dissemination is appropriate.

Accurate dissemination of your work is important for two reasons. Sharing your findings with the research and teaching communities may inform future studies and shape educational practices. As illustrated above, reading previous investigations will help researchers (a) conduct more rigorous studies addressing limitations recognized in earlier works and (b) extend the knowledge base by addressing unanswered questions. For example, constant time delay had been shown to be effective in teaching preschoolers many skills in multiple studies (e.g., Alig-Cybriwsky, Wolery, & Gast, 1990; Doyle, Wolery, Gast, Ault, & Wiley, 1990). However, in each study, the procedure's use was monitored carefully and occurred at high degrees of accuracy. The question became, would teachers actually implement it that way. As a result, two levels of procedural accuracy (i.e., high and low) were compared experimentally (Holcombe, Wolery, & Snyder, 1994). The data showed correct implementation was indeed important for nearly all children. Similarly, forward-thinking practitioners have the ability to glean information on "what works" (or does not work), which can then be used to improve practices. For example, a teacher working in an inclusive classroom who is struggling to support students with emotional and behavioral disorders could benefit from reading about function-based interventions conducted with similar students being educated in a similar setting (Lane, Weisenbach, Little, Phillips, & Wehby, 2006). Thus, research findings should be shared to benefit others and help influence research and practice. Many methods exist for distributing your information; some are informal (e.g., conversations with your colleagues or students; information sharing sessions with practitioners and parents) and others are more formal. Some formal venues include poster presentations, conference seminar presentations, web-based publications, and refereed journal articles. This section describes each of these. However, first, an important issue of all dissemination activities is discussed: deciding on authorship and order of authorship.

Deciding Authorship

The issue of authorship is important because it carries implications for scientific contribution and productivity, which potentially influence hiring, promotion, and tenure decisions. Moreover, authorship and the order of authorship can be conceptualized as expressions of intellectual property rights—individuals who contribute substantially have a right, because of ownership, to be included as authors. Authorship can be a point of contention between contributors if clear guidelines are not established to determine who should be an author and the order in which the names are listed. This section describes: (a) inappropriate practices, (b) assumptions, and (c) general guidelines.

Inappropriate Practices

Two practices violate ethical guidelines with respect to publication credit: under-inclusion (fraud) and over-inclusion. Under-inclusion or fraud refers to omitting individuals as authors who have contributed substantially to the research or product. In brief, this can be thought of as not giving sufficient credit when due. Over-inclusion refers to including individuals as authors who did not make a substantial contribution to the research or product. In brief, this can be thought of as giving undue credit.

Assumptions

Individuals should be included as an author if they made a substantial contribution to the study or product. These can be defined as *intellectual*—providing conceptualizations and making decisions related to the design, implementation, and/or analysis; *material*—providing funding, space, and/or resources to the research or activity; *operational*—conducting the research or activity under the guidance and supervision of another person; and *descriptive*—writing the actual product or parts of the product. In Table 3.4, we suggest operational definitions of contributions that may and may not meet threshold for substantial contributions. Combinations of the behaviors specified under the category of "not a substantial contribution" from a single individual may constitute a meaningful contribution.

Once the decision to include an individual as an author is made, the next step is to determine the order of authorship. In other words, who is first, second, or third author? The order of authorship should be guided by the amount of contribution to the research, activity, or product. The person who contributes the most should be placed in the first author position, and those with the least (but still substantial) contribution should be placed in the last author position. The statements in Table 3.5 are provided as suggestions in making this decision. Regardless of position, it is essential for all authors to consent to being an author. Everyone has the right to decline authorship for whatever reason, despite their contribution to the research, activity, or product. Individuals who do not contribute substantially to the research or activity, but who provide support for it, should be acknowledged in an author footnote.

Table 3.4 Behaviors Constituting Substantial and Not Substantial Contributions.

Substantial Contribution	1. Participating in the conceptualization, design, implementation, analysis, and description of a study or activity entitles a person to be an author on the products of that work. However, membership on a research team where the research or activity is discussed regularly but with no other involvement does not entitle a person to serve as an author. 2. Conceptualizing, developing an outline, or designing a product, activity, or study is sufficient contribution. 3. Conducting a substantial amount of the experimental sessions is sufficient contribution. 4. Supervising the day-to-day implementation of experimental sessions is sufficient contribution. 5. Securing funding and participating in decisions about the conceptualization, design, implementation, analysis, or description of the research, activity, or product is sufficient contribution. However, only securing funding is not a sufficient contribution. 6. Collecting, coding, entering, summarizing, *or* conducting statistical analyses of the data *and* participating in the decisions about the design and analysis are sufficient contribution. 7. Writing a major portion of a product based on the research or activities conducted by others is sufficient contribution.
Not a Substantial Contribution	1. Collecting data, coding data, entering data, summarizing data, maintaining a data base, conducting literature searches under the guidance of another, *or* conducting statistical analyses under the direction of another *without* participating in the decisions related to the design, implementation, or analysis of the research or activity do not constitute sufficient contribution. 2. Reading pre-submission/publication drafts of products and providing feedback do not constitute sufficient contributions. 3. Providing access to participants does not constitute sufficient contribution. 4. Serving as a participant or as a rater of some aspect of a product (e.g., social validity, validation of a questionnaire) does not constitute sufficient contribution. 5. Providing periodic consultation to a person or group on a study or on a product does not constitute sufficient contribution. 6. Providing funding for a study but not being involved in the study does not constitute sufficient contribution.

7. Providing individuals with opportunities to develop a product (e.g., introducing an author to a publisher) does not constitute sufficient contribution.

Table 3.5 Guidelines for Determining the Order of Authorship.

1. Intellectual contribution takes precedence over material, operational, or descriptive contributions.
2. Material contribution takes precedence over operational or descriptive contributions.
3. Operational contribution takes precedence over descriptive contributions.
4. Combinations of these contributions may take precedence over other single contributions (e.g., operational and descriptive contributions could take precedence over material contributions).
   When multiple products are completed by a set of authors and the contributions to those products are generally equal, then the order of authorship should be counterbalanced across products.
5. When a study is a student thesis or dissertation, then the student is always the first author, but others should be included based on their contribution (intellectual, material, operational, and/or descriptive).
6. When a report is based on a student thesis or dissertation, the student maintains rights such as deciding whether, when, and where the manuscript is submitted and who is included as a co-author and order of authorship. However, when the student's study was part of an investigator's funded project, then the investigator and student share in the rights to make these decisions and a responsibility to disseminate the findings.
7. When multiple products are completed by a set of authors and the contributions to those products are essentially equal, then the order of authorship should be counterbalanced across products.

General Guidelines

This section has some general guidelines for determining issues of authorship. First, decisions about authorship (inclusion and order) should be made based on the contributions of each individual with a goal of including only individuals who offer a substantial contribution. Second, if an error (over or under inclusion) is made, we encourage over-inclusion rather than under- inclusion, with fraud being the more serious violation of the two. Third, decisions regarding inclusion and order should be made *early* in the development of a product or a study rather than after it is completed. In our collaborations, we often draft a reference for our CVs that includes the initially agreed upon author listing and proposed title to avoid confusion or hurt feelings later. However, changes to the initial decision can be made if the planned contributions change over the course of the study. Fourth, disputes about authorship should be discussed first with the senior author and/or principal investigator and then with all persons concerned. Fifth, maintain a list of all individuals who contribute to the research, activity, or product. The list should include those who will be authors and those who

will be acknowledged. Sixth, allow each author to read the product before it is submitted or disseminated and ask each person to sign a statement indicating they recognize they are included as an author on the product and they affirm the accuracy and integrity of the product. Finally, when a product (e.g., book, chapter, manual) is developed for publication or distribution and may result in financial gain, then the amount of the financial rewards should be negotiated among the authors when the product is initially conceived. This amount should be reviewed when the product is completed to ensure it reflects actual contributions. A written agreement among authors is recommended.

Once the issues of authorship (inclusion and order) are addressed, it is time to move forward with dissemination activities. We offer brief input on the following dissemination activities: (a) poster presentations, (b) conference seminar presentations, (c) web-based publications, and (d) refereed journal articles.

Poster Presentations

Poster presentations at conferences provide an opportunity to disseminate studies in a format that is potentially less nerve wracking than a formal presentation. When presenting a poster, you prepare a visual display of the study and stand near it. The visual display has most of the same sections as a written study report: title, introduction, method, results, and discussion. The poster should not have dense prose; rather, less text, more bullets, large font size, and more graphics are recommended. Individuals walk by, look at the poster, read the information, and can ask questions. Be prepared to give a 2–3 minute overview of the study. It also is wise to prepare a one-to-two-page description of the study with graphs to give to people who are interested. This format allows for brief conversations as well as extended conversations with highly interested people.

Conference Seminar Presentation

A conference seminar presentation will include the same study components presented in a manuscript. Often you will need to prepare a visual presentation (e.g., PowerPoint™) to guide your oral discussion. Below are some suggestions for preparing and delivering presentations:

Preparation Activities

Prior to submitting a presentation, be certain the study will be sufficiently complete to have meaningful data to discuss. Also, be sure you will have sufficient time to analyze the data. When making the presentation, the slides must be readable (e.g., font size of 20 or more) from the back of the room. Usually, a brief background to the study is followed by research questions, overview of methods, and results. If possible, practice with an audience of your peers to get feedback and gain confidence in responding to questions.

Another reason to practice ahead of time is to ensure that you abide by the timeframe allotted. Often conference seminars will have three or four studies presented by different authors in an hour. If you are allotted 15 minutes, do not exceed that limit, because you will use the time devoted to others.

Delivering Presentations

When it is time to present, dress professionally, find your presentation room well before the start of your presentation, restrict your presentation to the time limit, and answer questions respectfully and thoughtfully. In terms of dress, some conferences are more casual than others. If possible, consult with those who have attended the conference before to determine the level of formality. Conferences often occur in large hotels with a variety of different rooms for presentations. Unless you find your room beforehand, you might miss your time slot. As mentioned above, pace the presentation to allow sufficient time to cover the intended content and still answer questions. People attend conferences to seek information and time needs to be devoted to answering their questions. Some individuals will pose questions that appear challenging or thought provoking, whereas others will pose questions or comments that may appear inane. Respond respectfully in all instances; it is acceptable to say, "I don't know" to questions for which you do not know the answer. Present your work completely, with candor, and with integrity.

At the conclusion of the conference, thank those involved. We often send a thank you note to the conference coordinators to acknowledge them for their time and effort in organizing the event. On our research team, we have a tradition in which the lead presenter (for poster and presentation) emails out the line item for everyone's CV to those who participated along with a thank you. This kindness also ensures the line items are consistent on each person's CV.

Web-Based Publishing

The internet is widely accessible to many individuals and this makes it a tempting means of disseminating information. As is widely known, nearly anyone can put nearly anything on the web. We do not recommend simply putting your study reports on the web on your own home page or some other non-scientific outlet. Some journals (e.g., *Journal of Early and Intensive Behavioral Intervention*) are only published on the web; they do not have a corresponding paper version of the journal. Some of these are legitimate outlets for scientific products. The defensible ones are similar to hard-copy journals in the following ways: they have an editor, they have an editorial board comprised of reputable scientists, their review process is described, and they use a peer-review process (described below). Such journals are likely to become progressively more common. While it is acceptable to publish in such web-based journals, we recommend avoiding those that do not have an editorial board, do not use the peer-reviewed process,

and require payment for publication.

Refereed (Peer-Reviewed) Journals

Conducting research and learning something relevant is a major accomplishment and the findings should be shared. One of the most prestigious and rigorous methods of study dissemination is publication in refereed journals. This section describes how to (a) select a journal for submission, (b) write the article and prepare for submission, (c) submit the manuscript, and (d) participate in the review process.

Select a Journal for Submission

Journals vary greatly in terms of the types of papers published, readership, type of research methods accepted, and perceived quality and rigor. We suggest the following steps for choosing a journal. First, identify the types of readers who are most appropriate for your manuscript and identify journals most apt to have such a readership. For example, if you are interested in sharing your behavioral research with other scholars with interest in rigorous applied behavior analytic studies, you might consider publishing in the *Journal of Applied Behavior Analysis*. However, if you are more interested in reaching both researchers and practitioners, you might consider publishing in the *Journal of Behavioral Education* or *Journal of Positive Behavior Supports*. Second, of those with an appropriate readership, identify the journals publishing papers similar to your study in terms of independent and dependent variables and similar in terms of research methods. Finally, of those with an appropriate readership and matching methods, identify the one with the highest perceived quality. Look at the reference list in your study; consider publishing in the journals where many of the authors you cited published.

Write the Article and Prepare for Submission

Once a journal is selected, examine a recent issue of the journal (or the journal's website) to obtain some key information such as (a) the desired length of manuscripts—most journals present some outer limit; (b) any unique formatting or presentation guidelines; (c) number of copies to submit (if they do not have an online submission process); and (d) how manuscripts are submitted (electronically, hard copy, etc.). Often there is a section titled, "guidelines for authors" or "information for authors." This section often is on the inside cover of the journal. Some journals (e.g., *Topics in Early Childhood Special Education, Topics in Language Disorders, Remedial and Special Education,* and *Journal of Positive Behavior Interventions*) have topical issues—meaning they are seeking papers on a specific topic and have specific due dates for such manuscripts.

Because most theses and dissertations exceed article length, students often need to do

a major revision in which the length of their initial document is decreased and reorganized to adhere to APA and journal guidelines. Generally, consider the following: (a) the introduction is between three and six pages in length; (b) the method section is completed with sufficient details about participants characteristics, participant selection criteria, overall procedures, intervention procedures, measures (including procedural fidelity and social validity), and experimental design; (c) the results section must be quite succinct with narrative and only the key figures and tables; and (d) the discussion section should be relatively short (three to five pages) focusing on how the study confirms or extends previous investigations, limitations, future directions, and educational implications. Horner et al. (2005) and Wolery et al. (2011) present recommendations and guidelines for conducting and reporting research using SCDs.

In preparing a paper for submission, attend carefully to the APA style manual and the specific journal's editorial guidelines (e.g., page length maximums, types of submissions considered, and focus of the journal). All authors should read the paper before it is submitted, give feedback, and give explicit approval for their name to be included on the paper. It is good practice to have someone unassociated with a manuscript read it prior to submission. The intent is to obtain feedback on the logic, readability, presentation, and mechanics. The goal is to submit a clean manuscript free of presentation errors (spelling, formatting, grammatical, punctuation).

Submit the Manuscript

Submitting the paper includes several steps: (1) assuming authorship was previously determined, confirm all authors agree to manuscript submission, (2) gather all required information, (3) write a cover letter, and (4) submit the manuscript and related files. Before beginning the submission process, it is wise to research requirements specific to the journal. For example, journals typically accept manuscripts of about 30 double-spaced pages, but this varies widely even among journals in a single field. Prior to submission, you will need to gather needed documents, usually including: (1) title page with authors' names and contact information; (2) blinded manuscript (i.e., with no title page or other identifying information), generally in a Microsoft Word file; (3) figures; and (4) tables, unless they are in the blinded manuscript file.

The cover letter is a request for a review of the paper, and is generally uploaded separately from the other documents. Sometimes editors want specific information addressed in the cover letter (contact information, information on approval from the institutional review board [IRB], etc.); specific information is generally available on each journal's website. The letter is usually addressed to the editor (see the author guidelines). The letter usually starts by saying, "Please consider [title] for review and possible publication in [Journal Title]." If the manuscript has multiple authors, it is often wise to state that all authors agreed to submit it to the journal. If it contains original data, state that that IRB approval was received before the study was initiated. State that you will not submit the paper to another journal during the time it is under review by the journal

to which you are submitting. It is an ethical violation to submit the same manuscript to more than one journal at a time, and it is a legal violation (copyright infringement) to publish the same article in more than one journal or other source. Generally, you may present a paper at a conference (as a poster or formal publication) and also submit it to a journal for publication. Give the contact information (address, email, phone number) of the corresponding author. Finally, thank them for their consideration of the paper. Most websites provide an automated reply when a paper is submitted. Following review (generally 90 days), you will be notified of the publication decision via an email, generally from the editor.

Participate in the Review Process

When editors receive a manuscript for review (which is often done electronically), they send it to three to five reviewers. Sometimes they send it to an associate editor who in turn sends it to the reviewers. The reviewers are typically given a date by which they are to submit their reviews and specific guidelines for conducting the review. After the editor has received the reviews a decision is made, which often includes one of the following: accept, accept contingent upon revisions, reject but invite resubmission, or reject. These are defined in Table 3.6. The editor makes the decision, communicates that in a letter to the author, and often describes the needed changes. The reviewers' comments also are sent to the author. Most journals use a "double blind" review process; specifically, the reviewers are not told who authored the paper and the authors are not told who reviewed the paper. This process ensures scientific integrity and reduces the chance of bias; thus, it is preferable to a non-blind review process, although some journals still conduct non-blind reviews (e.g., *Journal of Applied Behavior Analysis*).

Table 3.6 Description of Editorial Decisions.

Decision	Description
Accept	Very rarely is a paper accepted without revisions—once or twice in a lifetime of publishing, unless of course it is an invited commentary on something and even then revisions are likely.
Accept contingent upon revisions	This decision means the editor is willing to publish the paper if the authors are able to make the requested revisions in a satisfactory way. The editor often, but not always, is explicit about what revisions are needed. Even when the editor's letter details the needed revisions, it is useful to examine the reviewers' comments about the paper. When this is the decision, you should make the revisions quickly and submit the required number of copies. Often the editor's letter includes a specific date by which the revisions need to be done. Do not send it to another journal.
Reject but invite	This decision means the editor is not ready to accept it

resubmission (revise/resubmit)	and not ready to reject it. Usually, fairly major revisions are requested, and sometimes this involves reanalysis of the data, inclusion of additional data, inclusion of new information, or reframing the paper in a major way. Sometimes editors are explicit about what needs to be changed, and at other times they may just refer you to the reviewer's comments. If you get such a decision, carefully determine whether you can respond adequately to the concerns. If a revision is submitted, the manuscript is almost always sent back out for review, often to a couple of the original reviewers and one or two new reviewers. With a reject but resubmit decision, you are free to send it to another journal or revise and resubmit it to the original journal (but not both!). Often, but not always, giving the original journal another try is worth the effort.
Reject	This means the editor will not publish the paper. When this occurs, you are free to submit it to another journal, although you should attend to the reviewer's comments in making the decision to submit the paper to another journal.

After receiving the review, read the decision letter carefully. Sometimes (particularly when the review is less than favorable) it is helpful to set the review aside and read it after the initial reaction passes. You do not need to respond to the editor when a manuscript is rejected or when it is a rejection with an invitation to resubmit (unless asked to do so), although some authors will send a brief thank you expressing appreciation for a timely and thorough review (if that was the case). If you decide to refute a decision, we urge to you to be respectful in all communication.

When the decision is to accept with revisions, attend carefully to the reviews and to the editor's letter about the nature of the needed revisions. Make the revisions as quickly as possible, and send them back to the editors. Most editors want a letter describing what and how changes were made, and they want a justification for any requested revisions which were not done. Depending upon the editor and journal, there may be multiple rounds of revisions. Also, just because your paper was accepted contingent upon the revisions, it is not really accepted until the editor is satisfied with the revisions.

Once a paper is accepted, the editor generally sends a message to let you know the paper was accepted. Although this is a major hurdle, the process is not complete. You will receive page proofs which include queries to address final questions. Hopefully your final submission was highly polished (e.g., reference check complete, no grammar or APA errors). If not, those issues will be addressed here. They will ask you to read it again to make sure they have not changed your meaning and to catch any typographical errors. They rarely allow major changes in the paper at this stage. They often give a 24- or 48-hour turn around deadline. It is an important step; read your paper carefully, make

sure the tables and figures are correct, and make sure the references to the figures/tables are what you intended. This will be your last opportunity to edit your manuscript. Check the paper, respond to their queries, and return it as soon as possible.

Often with the page proofs, you and the other authors will be asked to complete a copyright transfer agreement. The copyright transfer agreement is a legal document transferring the copyright from the authors to the publisher. Journals will not publish a paper if the copyright transfer is not signed. Sometimes the corresponding author can sign this form, but most publishers require all authors to sign the form. Journals do not pay authors for publishing their articles; we publish to share information with colleagues. Most reputable journals do not request payment from authors for publication, although many web-based journals do so and this is seen as a weakening of the peer review process. At this point, you wait. Some months later the manuscript will appear in the journal as published form! Now, celebrate!

Summary

In this chapter, we provided an overview of some key standards for preparing documents adhering to scientific writing, the specific guidelines for reviewing the literature, writing different types of research questions, writing research proposals, and describing completed investigations. We also made some recommendations for disseminating what you have learned.

References

Alexander, J. L., Smith, K. A., Mataras, T., Shepley, S., & Ayres, K. M. (2015). A meta-analysis and systematic review of the literature to evaluate potential threats to internal validity in probe procedures for chained tasks. *The Journal of Special Education*, *49*, 135–145.

Alig-Cybriwsky, C. A., Wolery, M., & Gast, D. L. (1990). Use of a constant time delay procedure in teaching preschoolers in a group format. *Journal of Early Intervention*, *14*, 99–116.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Foundation.

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

Beller, E. M., Glasziou, P. P., Altman, D. G., Hopewell, S., Bastian, H., Chalmers, I.,. . PRISMA for Abstracts Group. (2013). PRISMA for abstracts: Reporting systematic reviews in journal and conference abstracts. *PLoS Med*, *10*(4), e1001419.

Billingsley, F. F., White, O. R., & Munson, R. (1980). Procedural reliability: An example and rationale. *Behavioral Assessment*, *2*, 129–140.

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

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. New York, NY: John Wiley.

Charlop, M. H., Kurtz, P. F., & Casey, F. G. (1990). Using aberrant behaviors as reinforcers for autistic children. *Journal of Applied Behavior Analysis*, *23*, 163–181.

Charlop-Christy, M. H., & Haymes, L. K. (1996). Using obsessions as reinforcers with and without mild reductive procedures to decrease inappropriate behaviors of children with autism. *Journal of Autism and Developmental Disorders*, *26*, 527–546.

Common, E. A., Lane, K. L., Pustejovsky, J. E., Johnson, A. H., & Johl, L. E. (2017). Functional assessment-based interventions for students with or at-risk for high incidence disabilities: Field-testing single-case syntheses. *Remedial and Special Education*. doi:10.1177/07419325176933

Cook, B. G. (2016). Reforms in academic publishing: Should behavioral disorders and special education journals embrace them? *Behavioral Disorders*, *41*, 161–172.

Cuneo, A. (2007). *Enhancing preschoolers' conversation during themed play using in-school play dates*. Unpublished master's degree thesis. Vanderbilt University.

Doyle, P. M., Wolery, M., Ault, M. J., & Gast, D. L. (1988). System of least prompts: A review of procedural parameters. *Journal of the Association for Persons With Severe Handicaps*, *13*, 28–40.

Doyle, P. M., Wolery, M., Gast, D. L., Ault, M. J., & Wiley, K. (1990). Comparison of

constant time delay and the system of least prompts in teaching preschoolers with developmental delays. *Research in Developmental Disabilities*, *11*, 1–22.

Filla, A., Wolery, M., & Anthony, L. (1999). Promoting children's conversations during play with adult prompts. *Journal of Early Intervention*, *22*, 93–108.

Green, V. A., Pituch, K. A., Itchon, J., Choi, A., O'Reilly, M., & Sigafoos, J. (2006). Internet survey of treatments used by parents of children with autism. *Research in Developmental Disabilities*, *27*, 70–84.

Gresham, F. M., Gansle, K. A., & Noell, G. H. (1993). Treatment integrity in applied behavior analysis with children. *Journal of Applied Behavior Analysis*, *26*, 257–263.

Griffen, A. K., Wolery, M., & Schuster, J. W. (1992). Triadic instruction of chained food preparation responses: Acquisition and observational learning. *Journal of Applied Behavior Analysis*, *25*, 193–204.

Hemmeter, M. L., Maxwell, K. L., Ault, M. J., & Schuster, J. W. (2001). *Assessment of practices in early elementary classrooms*. New York, NY: Teachers College Press.

Hitchcock, C. H., Dowrick, P. W., & Prater, M. A. (2003). Video self-modeling intervention in school-based settings: A review. *Remedial and Special Education*, *24*, 36–46.

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.

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

Johnston, J. M. (1988). Strategic and tactical limits of comparison studies. *The Behavior Analyst*, *11*, 1–9.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research* (2nd ed.). Hildale, NJ: Lawrence Erlbaum Associates.

Jolivette, K., Wehby, J. H., Canale, J., & Massey, N. G. (2001). Effects of choice-making opportunities on the behavior of students with emotional and behavioral disorders. *Behavioral Disorders*, *26*, 131–145.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Allyn and Bacon.

Kern, L., Vorndran, C. M., Hilt, A., Ringdahl, J. E., Adelman, B. E., & Dunlap, G. (1998). Choice as an intervention to improve behavior: A review of the literature. *Journal of Behavioral Education*, *8*, 151–170.

Kettler, R. J., & Lane, K. L. (2017). Methodological foundations of school psychology (MFSP) research and practice. *Under Review*.

Lane, K. L. (2004). Academic instruction and tutoring interventions for students with emotional/behavioral disorders: 1990 to present. In R. B. Rutherford, M. M. Quinn, & S. R. Mathur (Eds.), *Handbook of research in emotional and behavioral disorders* (pp. 462–486). New York, NY: Guilford Press.

Lane, K. L., Little, M. A., Rhodes, J. R., Phillips, A., & Welsh, M. T. (2007). Outcomes of a

teacher-led reading intervention for elementary students at-risk for behavioral disorders. *Exceptional Children*, *74*, 47–70.

Lane, K. L., Robertson, E. J., & Graham-Bailey, M. A. L. (2006). An examination of school-wide interventions with primary level efforts conducted in secondary schools: Methodological considerations. In T. E. Scruggs & M. A. Mastropieri (Eds.), *Applications of research methodology: Advances in learning and behavioral disabilities* (vol. 19). Oxford: Elsevier.

Lane, K. L., Weisenbach, J. L., Little, M. A., Phillips, A., & Wehby, J. (2006). Illustrations of function-based interventions implemented by general education teachers: Building capacity at the school site. *Education and Treatment of Children*, *29*, 549–671.

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

Ledford, J. R., & Gast, D. L. (2006). Feeding problems in children with autism spectrum disorders: A review. *Focus on Autism and Other Developmental Disabilities*, *21*, 153–166.

Ledford, J. R., Lane, J. D., Elam, K. L., & Wolery, M. (2012). Using response prompting procedures during small group direct instruction: Outcomes and procedural variations. *American Journal of Intellectual and Developmental Disabilities*, *117*, 413–434.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Oaks, CA: Sage.

Logan, K. R., & Gast, D. L. (2001). Conducting preference assessments and reinforcer testing for individuals with profound multiple disabilities: Issues and procedures. *Exceptionality*, *9*, 123–134.

Maggin, D. M., Talbott, E., Van Acker, E. Y., & Kumm, S. (2017). Quality indicators for systematic reviews in behavioral disorders. *Behavioral Disorders*, *42*, 52–64.

Messenger, M., Common, E. A, Lane, K. L., Oakes, W. P., Menzies, H. M., Cantwell, E. D., & Ennis, R. P. (2017). Increasing opportunities to respond for students with internalizing behaviors: The utility of choral and mixed responding. *Behavioral Disorders, 42*, 170–184.

Moher, D., Cook, D. J., Eastwood, S., Olkin, I., Rennie D., … for the QUOROM Group. (1999). Improving the quality of reporting of meta-analysis of randomized controlled trials: The QUOROM statement. *Lancet*, 354, 1896–1900.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Prisma Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS medicine*, *6*(7), e1000097.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., … PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Systematic Reviews*, *4*(1), 1–9.

Mulrow, C. D. (1987). The medical review article: State of the science. *Annals of Internal Medicine*, *106*, 485–488.

Munson, L. J., & Odom, S. L. (1996). Review of rating scales that measure parent-infant interaction. *Topics in Early Childhood Special Education*, *16*, 1–25.

Premack, D. (1959). Toward empirical behavioral laws: I. Positive reinforce. *Psychological Review*, *66*, 219–233.

Royer, D. J., Lane, K. L., Cantwell, E. D., & Messenger, M. (2017). A systematic review of the evidence base for instructional choice in K-12 settings. *Behavior Disorders*, *42*. doi:10.1177/0198742916688655

Sacks, H. S., Reitman, D., Pagano, D., & Kupelnick, B. (1996). Meta-analysis: an update. *The Mount Sinai Journal of Medicine, New York*, *63*, 216–224.

Schuster, J. W., Morse, T. E., Ault, M. J., Doyle, P. M., Crawford, M. R., & Wolery, M. (1998). Constant time delay with chained tasks: A review of the literature. *Education and Treatment of Children*, *21*, 74–106.

Shogren, K. A., Faggella-Luby, M. N., Bae, S. J., & Wehmeyer, M. L. (2004). The effects of choice-making as an intervention for problem behavior: A meta analysis. *Journal of Positive Behavior Interventions*, *6*, 228–237.

Stewart, L. A., Clarke, M., Rovers, M., Riley, R. D., Simmonds, M., Stewart, G., & Tierney, J. F. (2015). Preferred reporting items for a systematic review and meta-analysis of individual participant data: the PRISMA-IPD statement. *JAMA*, *313*(16), 1657–1665.

Stokes, T. F., & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis*, *10*, 349–367.

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

Wendt, O., & Miller, B. (2012). Quality appraisal of single-subject experimental designs: An overview and comparison of different appraisal tools. *Education and Treatment of Children*, *35*, 235–268.

Werts, M. G., Wolery, M., Holcombe, A., Gast, D. L. (1995). Instructive feedback: Review of parameters and effects. *Journal of Behavioral Education*, *5*, 55–75.

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

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

Wolery, M., Kirk, K., & Gast, D. L. (1985). Stereotypic behavior as a reinforcer: Effects and side-effects. *Journal of Autism and Developmental Disorders*, *15*, 149–161.

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

4
Replication

David L. Gast and Jennifer R. Ledford

Important Terms

replication, external validity, sequential introduction and withdrawal designs, time-lagged designs, rapid iterative alternation designs, intra-participant direct replication, inter-participant direct replication, clinical replication, systematic replication

Replication

In both applied and basic research, **replication** can be described as an investigator's ability to repeat the effect an independent variable has on the dependent variable(s). Replication is important in all research paradigms, but the replication rate in published single case design (SCD) research is higher than that of between-groups research (Lemons et al., 2016). Sidman (1960), in *Tactics of Scientific Research*, provided the definitive word regarding replication, in which he discussed two types of replication: direct replication and systematic replication. Although much of his discussion was written in the context of basic experimental psychology laboratory studies, the impact his insightful discussion of replication has had on applied behavioral research has been far-reaching. It is not likely that you will find a text devoted to SCD research methodology, or applied behavioral analysis, that doesn't reference Sidman (1960) and his contributions to behavioral science. His differentiation of direct replication, and its importance for evaluating the reliability of findings, and systematic replication, and its importance for evaluating the generality of findings, has provided applied researchers with a framework for evaluating research. It is important to recognize that replication is at the heart of all science, and only through successful replication attempts can we gain confidence that experimental findings have both internal validity (likelihood outcomes observed in a study are due to the intended differences between conditions rather than other plausible factors) and **external validity** (degree to which outcomes from a study are likely to be generalizable outside of the study context; i.e., generality). Our confidence in research findings is directly related to the consistency within and across research attempts. This chapter addresses the parameters of replication you should attend to in the design, implementation, and evaluation of research.

"The soundest empirical test of the reliability of data is provided by replicating" (Sidman, 1960, p. 70). Any study should be situated into the context of other studies that have addressed the same or similar presenting problem, and manipulated the same or similar independent variable. Through a comprehensive literature review you will gain insight into what interventions have been effective, what modifications have been made to make the original independent variable effective, and what gaps there are in the research conducted to date. There are few research studies that are truly novel or unique; most are extensions or modifications of previous research. That is, most research investigations attempt to expand our knowledge beyond what we currently know. This is accomplished by identifying gaps in the research literature and investigating whether a previously studied intervention will be as effective with other populations, or with other behaviors, or can be modified to be more effective and/or efficient.

Most studies are conducted to answer simple questions. Suppose that you conduct one study with one participant and that the data show a clear effect. Your reaction will be an enthusiastic and joyful "I did it and it worked!" to which the scientific community will

(or should) reply "Yes, but … ?" You have demonstrated the reliability of effect, but with only one participant. The next question, then, is "If I repeat the experiment with similar but different participants, holding all other variables constant used with the original participant, will I get the same effect?" In other words, "Are my results reliable across other similar participants?" This question addresses both the reliability of effect and, to the extent to which participants (and conditions) differ, generality of effect. Suppose that you were successful in your replication attempt with three similar participants and you were able to keep all pertinent condition variables the same. Will these data quiet the scientific community? Possibly … but not for long. A single study, conducted with three or four similar participants, in which the independent variable consistently has a positive effect on behavior will gain the attention of the research community, and may be published in a refereed journal, assuming all threats to internal validity were controlled for, procedures were clearly described, data were accurately analyzed, and information was reported according to publication guidelines. The questions others in the research community will ask, having read your research report, are: "Will I get the same results that you got with different participants, at a different research site?" and "How broadly will the results generalize beyond the original experiment?" When replicating a previously-conducted study, the greater the number of differences from the original study, the greater the risk of not replicating the effect, but if successful, the greater the evidence for generality.

If the results of an intervention are spurious and cannot be reproduced reliably across participants in a single investigation, it is unlikely anyone will attempt to use the intervention. However, you should not be discouraged when there is a "failure to replicate"; rather it should inspire you to attempt to identify the reasons for failure. As an applied SCD researcher, it is your responsibility to see that participants in your study benefit from their participation; thus, if your originally-implemented intervention is ineffective, you should implement a modification to that intervention, or an alternative intervention, that brings about positive behavior change.

There are three primary reasons or purposes for attempting to replicate the findings of a study or series of studies:

1. Assess the reliability of findings (internal validity).
2. Assess the generality of the findings (external validity).
3. Look for exceptions.

Each of these reasons for replication will be addressed in the discussion of direct and systematic replication. Sidman (1960) succinctly and cogently addressed the importance of replication when he wrote:

> To the neutral observer it will be obvious that science is far from free of human bias, even in its evaluation of factual evidence. Experimental findings, furthermore, are so fragile when considered within the total matrix of natural phenomena from which they are lifted, and conclusions from such data often so tenuous, that one can only feel surprise at the actual achievements of experimental methodology. What must we work with in any experiment? Uncontrolled, and even unknown, variables; the errors of selective perception arising out of

> theoretical and observational bias; indirect measurements; the theory involved in the measurement techniques themselves; the assumptions involved in making the leap from data to interpretation. In short, we have a margin of error so great that any true advance might be considered an accident were it not for the fact that too many genuine advances have occurred in too short a time for the hypothesis to be entertained seriously.
>
> (p. 70)

It is through replication that we reduce the margin of error and increase confidence that findings that withstand repeated tests are real, not accidental.

Direct Replication

Sidman (1960) defines direct replication 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 participant under each of several conditions" (p. 73). Two types of direct replications are described: intra-participant direct replication and inter-participant direct replication (historically—"intra-subject" and "inter-subject"). Both intra-participant and inter-participant direct replications refer to an investigator's attempts to repeat an experimental effect with the same participant (intra-participant), or across participants in the same study (inter-participant). In SCD studies with more than one participant, your investigation may address both intra- and inter-participant replication. In its narrowest and most conservative definition, direct replication is only possible in laboratory studies using infrahuman subjects (rats, pigeons etc.); however, in applied research with human participants direct replication is more broadly defined.

Condition Ordering and Direct Replication in Single Case Design

In SCD research, there are three primary ways to ensure within-study replication (Kratochwill et al., 2010), all of which involve repeating a change between two adjacent conditions. **Sequential introduction and withdrawal designs** include the repetition of the basic A-B comparison within a single participant (e.g., A-B-A-B). **Time lagged designs** include the repetition of the basic A-B comparison across a set of three or more participants, behaviors, or contexts. **Rapid iterative alternation** designs include repetition of an A-B comparison, with single session replication and comparisons. See Figure 4.1 for illustrations of replication in basic SCD types. Using replication by adhering to these prescribed condition ordering types increases internal validity by reducing the possibility that outcomes are related to extraneous variables rather than to your independent variables. When a researcher conducts sufficient replications and demonstrates consistent effects, he or she has demonstrated experimental control of the independent variable on the dependent variable.

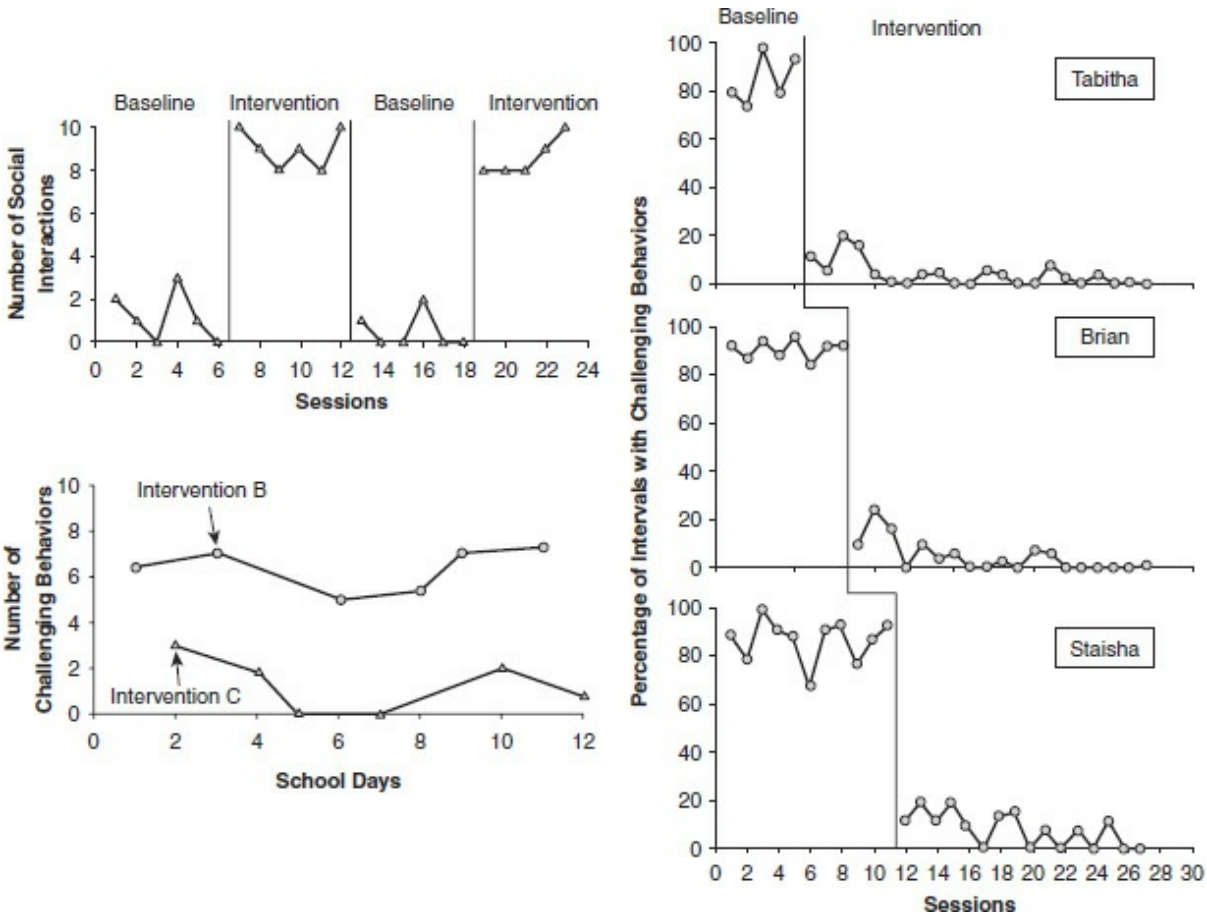


Figure 4.1 Single case designs illustrating three methods for ordering conditions: (1) Sequential introduction and withdrawal (top left), (2) Rapid iterative alternation (bottom left), and (3) time-lagged introduction (right).

Direct Intra-participant Replication

Direct intra-participant replication (historically termed "intra-subject") refers to *repeating the experimental effect with the same participant* more than once in the same study. For example, when an investigator uses an A_1-B_1-A_2-B_2 withdrawal experimental design, if the removal in the independent variable in A_2 results in a return to levels observed in A_1, and the re-introduction of the independent variable in B_2 results in a return to observed levels in B_1, intra-participant replication has been achieved (i.e., the investigator has been able to show that the presence or absence of the independent variable will determine the level of the dependent variable). If an investigator uses a multiple baseline design across behaviors, in which the intervention is systematically introduced across three or more similar but independent behaviors, intra-participant replication is achieved if there is a change in the dependent variable only upon introduction of the independent variable, not before. Cooper, Heron, and Heward (2007) have described the intra-participant replication process in four stages: prediction (A_1), affirmation (B_1), verification (A_2), and replication (B_2). During A_1 the data pattern is used to *predict* the data pattern if there is no change in experimental condition; B_1 data

pattern *affirms* that the independent variable may have had an effect on the behavior; A_2 data pattern *verifies* there is a cause-effect relation between independent and dependent variables at the simplest level; B_2 data pattern *replicates*, or repeats the effect that the independent variable has on the dependent variable, thus increasing confidence that there is functional relation between independent and dependent variables. Often, these changes in behavior across three changes in conditions (A_1 to B_1, B_1 to A_2, and A_2 to B_2) are simply referred to as three "demonstrations of effect" or "replications of effect." Through the intra-participant direct replication process you gain confidence that you have demonstrated "reliability of effect" (i.e., internal validity, with this one participant). Your objective now is to establish this same effect with other participants included in your study.

Inter-Participant Direct Replication

While intra-participant direct replication refers to repeating the effect with the same participant, **inter-participant direct replication** (historically inter-subject replication) refers to *repeating the experimental effect with different participants*. The importance of inter-participant replication was concisely stated by Sidman (1960) when he wrote, "When an experiment is performed with a single organism as the subject, inter-subject replication is often demanded on the grounds that the original subject may have been a 'freak'," and he went on to write, "The purpose of inter-subject replication is to determine whether uncontrolled and/or unknown variables might be powerful enough to prevent successful replication" (p. 74). It is important to realize that the level of confidence you can have in a study, yours or others, is limited with only one participant. Although "N=1 studies" appear in refereed research journals, findings of these studies must be accepted with caution since the generality across other individuals and conditions has not been established. A common reason for the publication of single participant studies is that they frequently address novel interventions or unusual challenges that journal editors and reviewers believe warrant dissemination with the hope of encouraging others to attempt a replication. Although a study with only one participant is acceptable under some circumstances, we recommend that you start your investigation with a minimum of three participants regardless of the specific design you use.

Within an SCD study, researchers must include at least three inter- or intra-participant replications; additional designs can be used to add additional inter-participant replications. Generality of findings is primarily established through systematic replication across designs. Figure 4.2 presents three graphic displays that illustrate inter-participant replication with an A-B-A-B design, multiple baseline design across behaviors, and multiple baseline design across participants, respectively. In Figure 4.2a the effectiveness of the independent variable is repeated across Participant 1, Participant 2, and Participant 3 in the same study with the same investigator. Figure 4.2b illustrates how inter-participant replication is addressed when using a multiple baseline

design across behaviors. The effect of the independent variable on the dependent variable is repeated across each of the three participants. As with studies that employ an A-B-A-B design, intra-participant and inter-participant replications are addressed by repeating the effectiveness of the intervention with each participant and across participants. In both of these cases (A-B-A-B and multiple baseline design across behaviors), inter-participant replication is completed in the context of a separate experimental design for each participant; in both cases, there are nine total demonstrations of effect (denoted by circled numbers on graphs).

Figure 4.2c illustrates a multiple baseline design across participants. Unlike the A-B-A-B design and multiple baseline design across behaviors, intra-participant replication is not addressed. As will be discussed further in Chapter 10, the demonstration of experimental control (reliability of effect) and the generality of the findings with this design rests with the number of successful inter-participant replications out of the number attempted, and similarity in the data patterns (level and trend) across participants. Many, if not most, behavioral researchers would consider a multiple baseline design across participants a "weaker" evaluation and demonstration of experimental control compared to a multiple baseline design across behaviors or an A-B-A-B design because of the lack of intra-participant replication. From a strictly numerical perspective, Figure 4.2c illustrates that with the same number of study participants as shown in Figure 4.2a and 4.2b, three, there are only three demonstrations of independent variable effectiveness (indicated via circled numbers), and for that reason we recommend that more than the minimum number participants be included in your study when using a multiple baseline design across participants.

Sidman (1960) discusses a variation of inter-participant replication that he labels "inter-group" replication. Inter-group replication refers to repeating the effects of an intervention with different groups of individuals by comparing measures of central tendency (mean, median, mode). As with any comparison using measures of central tendency (including group design research), the findings will both underestimate and overestimate the effectiveness of the intervention by not reporting the individual data of "outliers"; there may be some members of the group whose behavior did not change when the intervention was implemented; some group members' behavior may have changed considerably more than the reported average. In this context, and in contrast to an inter-participant replication in which individual data are reported, Sidman writes:

> As a criterion of reliability and generality, inter-subject replication is a more powerful tool than intergroup replication. Intergroup replication provides an indicator of reliability insofar as it demonstrates that changes in central tendency for a group can be repeated. With respect to generality, however, intergroup replication does not answer the question of how many individuals the data actually represent. With inter-subject replication, on the other hand, each additional experiment increases the representativeness of the findings.
>
> (p. 75)

In light of this limitation associated with inter-group replication, applied researchers who study the behavior of groups of individuals, and who make research decisions based on group performance rather than each individual's performance, will often present,

analyze, and report individual data on each member of the group. Stinson, Gast, Wolery, and Collins (1991) exemplified this in their study of observational and incidental learning by four students with moderate intellectual disabilities. They presented one graph in which the mean performance of the group was plotted, and four graphs in which individual performance was plotted. Two tables were also used to summarize each individual's acquisition of incidental and observational information. It is important to remember, if research decisions are made based on some measure of central tendency of the group, it is the group's data that should "take center stage" and be graphically displayed and analyzed. By supplementing these primary data with each individual's data, you allow readers to independently analyze the data and draw their own conclusions regarding the extent of inter-participant replication.

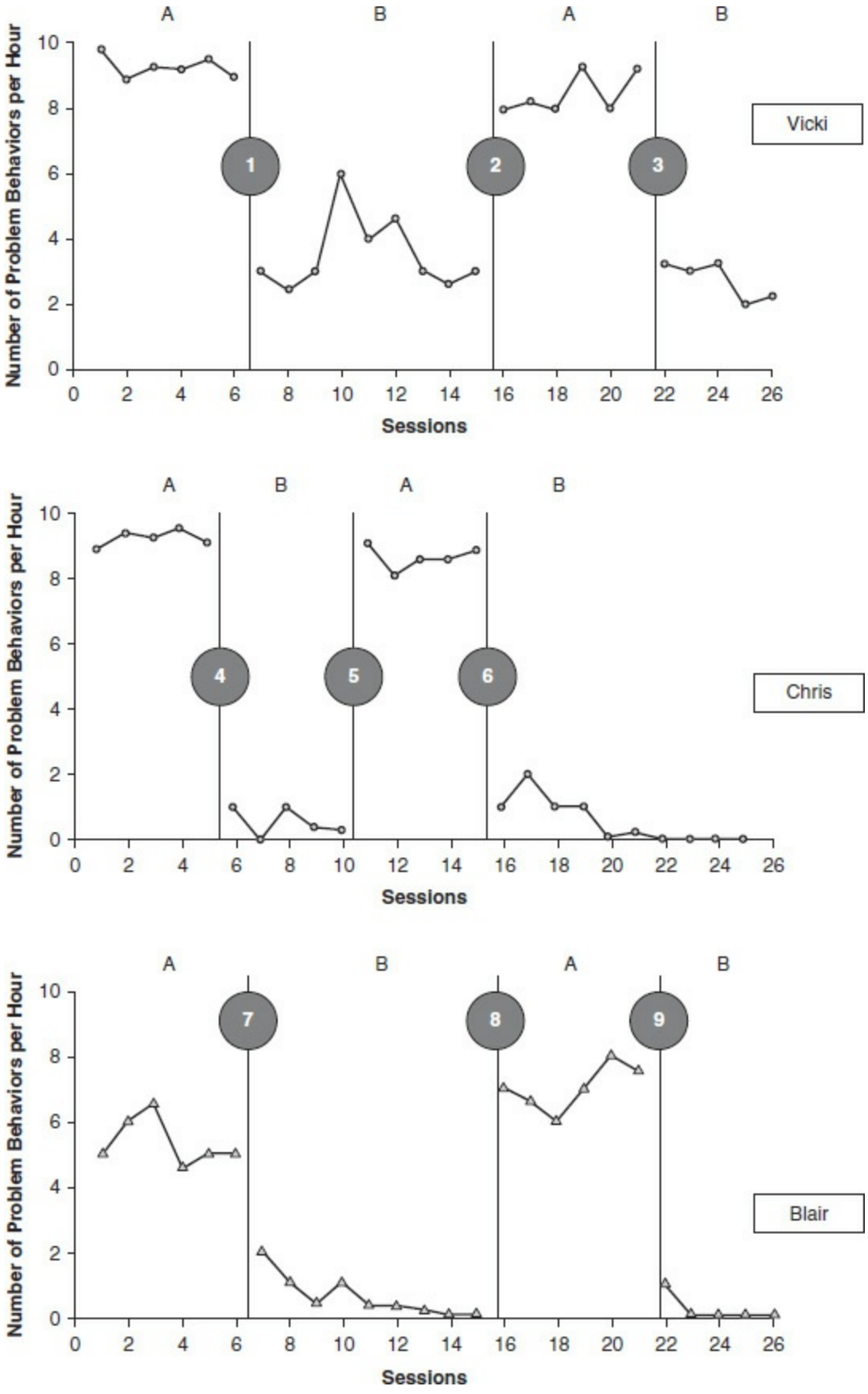


Figure 4.2a Three A-B-A-B designs showing three intra-participant replications for each participant (n=3), for a total of nine demonstrations of effect.

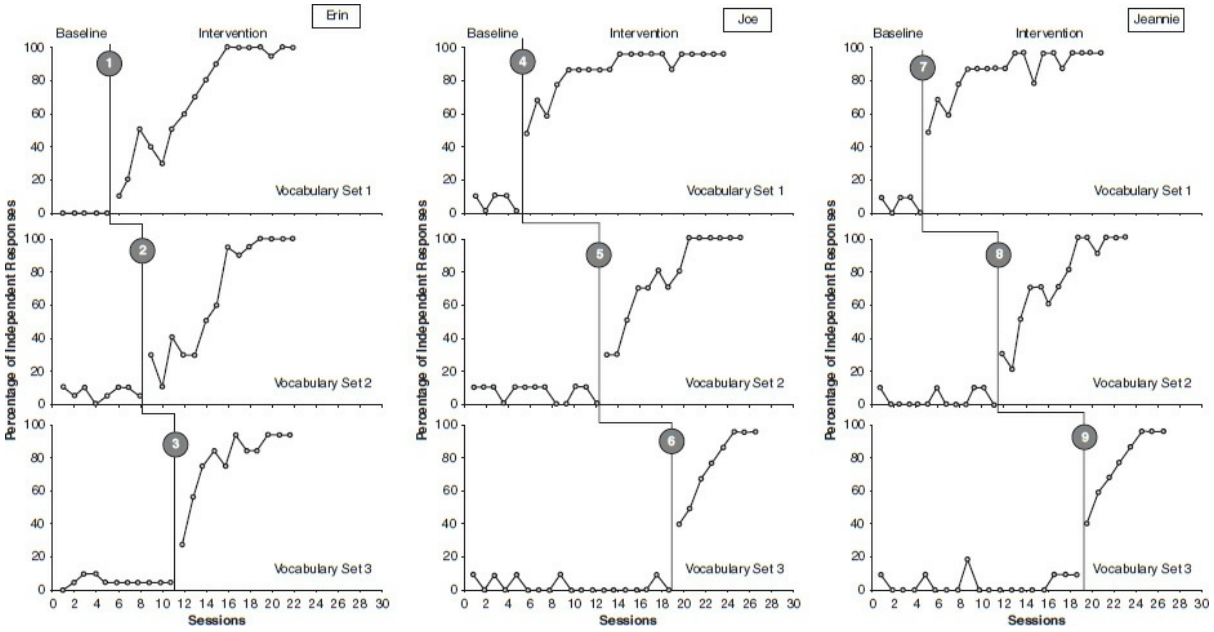


Figure 4.2b Three multiple baseline across behaviors designs showing three intra-participant replications for each participant (n=3), for a total of nine demonstrations of effect.

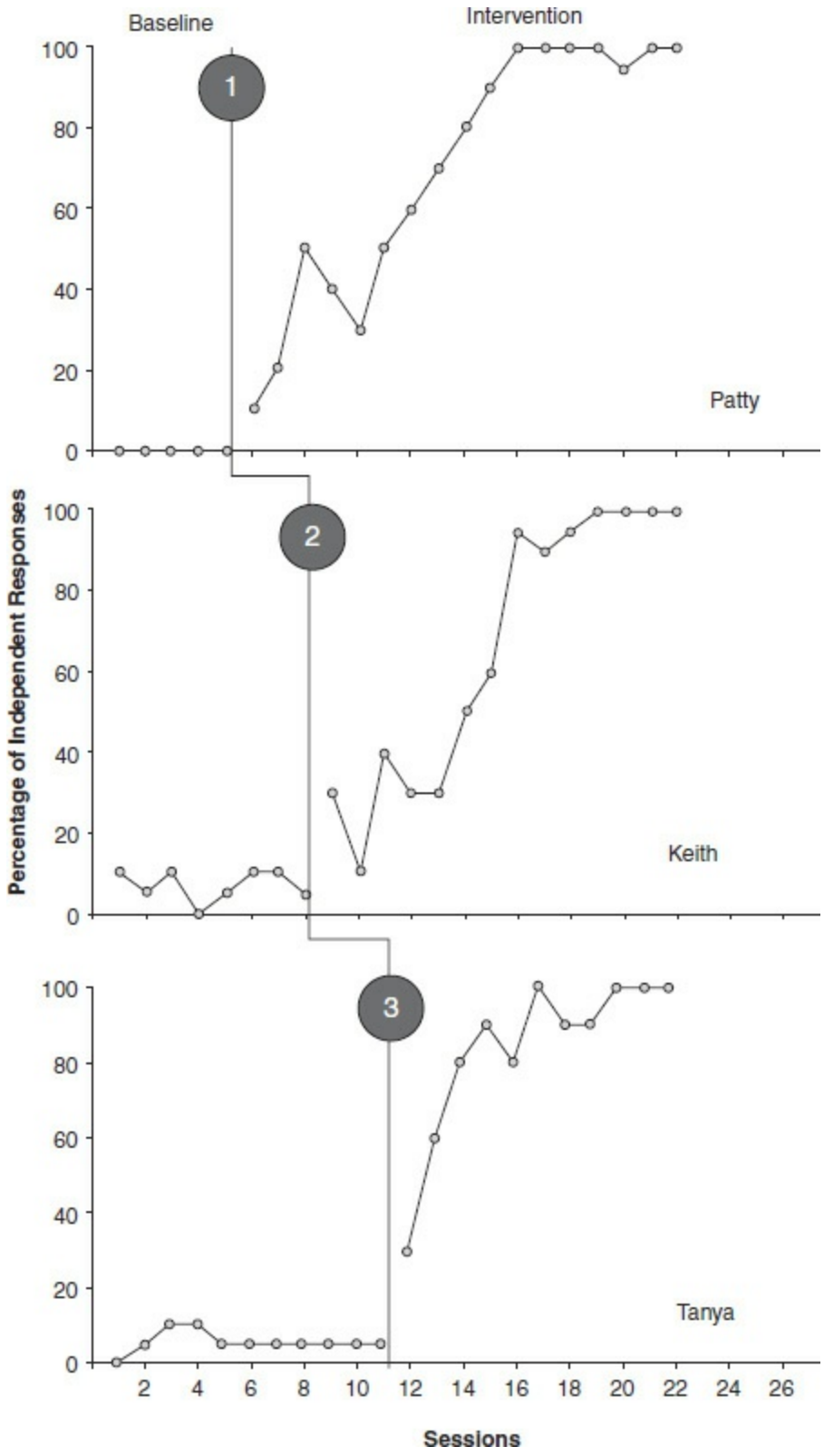


Figure 4.2c One multiple baseline across participants design showing three inter-participant replications with three participants, for a total of three demonstrations of effect.

Direct Replication Guidelines

The guidelines that follow are based on those presented by Barlow and Hersen (1984, p. 346) since they address direct replication in the context of applied behavior analysis research.

1. Investigator(s), setting(s), material(s), instructional arrangement(s), format(s) etc. should remain constant across replication attempts with the same participant and across participants in the same study (i.e., for intra- and inter-participant replication).
2. Dependent variable (target behavior and measure) should be similar across participants, but it need not be identical. For example, in a study in which you want to evaluate the effectiveness of a system of least prompts (SLP) procedure to teach chain task skills to three children with moderate intellectual disabilities, you may identify three different chain task skills for each of the three students. The SLP procedure must be the same across students and behaviors until the ineffectiveness of the original procedure occurs, at which time you may modify the original procedure. In the case of a study that addresses aberrant behaviors, it is recommended that behaviors serve the same function (for a discussion of functions of behavior, see Cooper et al., 2007) as identified through a Functional Behavior Assessment (FBA).
3. Participants should have similar abilities related to functional inclusion criteria (e.g., characteristics that may influence the effectiveness of the intervention). It is generally believed that replication failures are more likely when there are large differences between participants. Birnbrauer (1981) and Wolery and Ezell (1993) address this notion of individual characteristics (status variables) and their importance, or lack of importance, in predicting and evaluating when an intervention is likely to be effective. The topic of participant descriptions and matching study participants on the basis of status variables will be addressed later in this chapter. Suffice it to say here, that in educational and clinical research, the pool of possible study participants will likely be based on your teaching or clinical assignment. That is, you will likely be working with individuals within a certain age range, cognitive level etc., and it will be these individuals with whom you will conduct your study. It is important that you identify and report the similarities and differences between participants, especially in relation to functional similarities that are likely important to intervention success (e.g., an intervention to increase social play should likely include participants with similar play skills and social behaviors; their age, race, and diagnoses are likely less important).
4. The independent variable should be the same across participants unless progress toward the therapeutic or instructional objective stalls, at which time you may choose to modify the original intervention or replace it with a new intervention.
5. Three direct replications are generally considered the minimum acceptable number for determination of a functional relation; many published studies include both three intra- and three inter-participant replications.

When employing SCDs three successful inter-participant (or inter-group) replications are considered minimally acceptable before moving on to a systematic replication attempt. Variables you should consider in determining whether three replications are an adequate number include: (a) baseline data stability, (b) consistency of effect with related findings, (c) magnitude of effect, and (d) adequacy of controlling threats to internal validity. Mixed results will require additional replication attempts.

Clinical Replication

Hersen and Barlow (1976) introduced a third type of replication that they called "clinical replication," a form of direct replication, in which direct replication guidelines are followed. **Clinical replication**, as defined by Hersen and Barlow, refers to "the administration of a treatment package containing two or more distinct treatment procedures by the same investigator or group of investigators … administered in a specific setting to a series of clients presenting similar combinations of multiple behavioral and emotional problems, which usually cluster together" (p. 336). They refer to this as an advanced process, the end of years of research in "technique building." Their context was the clinical setting and their participants, individuals with many types of emotional and behavioral problems, thus the use of the term *clinical*, rather than *educational*, in their labeling this type of replication. Within this context we can observe their view of the scientific process as it relates to the field of clinical psychology. It is a three-stage process. First, a researcher working with a series of clients with a similar problem establishes that an intervention produces behavior change. This is direct replication. Next, in clinical replications the researcher (and associates), combining techniques, demonstrate the effectiveness of an intervention package with participants who demonstrate similar clusters of problem behaviors (e.g., children with autism). One example of this clinical replication process is apparent when comparing *focused intervention practices* and *comprehensive treatment models* (CTMs) for individuals with autism spectrum disorders (Odom, Boyd, Hall, & Hume, 2010). CTMs include many components, generally previously researched in isolation, shown to improve a specific behavior; in sum, CTMs are designed to improve a variety of behaviors across domains (e.g., communication, social behaviors, adaptive skills). Much of applied research today, whether clinical or educational in nature, is the study of a treatment or educational package. Although it is ideal to change only one variable at a time when moving from one experimental condition to the next (e.g., baseline to intervention), research conducted in community settings (schools, mental health clinics, therapeutic recreation programs) frequently investigate the effectiveness of intervention packages (e.g., video modeling, prompting, reinforcement; Smith et al., 2016). In such cases, at minimum, it is the responsibility of the applied researcher to identify and report all differences, procedural and otherwise, between experimental conditions. Only through such disclosure will it be possible to identify those variables that may have contributed to observed behavior changes. As discussed in Chapter 11, there are SCDs that can be used to evaluate the relative contribution, if any, of intervention package components.

Systematic Replication

Sidman (1960, p. 111) notes that the fundamental dictum of science is that all research participants be treated alike except for in regards to the independent variable; however, if adhered to, this rule would strangle "systematic replication as a primary method for establishing reliability and generality." He explains, "If the psychologist's experience has given him confidence in his techniques, he will choose systematic replication rather than direct replication as his tool for establishing reliability. Instead of simply repeating the experiment, he will use the data he has collected as a basis for performing new experiments and obtaining additional related data." He continues, "systematic replication demonstrates that the finding … can be observed under conditions different from those prevailing in the original experiment," and suggests that the experimenter's judgment (history) will dictate how far he can move from the original experiment. Systematic replication is a gamble, one that if successful, "will buy reliability, generality, and additional information" (p. 112).

What constitutes a **systematic replication** in applied research? When a researcher carries out a planned series of studies that incorporate systematic changes from one study to the next and identifies them as a replication series, a systematic replication clearly exists. If a researcher tries another researcher's procedure and states his intent to replicate, that is another instance (although this occurs less often; Lemons et al., 2016). Suppose that a researcher initiates a study based on current findings in an area, such as time delay transfer of stimulus control procedure, and develops an intervention that contains several elements of existing procedures. Is this an instance of systematic replication? It is at this point that the definition of systematic replication is in the mind of the beholder. Suppose the researcher combines elements of three time delay studies as a foundation for a new intervention. Then nothing is the same; we have a different researcher, different study participants, a different environment, and a different intervention. Some researchers may not consider this an instance of systematic replication. While there may be a link to previous research, there is no single common element. The situation is different if (a) the researcher sets out to replicate, (b) states an intent to replicate, (c) contacts the researcher whose work she wishes to replicate in order to verify correspondence with a published procedure, (d) carries out the study, and then (e) reports results that can be evaluated in relation to the original work. This is a more restricted definition than what Sidman offers. However, in the experimental laboratory serendipity plays a larger part than it does in the educational and clinical settings. While the basic researcher approaches a problem with the question "What will happen if… ?" the applied researcher, especially in classroom and therapy environments, will approach the problem of behavior change with the question, "How can I make X work?" or, as noted, "How can I do X better?" Or, "If X worked for someone else, how can I produce a more powerful effect?"

Systematic replication was defined by Hersen and Barlow (1976) as "any attempt to replicate findings from a direct replication series, varying settings, behavior change agents, behavior disorders, or any combination thereof" (p. 339). As pointed out by Tawney and Gast (1984), this definition presents some problems as it relates to their use of the word "series," in that their definition requires that systematic replication follow from a series of direct replication studies. This qualification, however, places a severe limitation on the definition (i.e., if systematic observation can only follow from a direct replication study, what is the status of studies designed to replicate another researcher's single study—one that has shown interesting and promising results?). This restriction notwithstanding, the phrase "any attempt to replicate" is, at the same time, perhaps too broad. To illustrate, Hersen and Barlow (1976) presented a table of systematic replication studies in the reinforcement of children's differential attention (pp. 346–349). These 55 studies were conducted by many investigators and were reported from 1959 through 1972. It is doubtful that these studies meet Hersen and Barlow's definition of systematic replication. Whether, collectively, they are systematic replication is a matter of personal opinion. Perhaps Jones' (1978) analysis of Hersen and Barlow (1976) will clarify the point:

> Replication is clearly a canon of applied behavioral science, and is discussed frequently, but executed less frequently. Absolutely pure replication probably seldom happens, if ever. Pure replication would require a point-by-point duplication of a research design, varying nothing except the time the study was conducted. Such replication is considered trivial by most researchers and may not be publishable. When behavioral interventions lead to large and dramatic effects, and there is no question about the experimental control demonstrated in the study, then such pure replication is trivial. But, when researchers change procedures (the inherent flexibility of single-case designs), plan to use the technique with different kinds of subjects in different settings, or anticipate changing any salient aspect of the design, then pure replication, of course, is impossible. Replication then becomes more a matter of repeating the work with systematic modifications. Modified procedures, subject populations, measurement systems, etc., are tested to see if comparable results occur. The value of replication in single-case experimentation occurs when there is a substantial accumulation of parallel or convergent findings from a set of similar, but not identical, procedures, techniques, measurement devices, subject samples, etc. In the end, convergence among results from many such replications determines the generality of findings. This is the big goal to be achieved by the field of applied behavior analysis.
>
> (p. 313)

Suffice it to say, systematic replication, as discussed by applied researchers today, is more broadly defined than the definition offered by Hersen and Barlow (1976), in that, (a) a systematic replication attempt may follow a single study, and (b) variations (i.e., systematic modifications) from the original study or studies are included in the definition and, in fact, are encouraged as means for extending the generality of experimental findings. On the topic of systematic replication Tawney and Gast (1984) wrote,

> systematic replication, as applied to research conducted in educational settings, is an attempt by a researcher to repeat his own procedure, employing variations in the procedure, with the same or different subjects. Or, it is a series of planned experiments, conducted by one researcher that utilizes the same basic procedure, but systematically varies it based on results of the first experiments. Or, it is an attempt by a researcher to reproduce the published findings of others, adhering closely to the original procedure.
>
> (pp. 97–98)

In writing this we considered our definition to reflect the reality that in classroom-based research, as in clinical research, very little is the same from day to day, and from study to study. It focuses on the goal of the researcher to repeat a procedure that has been successful (or at least seems promising). Or, viewed from the perspective of the scientist-practitioner, "If intervention X has been used effectively with students like mine, will it work with my students?" By asking such a question you will address the three purposes or goals of systematic replication: (a) demonstrate reliability of effect, (b) extend generality of findings, or (c) identify exceptions. Whatever the outcome of a systematic replication attempt, our understanding of the phenomenon being studied has been enhanced.

Failure to replicate can, and has, led to the discovery of limitations of current interventions and the discovery of new interventions. Regardless of whether a failure to replicate occurs within a direct or systematic replication attempt, the failure 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 successful and beneficial to the participant. It is not acceptable to simply note that there was a failure to replicate and move on. Baer, Wolf, and Risley (1968), in their description of applied behavior analysis, were clear in assigning behavior analysts the responsibility of ensuring that study participants, or society, benefit from research involvement. Thus, after a failure of an intervention to bring about the desired and expected therapeutic behavior change, the appropriate question you should ask is, "What modification can I make to the original intervention, to make it successful?" or, "What other intervention can I employ to bring about behavior change?" Failures should stimulate interest in why the failure occurred and what can be done to bring about success. Modification of the original intervention is advised as the first course of action, rather than abandoning the original intervention and replacing it with a new and different intervention. The likelihood of making the correct decision will be directly dependent on familiarity with the literature.

Systematic Replication Guidelines

Different applied researchers may suggest slightly different definitions of systematic replication (e.g., Hersen & Barlow, 1976, 1984; Jones, 1978; Tawney & Gast, 1984), however, general guidelines on when and how to proceed with a systematic replication attempt are quite similar.

1. Begin a systematic replication study when reliability of effect has been established through a direct replication study or series of studies. It doesn't matter whether the replication attempt follows a single study by one researcher or several studies conducted by several researchers over a number of years. The important factor in

deciding when to initiate a systematic replication of an earlier study is the belief that threats to internal validity in the original study were evaluated and controlled for, and the findings are accurate (reliable) and true (valid).

2. Identify and report the differences between the systematic replication attempt and the original study or studies; it is likely that many authors fail to identify when a study is a replication (Lemons et al., 2016). In the case of a replication attempt following a series of studies it is important to identify the number and types of differences (researcher or research team, participants, variations in the independent variable, dependent measures, experimental design etc.). Only through reporting these differences will we identify the extent to which earlier findings generalize and the potential reasons for failure to replicate. It is important to remember that generalization is not an "all or none" phenomenon, but a matter of degree across different variables. It is your responsibility to identify and report these variables after a successful systematic replication attempt.
3. After a failure to replicate, first modify the original intervention, and if necessary employ a different intervention to bring about the desired therapeutic or educational effect. Much is learned by failures to replicate if we can identify the cause of the failure and identify modifications or alternatives to the original independent variable. Surely, one participant's failure to respond as expected is so unique that other individuals won't respond in a similar fashion. Isn't that what is special about special education and clinical practice … interest in identifying procedural adaptations, accommodations, and alternatives to that which is considered the norm?
4. Systematic replication attempts are never over. In addition to strengthening the reliability and generality of findings, "systematic replication is essentially a search for exceptions" (Barlow & Hersen, 1984, p. 364), thus there is no predetermined time to stop, regardless of the number of studies that have successfully demonstrated the reliability and generality of effect. In Sidman's (1960) words, " a negative instance may just be around the corner" (p. 132).

Replication and External Validity

A common criticism directed at SCD research methodology always has been that findings can't generalize beyond the individual … there simply are too few participants in studies that employ SCDs. By contrast, group research methodology, which randomly assigns a large number of participants to two or more groups (one that serves as a control group and the other(s) experimental comparison group(s)), is considered the "gold standard" for establishing external validity. Few would argue that findings generated by large group research generalize better to other large unstudied groups, 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). At first glance these positions regarding research methodology and external validity seem to make sense, however, what if your interest is 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 respond to the independent variable. Their focus is on the group, not the individual. Sidman (1960) was clear in his position regarding the importance of reporting individual participant data and the reliability and generality of findings between inter-participant and inter-group replication when he wrote:

> Indeed, replication of an experiment with two subjects establishes greater generality for data among the individuals of a population than does replication with two groups of subjects whose individual data have been combined.
>
> (Sidman, 1960, p. 75)

In addition, the dynamic nature of SCD research may improve generality to clinical and educational contexts. For example, when an intervention condition is not successful in the context of an SCD, researchers modify or change the intervention until acceptable behavior change occurs. Thus, adaptable procedures may be more generalizable to contexts in which data-based modifications are likely (e.g., educational and clinical settings).

A final point regarding limitations of large group research is that intra-participant replications are seldom attempted. In a typical group research investigation individuals in the experimental group are exposed to the independent variable with no attempt to repeat its effect by either staggering its introduction across behaviors, or withdrawing and then re-introducing it to see if the effect can be repeated. Most behavior analysts would agree that intra-participant replication is a more convincing demonstration of

reliability than inter-participant replication, a design characteristic and limitation of many, if not most large group research designs, as well as some SCDs (multiple baseline and multiple probe designs across participants; see Chapter 10).

SCD research methodology has a long history in which the primary focus has been on the individual. Even when the focus of a research investigation has been on changing the behavior of a group, individual data of group members have been reported. There is a clear understanding among behavioral researchers that if your interest is in designing and implementing effective interventions for individuals, many of whom differ from the norm, it is imperative that you study the behavior of individuals. As previously discussed, direct intra-participant replication is the primary means by which SCD researchers establish the reliability of their findings and, to the extent that study participants differ, address the generality of their findings through direct inter-participant replication. External validity in SCD research is primarily accomplished through a series of systematic replication studies in which some characteristics (e.g., investigators, participants, settings) differ from previous studies and yield the same outcome. The question for you, as you attempt a systematic replication, or are considering using an intervention with a student or client is, "What individual characteristics or variables should I consider in determining the likelihood that the intervention under consideration will be successful?"

There are several variables that you may consider when attempting to determine the similarities and differences between research and "service" populations, the most common being status variables. Status variables are participant descriptors including gender, chronological age, ethnicity, intelligence quotient, academic achievement level, grade level, educational placement, and geographic location; these descriptors were considered "minimal" by the Research Committee of the Council for Learning disabilities (Rosenberg et al., 1992) when conducting studies with fewer than 10 participants. 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) hold 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). Through a brief review of constant time delay (CTD) research they concluded that in spite of consistent findings across several studies, procedural modifications were necessary even though participants "were nearly identical on status variables" and that the procedure's success was independent of status variables and was likely due to students' different learning histories" (p. 644). You need only look at published literature reviews and meta-analyses to appreciate the success of SCD research predicting and confirming the reliability and generality findings to other individuals. But if status variables are not the best predictors of generalization, what variables are?

SCD researchers support the position that external validity is directly related to

baseline condition performance, that predicting the effectiveness of an intervention will be determined by the similarities in response patterns by two individuals under the same or similar environmental conditions. Birnbrauer (1981) summarized the position when he wrote, "we should look for similarities in baseline conditions, the functional relations that appear to be operative during those pretreatment conditions, and the functional changes that implementation of treatment entailed for previous subjects. These are the keys to generalizing from single subject studies" (p. 129). This is not to say you should discount status variables when writing your research report. As discussed in Chapter 3, detailed participant descriptions are important, including reporting on status variables, but when it comes to predicting generalization success, emphasis should be placed on what Wolery and Ezell (1993, p. 645) termed functional variables (i.e., "the effects of specific environmental-participant interactions"). Specifically, you should describe the characteristics of the baseline condition (e.g., response contingencies, number of opportunities to respond) and the behavior patterns generated by your study participants to predict, with greater confidence, whether your intervention will or will not be effective. Prediction of inter-participant replication success, be it a direct or systematic replication attempt, is more about your attention to baseline condition data, experience with the independent variable, and visual analysis skills, than it is about matching participants on status variables. 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 (Juan 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 potential need for a focused academic intervention). Thus, information about 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. To determine what variables are critical, you must identify a theory of change for your independent variable (see Chapter 6).

Generalization Continuum

Generalization is not an all-or-none proposition. The generality of experimental findings are viewed along a continuum in which the number of variables that change between studies will determine the extent of generalization. Figure 4.3 illustrates the point: at the far left of the horizontal line is direct inter-participant replication. As discussed, few, if any, condition variables are changed between participants in the same investigation,

with the same investigator, conducted during the same time period. Study participants are different, but they tend to be similar in age, cognitive abilities, entry skills, need for behavioral intervention etc. On status variables they look quite similar. If you will, generalization is "close-in" and, therefore, limited to the degree to which participants differ, which often isn't very much. At the far right of the horizontal line is systematic replication with a multitude of differences from previous studies. Systematic replication, at the extreme, has a different investigator, different research site, different types of participants based on status variables, different target behavior or class of behaviors, different dependent measure(s), different SCDs, variation of the independent variable etc. The differences are many; the similarities are few. Except for the independent variable being "similar" to independent variables studied in previous investigations, it is close to being considered a novel study. At these two extremes the risks of replication are quite different, direct inter- participant replications are much less of a gamble than are systematic replications in which numerous variables are changed from earlier studies. Most replication attempts, however, fall somewhere between these two extremes. The number and types of variables that are changed between separate studies will determine the degree of risk and the extent of generalization. For this reason it is imperative that researchers delineate each and every difference between their study and those that preceded it.

"N of 1" Single Case Studies and Their Contribution

Some journals publish SCD research investigations that have been conducted with only one participant, although this may be decreasing over time. They are truly "N=1" studies and, thus, in and of themselves contribute little to the external validity of the independent variable under study. Their "stand alone" contribution is a quantitative evaluation and demonstration of intervention effectiveness, in which threats to internal validity have been adequately evaluated and controlled for through direct intra-participant replication, thereby lending support for the intervention that addressed a novel or rare challenging behavior. Through the publication of such research reports, systematic replication is encouraged, which in turn will increase our understanding of the reliability and generality of the intervention. However, as a consumer of research, you should proceed with caution before implementing an intervention with your client or student that has been conducted with only one individual. As an applied researcher you are encouraged to attempt to replicate the effect. It is important not to discount the findings of these studies but you need to understand their limitations and need for replication to build confidence in their findings. On this topic, Sidman (1960) wrote:

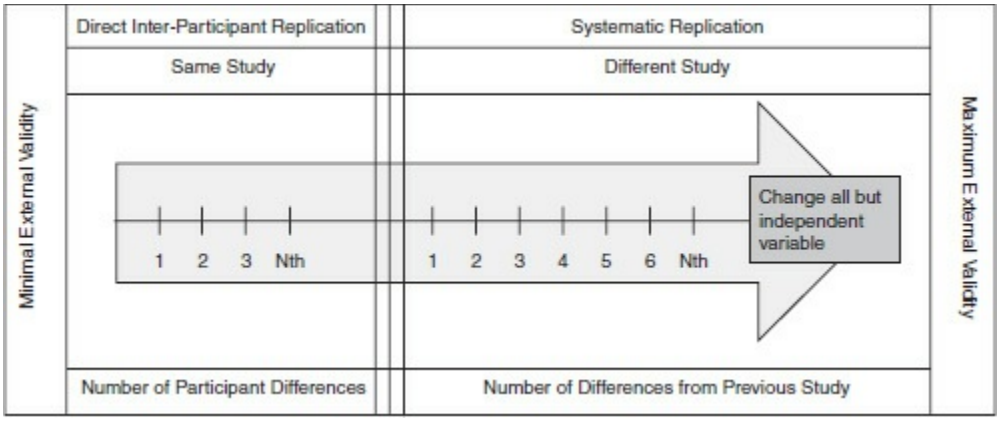


Figure 4.3 External validity continuum.

> Often, especially in a young science, an experiment is performed for the sole purpose of determining if it is possible to obtain a certain phenomenon. In such an experiment, demonstration of the phenomenon in one organism, with reliability established by intra-subject replication is all that is necessary. Such studies are the impetus for further research.
>
> (p. 93).

General Recommendations for Starting a Systematic Replication

If you wish to initiate a systematic replication attempt, here are a few suggestions on how to proceed:

1. Read studies that relate to your research interest(s) or question(s). Look for recently published literature reviews and meta-analyses on your topic as they may provide you with a comprehensive reference list of empirical investigations that addressed the same or similar research question(s).
2. Develop two tables, one that identifies similar elements, and the other dissimilar elements, across studies you identified.
3. Analyze the data entered in these tables and identify the similarities and differences between studies.
4. Read and list researchers' suggestions for future research on the topic. These are commonly found in the discussion section of research reports.
5. Write your research question(s), if you haven't already, taking into account researchers' suggestions and the practical constraints of your setting (e.g., access to participants, daily schedule, availability to materials and equipment, control of contingencies etc.).
6. Explicitly state that the study is a replication attempt, and report the specific differences between your proposed study and those that have preceded it.

Once your study begins you should note whether the effects of the independent variable were replicated with all or only some participants. Regardless of whether your replication attempt was a "success" or "failure," your ability to identify the differences between participants who responded positively to the intervention, and those who did not, is important. In cases of "failure to replicate," your ability to implement a successful variation of, or alternative to the original intervention, will advance our understanding of the reliability, generality, and limitations of the intervention. This contribution is further expanded when you analyze your findings, including the functional and status variables associated with intervention success or failure, with participants in earlier studies.

Summary

Replication is essential for both internal and external validity. Through direct intra-participant replication the reliability of research findings is established. By including multiple participants in the same study an investigator extends the generality of findings to the extent that participants differ on both status variables and functional variables. In SCD research the generality of research findings is primarily established through systematic replication, a series of studies conducted over several years in which the investigator, target population, behavior, dependent measures etc. differ from earlier studies. Systematic replication is ongoing, never over, as a failure to replicate may be just around the corner. When a "failure to replicate" is evidenced and an exception to previous research findings identified, the limitation of the intervention is revealed. Applied behavioral researchers approach such failures as a challenge and attempt to identify their cause, as well as 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

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.

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

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

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

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

Jones, R. R. (1978). Invited book review of *Single-case experimental designs: Strategies for studying behavior change* by M. Hersen and D. H. Barlow. *Journal of Applied Behavior Analysis*, *11*, 309–313.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from What Works Clearinghouse http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Lemons, C. J., King, S. A., Davidson, K. A., Berryessa, T. L., Gajjar, S. A., & Sacks, L. H. (2016). An inadvertent concurrent replication: Same roadmap, different journey. *Remedial and Special Education*, *37*, 213–222.

Odom, S. L., Boyd, B. A., Hall, L. J., & Hume, K. (2010). Evaluation of comprehensive treatment models for individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *40*, 425–436.

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.

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

Smith, K. A., Ayres, K. A., Alexander, J., Ledford, J. R., Shepley, C., & Shepley, S. B. (2016). Initiation and generalization of self-instructional skills in adolescents with autism and intellectual disability. *Journal of Autism and Developmental Disabilities*, *46*, 1196–1209.

Stinson, D. M., Gast, D. L., Wolery, M., & Collins, B. (1991). Acquisition of nontargeted information during small group instruction. *Exceptionality*, *2*, 65–80.

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

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

5
Dependent Variables, Measurement, and Reliability

Jennifer R. Ledford, Justin D. Lane, and David L. Gast

Important Terms

reversible, non-reversible, continuous recording, non-continuous recording, onset, offset, count, duration, latency, inter-response time, event recording, timed event recording, free-operant, trial based, partial interval recording, whole interval recording, momentary time sampling, construct validity, observer drift, observer bias, blind observer, interobserver agreement, discrepancy discussion, occurrence agreement, non-occurrence agreement, gross agreement

Measurement may be defined as the systematic and objective quantification of objects, events, or behaviors according to a set of rules. In the next two chapters, we discuss measurement of dependent and independent variables, with a focus on ensuring the

accuracy, believability, and meaningfulness for both procedures and targeted behaviors.

Choosing, Defining, and Characterizing Behaviors

Choosing Behaviors

As an applied researcher what you decide to measure will depend directly on your research question and objective. Several sources are available to help you determine what to measure. In addition to using personal observations, you can consult with significant others (e.g., parents, teachers and therapists, psychologists), and examine previous assessments. You can also consult a current individual education program (IEP), individual family service plan (IFSP), or treatment plan. Often in applied research, there is an apparent problem that needs to be solved for an individual (e.g., a scientist-practitioner has a client who has reported a specific need) or a population of individuals (e.g., a review of the extant research shows inadequate research support for the use of social narrative interventions for young children without autism; Zimmerman & Ledford, 2017).

After choosing a target behavior, you must determine what dimension of the behavior is of interest (Barlow & Hersen, 1984). There are two primary dimensions: time and number. For example, you may be interested in reducing the number of tantrums a child engages in during each school day, or you may be interested in reducing the amount of time in which a child engages in tantrums. Similarly, for the child who displays tantrums, you may want to increase the number of prosocial interactions with peers and simultaneously increase the duration of appropriate play during the school day. Often, but not always, a change in time or count results in a corresponding change in the other dimension. The procedures, difficulties, and benefits of measuring and estimating each are different; thus, it is important to carefully select the dimension of interest before defining behavior occurrence and choosing measurement procedures.

Defining Behaviors

In accordance with the behavioral approach to teaching and clinical practice, you should define target behaviors in observable and measurable terms (Barlow & Hersen, 1984). Rather than define a child's behavior using ambiguous global terms, such as "disruptive," "bored," or "passive," describe the behavior(s) in specific terms. For example, if a teacher frequently has observed a student leaving his desk without permission, talking with classmates during class presentations, and dropping pencils and books, you would have a much clearer idea as to what the teacher considers disruptive behavior. When writing operational definitions for behaviors, you should also provide examples and non-examples to ensure that all relevant behaviors are coded and that all non-relevant behaviors are not. Examples and non-examples should include close examples and close non-examples (what Barlow and Hersen called "questionable instances," p. 112). For

example, disruption might include *being more than 1 meter away from desk for at least 3 seconds* and a close non-example of the behavior would be *leaving the desk area within 10 seconds of a teacher instruction or permission to do so.* With greater behavioral specificity and clarity you can more effectively document observations and communicate them to participants, readers, or other stakeholders such as parents. Examples and non-examples should be written and used to clarify, rather than as exhaustive lists. See Table 5.1 for several examples of definitions, examples, and non-examples used in a study designed to assess the effects of a playground-based intervention on physical activity behaviors (Ledford, Lane, Shepley, & Kroll, 2016).

Characterizing Behaviors

For the purposes of measuring dependent variables in the context of a single case design (SCD) study, you will need to decide whether they are reversible or non-reversible (not readily reversible). **Reversible** dependent variables are those behaviors that are likely to revert to baseline levels if an intervention is removed. Examples may include problem behaviors like aggression, on-task behavior, active student responding, and social interactions. Changes in **non-reversible** dependent variables are not truly permanent, but these changes may be likely to maintain in the absence of an intervention condition. Examples may include most academic behaviors (e.g., sight word reading, picture-naming), some functional behaviors (e.g., learning how to use an iPod to access games), and motor behaviors (e.g., learning how to ride a bike). Appropriate designs that may be used for reversible and non-reversible behaviors are shown in Table 5.2.

Table 5.1 Example Coding Definitions, Examples, and Non-Examples.

Code	Definition	Examples	Non-examples
Social Interaction	Verbal or non-verbal initiations or responses that are directed toward a peer and that are neutral or positive in nature	Calling a peer's name Responding to a peer initiation by looking Responding to a peer request to give an item Calling multiple peers at once (e.g., Hey everyone!)	Any interactions directed to an adult Any negative interaction (aggression, threats, and other actions or words considered "not nice" by classroom staff such as "shut up" or "I

			hate you")
Engagement	Appropriately playing with materials or peers or engaging in purposeful physical activity	Playing chase Playing with a bat to hit a ball Running toward the slide	Wandering Walking in a repetitive sequence Sitting at the top of the slide for more than 2 s
Proximal Play	(a) Being within 5 ft of another child while playing with the same materials or activities *and* (b) either oriented to the same object/action/direction or oriented toward each other	Standing next to a peer, both watching bubbles Rolling a ball to a peer Playing on the same structure (if within 5 ft)	Any behavior while swinging Playing on opposite sides of the same structure Running more than 5 ft apart

Ledford, J. R., Lane, J. D., Shepley, C., & Kroll, S. (2016). Using teacher-implemented playground interventions to increase engagement, social behaviors, and physical activity for young children with autism. *Focus on Autism and Other Developmental Disabilities*, *31*, 163–173.

Table 5.2 Design Types to Be Used When Demonstrating Efficacy or Comparing Interventions, for Reversible and Non-Reversible Behaviors.

	Reversible	**Non-reversible**
Demonstration	A-B-A-B Multiple baseline Changing criterion	Multiple probe
Comparative	Alternating treatments Multitreatment Simultaneous treatments Multielement	Adapted alternating treatments Parallel treatments Repeated acquisition

In addition to characterizing behaviors according to reversibility, you should also characterize your behaviors of interest according to whether they occur briefly or for at least a few seconds at a time. Some behaviors last a very brief (trivial) amount of time, such as hitting or scratching peers, cursing, imitating a child's utterance, choosing a response from a field of four by pointing, or responding to a multiple choice question. That is, they occur for less than a second, and the time it takes for them to occur is

generally not of interest. We will refer to these behaviors as *short duration* behaviors or *behaviors of trivial duration* (Yoder, Ledford, Harbison, & Tapp, 2017). Other behaviors tend to last for at least a few seconds at a time. Examples of these *long duration* behaviors include off-task behavior, tantrum behavior, engagement, parallel play, and physical activity. Some behaviors may be short duration or long duration behaviors depending on the context—for example, measuring conversational turns for a 3-year-old with autism and limited verbal skills (short duration) versus measuring conversational turns for typically developing teenage participants (long duration) or measuring correct responses to sight words (short duration) versus measuring how long it takes a child to read a given passage (long duration). Once you have determined the type of behavior you are interested in measuring, you can select a data recording procedure.

Selecting a Data Recording Procedure

After identifying the behavior to be measured, defining it in observable terms, and determining whether it is reversible, you must decide on a method for quantifying the behavior. There are a variety of recording procedures available to SCD researchers, each with its own advantages and disadvantages. You must decide the behavior characteristic that deserves attention (e.g., how often it occurs, how long it lasts, or percentage of opportunities for which it is done correctly) and then select a recording procedure that will capture the characteristic of interest, is feasible for use, and can be used accurately. Variables that require consideration include the (a) target behavior, (b) objective of the intervention program, (c) practical constraints of the setting(s) in which the behavior is to be measured, and (d) sensitivity to document behavior change.

The most common type of dependent variable assessment in SCD research is direct, systematic observation and recording (DSOR). That is, humans watch their participants and measure what they do, in a rule-bound and systematic fashion (Wolery & Ledford, 2013). We will spend the remainder of the chapter focusing on DSOR, but two additional methods for measuring behavior are worth noting. First, the use of automated recording devices, including bio-behavioral records like electroencephalography (EEG; cf. Au et al., 2014) and physical activity trackers (Ledford et al., 2016), may become more common as these measures become pervasive in practice and feasible for use (e.g., as costs decrease). Additional research is needed to determine to what extent these measures correlate with observed behavior, but the decreased resource needs for human data collection make automated measurement appealing. Secondly, permanent products are sometimes used to measure behaviors, particularly related to acquisition of academic skills (Tawney & Gast, 1984). For example, without watching a child perform the task, you could assign an accuracy score to a math test. This permanent product is typical in educational and clinical settings, but is less common in SCD research, in part because of the risk of testing effects (see Chapter 1). Nevertheless, it is a reasonable and feasible option when measuring behaviors of participants that result in a product.

When using DSOR, there are two options when recording behavior occurrence: continuous recording and non-continuous recording (Johnson & Pennypacker, 2009). Continuous recording *quantifies* the occurrence of behavior; and non-continuous recording *estimates* the occurrence. **Continuous recording** requires counting or timing each behavior occurrence. For example, you might tally the number of words a child correctly reads (count) or time how long it takes her to read a passage of a given length (time). **Non-continuous recording** involves *sampling* behavior occurrence in order to estimate the actual count or time. Generally, non-continuous recording involves selecting an interval length (e.g., 30 seconds), determining the rules to code whether or not a behavior occurrence is scored for the interval, and using the rules to estimate behavior occurrence. Continuous recording is generally superior to non-continuous

recording, since it does not rely on behavior sampling, which can introduce error. However, continuous recording may prove to be infeasible or prohibitively resource intensive (e.g., a teacher in a classroom may not be able to time the duration of a tantrum while engaging in teaching tasks) or too difficult (e.g., it may be difficult to define on-task behavior in a way in which observers can accurately identify the **onset** and **offset** of the behavior). The onset of a behavior is the moment a behavior starts to occur, and the offset is the moment it stops.

Before choosing a procedure, you should identify the dimension of the dependent variable (i.e., target behavior) that is of interest. The two most commonly measured dimensions are *time* and *number*. If the primary interest is number, the measurement system will be based on **count** (the number of times a behavior occurs). Time-related measures include **duration** (amount of time for which the behavior occurs, or the time between the onset and the offset; Johnson & Pennypacker, 2009; Wolery & Ledford, 2013), **latency** (amount of time between a signal or cue and the onset of the target behavior; Johnson & Pennypacker, 2009; Wolery & Ledford, 2013), and **inter-response time** (amount of time that passes between the offset of a behavior and the onset of the next behavior occurrence; Johnson & Pennypacker, 2009). See Table 5.3 for examples of the use of count, duration, and latency measures in applied research. Inter-response time is rarely used as a dependent variable, although it is sometimes used in behavioral definitions (e.g., a new occurrence is counted if the onset of the behavior is more than 2 seconds from the offset of a previous occurrence).

Event and Timed Event Recording to Measure Count

Perhaps the simplest option for measuring behavior is to count the number of times it happens; this is an intuitive metric and one often used in typical non-research settings (e.g., counting the number of correct responses on a test, number of social interactions, or number of discipline referrals for a child). When using count, you must attend to (a) carefully defining a behavior in such a way that two independent observers can agree whether a potential instance of a behavior should be recorded, and (b) under what conditions a new occurrence happens. As previously described, careful consideration of examples and non-examples will assist with the first task of defining the behavior. The conditions for a new occurrence may be simple (e.g., each successive hit counts as an occurrence of self-injurious behavior; each item correctly answered on a worksheet), but are sometimes more complicated (e.g., two statements count as two separate social interactions if they are separated by at least 2 seconds in time *or* if they are separated by a related peer response).

Table 5.3 Examples of Use of Count, Duration, and Latency Measures in Applied Research.

Citation	Behavior	Recording System	DV
Measuring Count			

Kamps et al., 2014	Communicative acts	Free operant timed event recording	Number per session
Shepley, Lane, & Shepley, 2016	Correctly labeling actions	Trial-based event recording	Percentage correct
Chazin, Bartelmay, Lambert, & Houchins-Juarez, 2017	Correctly completing steps for cooking task	Trial-based event recording	Percentage of steps correctly performed
Sutherland, Alder, & Gunter, 2003	Opportunities to respond, correct responses, disruptive behaviors	Free operant event recording	Number (rate) per minute
Measuring Time			
Leatherby, Gast, Wolery, & Collins, 1992	Switch activation for toy access	Duration per occurrence	Number of seconds + Number of occurrences
Green et al., 2013	Peer interaction	Total duration	Number of seconds
Kamps, Conklin, & Wills, 2015	On-task behavior	Total duration	Percentage of session
Wehby & Hollahan, 2000	Compliance with low-probability demand	Latency per occurrence	Seconds to compliance
Estimating Count			
Zimmerman, Ledford, & Barton, 2017	Problem behaviors	Partial interval recording (10 s)	Estimated number per session
Estimating Time			
Reichow, Barton, Good, & Wolery, 2009	Engagement, problem behavior	Momentary time sampling (10 s)	Percentage of intervals
Luke, Vail, & Ayres, 2014	On-task behavior	Momentary time sampling (15 s)	Percentage of intervals

Event Recording Versus Timed Event Recording

The simplest way to measure events is by denoting how many occur (i.e., a tally); this is referred to as **event recording** (Tawney & Gast, 1984). This procedure and variations discussed below are appropriate when you are interested in *number* rather than *time*. A more sophisticated measure, **timed event recording**, involves denoting that an event has occurred *and* noting the time of the event (Yoder & Symons, 2010). Electronic data collection applications have made this type of recording, which was historically rare, more common. Specifically, the ease of video recording and/or the use of electronic data collection applications make timed event recording feasible for use in many research studies. The benefits of timed event recording include: (1) information about timing of behaviors may be important (e.g., if a child engages in challenging behavior near the beginning of each session but not late in the session, this may indicate that the child might benefit from a contingency review prior to the session) and (2) more precise agreement calculations are possible (Yoder & Symons, 2010; see "Ensuring Reliability of Data Collection").

Another variation of event recording can be used when timed event recording is not possible. In this case, researchers can use event recording, but can "group" events based on time. To do this, you (a) determine the smallest period of time that is feasible for measurement (e.g., 10-second or 1-minute intervals); (b) set a timer or other device to alert the data collector at regular intervals; (c) count the number of occurrences that occur between alerts (e.g., from timer start to 1 min, from 1:01 to 2 min, etc.). This data collection allows for more precision than using event recording alone, but less precision than using timed event recording. Using intervals to divide counts *does not constitute using non-continuous, interval-based recording*. It is simply a strategy used to improve the precision of the data collection—for example, similar to timed event recording, using event recording within intervals allows you to identify the temporal characteristics of the behavior (i.e., at approximately when they occur) and offers superior evaluation of agreement between raters. When this variation is used, the total number of occurrences is reported (cf. Barton, Pribble, & Chen, 2013).

Trial-Based and Free-Operant Events

Some behaviors can occur at any time during a measurement occasion (e.g., number of social initiations during free play), while others are dependent on specific antecedent events (e.g., number of correct responses on a word-reading task). Events that are free to occur at any time are referred to as **free-operant** events (Yoder & Symons, 2010); we will refer to events that have specific antecedent conditions (e.g., researcher task direction, peer initiation) as **trial-based** events. Generally, event recording can be used when trial-based events are of interest, because there is an anchor for each event (e.g., the first response is associated with the presentation of the first word). For each opportunity, an occurrence or non-occurrence is usually recorded (see below for information on exhaustive coding). Free-operant events can be more difficult to measure because there is no specification regarding when a behavior should occur; when using event recording

for free-operant behaviors, only responses (not non-responses) are recorded. Because free-operant responses can be more difficult to measure using event recording, researchers often use interval-based systems to estimate behavior occurrence (see *Estimating Duration and Count With Interval-Based Systems* section in this chapter).

Transforming Count

When count is used to measure behavior occurrence, it can be transformed for data presentation for ease of comparison between measurement occasions. Specifically, it is often transformed into a percentage or a rate.

Percentage

Trial-based counts are often transformed to a percentage of opportunities. For example, authors might report a percentage of trials during which a student independently and correctly responded to a query related to multiplication facts or a percentage of words read correctly in a reading passage. When differences among measurement occasions exist (e.g., the number of words in passages vary), using percentages allow for fair comparisons between sessions. In addition, percentage is often used and well understood outside of research contexts (e.g., is the basis of grades received in school, restaurant health scores). Using a percentage also facilitates comprehension because there is less need to understand context (Cooper, 1981; Gentry & Haring, 1976). For example, if the number of correct responses were reported on a graph as 10 (count), the reader would need to determine the maximum number of correct responses (e.g., scores of 10/10 and 10/20 are quite different). Percentage is calculated as the number of behaviors (or number of correct behaviors) divided by the total number of opportunities or trials, multiplied by 100. Note that free-operant behaviors without discriminative stimuli (i.e., cues that the behavior should occur) cannot be transformed into percentages.

Rate

Free-operant behaviors can be reported as a simple count, but if the measurement occasions differ in length, they are often converted to rate. Rate refers to the number of occurrences measured within a specific period of time. For example, you might report number of words read per minute or number of problem behaviors per hour. As with percentage, even when the measurement occasion is consistent in duration, rate facilitates quick understanding regardless of session length (Gentry & Haring, 1976). Rate is calculated as number of occurrences divided by duration of the measurement occasion (e.g., session); if 11 problem behaviors occurred during a 5-minute session, the reported rate would be 2.2 problem behaviors per minute. Trial-based behaviors should not be reported using rates because a non-participant (i.e., researcher, implementer, peer)

controls the rate of trial presentation, which constrains the rate of responding.

Duration and Latency Recording to Measure Time

Sometimes, the number of times a particular behavior occurs is less important than the amount of time for which it occurs. For example, suppose two children, Lauren and Andrew, were both on task 3 times during a short math activity. Without knowing the *duration* of the on-task behavior, knowing the number is relatively unhelpful (e.g., Lauren may have been on task for three 1-minute intervals; Andrew may have been on task for three 5-minute intervals). Whether the interest is duration or latency, there are two options for measuring time: time per occurrence and total time.

Time per Occurrence

Time per occurrence is measured by using a timing device to count the number of seconds of occurrence for each instance of the behavior. Historically, time per occurrence was unwieldy because for each behavior occurrence, researchers needed to start a timer at the onset of the behavior, stop the timer at the offset of the behavior, and record the time. Applications for electronic devices available for free or at low cost makes time per occurrence relatively simple to record. For some applications, for example, you can toggle a code "on" at the behavior onset and toggle it "off" when the behavior is discontinued; the program itself calculates the number of seconds per occurrence (e.g., Countee application for iPhone). Whether collected by hand or via an electronic device, time per occurrence data yields a number of potentially useful statistics: number of occurrences, average duration per occurrence, and total duration.

Total Time

Total time recording involves starting a timing device at each behavior onset and stopping the timing device at each behavior offset, without recording the time for each occurrence. At the end of a measurement occasion (e.g., session, class period), the total time is recorded. Unlike time per occurrence, no information is available regarding the number of occurrences or mean time per occurrence. However, especially if electronic recording devices are not feasible or available, this method is sufficient for determining the overall amount of time for which a behavior occurs.

Transforming Duration

As with count, duration measures can be, and often are, transformed into percentage statistics. You can calculate percentage by dividing the number of seconds of behavior occurrence by the total number of seconds in a measurement occasion (e.g., 600 seconds

in a 10-minute session) and multiplying by 100. Thus, if 60 seconds of off-task behaviors occurred in a 10-minute session, you could report that it occurred for 10% of the session ([60/600]×100).

Estimating Count and Duration With Interval-Based Systems

Although it is often possible to directly measure number and time variables, it is sometimes difficult or infeasible, especially in applied contexts. In these cases, researchers often choose to *estimate* behavior occurrence using interval-based systems with the assumption that estimation systems parallel a continuous measure of behavior, representing an approximation of the true value of a given behavior in context. These non-continuous recording systems all involve use of pre-determined intervals and systematic rules for counting occurrences within intervals (Powell, Martindale, & Kulp, 1975). Across interval-based systems, intervals tend to be between 5 and 30 seconds in length (Lane & Ledford, 2014); in general, you should use the shortest interval that is feasible given resource constraints. When using these systems, an interval timer (i.e., a timing device that will provide a notification on a regular schedule) is needed; many are available for electronic devices (e.g., Interval Timer; Deltaworks, 2016; Simple Interval Timer, Kazarova, 2017). Physical interval timers are also available from sporting goods stores and online retailers (e.g., GymBoss®, MotivAider®).

We describe interval systems below without including a separate "record" interval (Barlow & Hersen, 1984); that is, you record behaviors as occurring for one interval as the next interval starts, without taking a break. Separate record intervals can also be used such that, for example, you record an occurrence or non-occurrence at the end of the first interval during a 5-second break, before you begin the second interval. These record intervals have been used somewhat often and may be most useful when recording in-situ. When using this variation, less data are available than when using interval based systems with no record interval; of course, *less data* are preferable to *inaccurate recording*, so these intervals should be used when they are necessary for accuracy.

We caution researchers to only use these non-continuous systems if continuous measurement is not possible or feasible, since all systems are associated with estimation error (e.g., estimating time or number using these systems results in reliably *inaccurate* measurement). If you must use one of these systems, follow the recommendations below to ensure you choose the best system for estimating the dimension of interest, choose reasonable parameters, and make necessary corrections to improve estimations. For all interval systems, an estimated *count* should be reported when *number* is the dimension of interest (e.g., number of intervals in which the behavior occurred estimates number of occurrences) and *percentage of intervals* should be reported when *duration* is the dimension of interest (e.g., percentage of intervals in which the behavior occurred estimates percentage of time the behavior occurred). We note that authors almost exclusively report percentage of intervals, even when behaviors of interest are of trivial duration (and thus, researchers are unlikely to be interested in duration). Below, we

describe procedures, weaknesses, and recommendations for each of the three interval-based systems; following, we describe problems associated with the use of interval-based systems.

Partial Interval Recording

Partial interval recording (PIR) is the most widely used interval-based system (Lane & Ledford, 2014; Lloyd, Weaver, & Staubitz, 2016; Mudford, Taylor, & Martin, 2009). When PIR is used, the observer (data collector) records an occurrence if the target behavior occurs at any time during the interval. Thus, a behavior is recorded as occurring in the interval regardless of whether it occurred for the whole interval or for a very small part of the interval and whether the behavior occurs once or many times during the interval.

Benefits and Weaknesses

Benefits of PIR include ease of use and historical precedent. As mentioned above, PIR has been widely used in behavioral sciences for estimating behavior occurrence for more than 40 years. In addition, PIR may be easier to use than continuous recording because once a behavior has occurred for an interval, additional observation is extraneous and behavior is only recorded once per interval regardless of the number of occurrences. Serious weaknesses of PIR include inaccurate estimates of both count and duration and the need for very small interval lengths and statistical corrections to minimize these shortcomings.

Steps for use of PIR

If you decide to use PIR to estimate count or duration, we advise you to follow these guidelines:

1. Operationally define behavior occurrence.
2. Choose an interval length that is as short as is feasible given measurement and resource constraints (e.g., 5 seconds).
3. Set up a data collection system that allows for coding of a behavior occurrence (or non-occurrence) during each interval.
4. Set an interval timer to alert you via alarm or vibration at the end of each interval.
5. Record occurrences and non-occurrences:
    a. Record a behavior occurrence if the behavior occurs *at any time during the interval.* Only record one occurrence per interval, regardless of the number of times the behavior occurs.
    b. Record a behavior non-occurrence if the behavior does not occur at all during the interval.

6. Following session completion, summarize the data:
    a. If you are interested in *number*, count the number of intervals in which the behavior occurs. Use the Poisson correction to reduce error (see below). Report this number as an *estimated count*.
    b. If you are interested in *time*, count the number of intervals in which the behavior occurs and divide that number by the total number of intervals to get a percentage of intervals in which the behavior occurred. Report this percentage as an *estimated duration*.

Suggestions for use. We suggest use of PIR systems when (a) it is feasible to use small interval lengths, (b) the behavior is of short duration, (c) the dimension of interest is count, and (d) it is reported as an estimated count rather than percentage of intervals. Two variations are possible with PIR: counting across intervals and counting onset only. The first is the historical procedure, in which any behavior occurring across intervals is counted in both. The second is preferable; in this variation, count only behavior onsets (e.g., an occurrence is counted if the onset of the behavior occurs during the interval).

Whole Interval Recording

Whole interval recording (WIR) is the least widely used interval-based recording system (Lane & Ledford, 2014; Lloyd et al., 2016; Mudford et al., 2009), perhaps given the common acknowledgement that it performs poorly under most conditions (Ledford, Ayres, Lane, & Lam, 2015). When WIR is used, the observer (data collector) records an occurrence if the target behavior occurs for the entire duration of the interval. Thus, a behavior is *only* recorded as occurring if the behavior begins at or before the interval onset and continues until the interval is complete.

Benefits and Weaknesses

WIR has no notable benefits, since it is more resource intensive than simple timing or counting and is largely inappropriate for estimating count and duration.

Steps for Use of WIR

Although we do not recommend the use of WIR, it is important to understand the procedures used in order to better interpret the data from studies that used this measurement system, thus we have outlined them below:

1. Operationally define behavior occurrence.
2. Choose an interval length that is as short as is feasible given measurement and resource constraints (e.g., 5 seconds).

3. Set up a data collection system that allows for coding of a behavior occurrence (or non-occurrence) during each interval.
4. Set an interval timer to alert you via alarm or vibration at the end of each interval.
5. Record occurrences and non-occurrences:
    a. Record a behavior occurrence if the behavior occurs *for the entire duration of the interval.*
    b. Record a behavior non-occurrence if the behavior does not occur for the entire interval; non-occurrences are recorded for intervals in which the behavior does not occur at all *and* for intervals in which the behavior occurs for some but not the entire interval (including intervals in which the behavior occurs for most but not all of the interval).
6. Following session completion, summarize the data:
    a. If you are interested in *number*, count the number of intervals in which the behavior occurs. Report this number as an *estimated count.*
    b. If you are interested in *time*, count the number of intervals in which the behavior occurs and divide that number by the total number of intervals to get a percentage of intervals in which the behavior occurred. Report this percentage as an *estimated duration.*

Momentary Time Sampling

Momentary time sampling (MTS), like PIR, is widely used in SCD research (Lane & Ledford, 2014; Lloyd et al., 2016; Mudford et al., 2009). When MTS is used, the observer (data collector) records an occurrence if the target behavior is occurring at the moment the interval ends. The occurrence or non-occurrence of the behavior at any other time during the interval is disregarded. A variation of MTS, dubbed the PLA-CHECK, measures the behavior of a group of participants by counting the number of engaged participants out of the total number of participants at the end of each interval (Doke & Risley, 1972)—in this variation, the "case" is the *group* of participants.

Benefits and Weaknesses

MTS is likely the easiest-to-use interval-based system because it requires attending to the presence or absence of a target behavior at a single point in time for each interval; however, it is most accurate when small intervals are used (e.g., 5 seconds), perhaps minimizing this advantage. MTS is the most accurate interval-based system for estimating duration (Ledford et al., 2015). MTS should not be used for estimating count unless the behaviors (a) have clear onsets and offsets and (b) are long duration behaviors (behaviors with non-trivial durations; see above).

Steps for use of MTS

If you decide to use MTS to estimate duration, you should follow guidelines below for use:

1. Operationally define behavior occurrence.
2. Choose an interval length that is as short as is feasible given measurement and resource constraints (e.g., 5 seconds).
3. Set up a data collection system that allows for coding of a behavior occurrence (or non-occurrence) at the end of each interval.
4. Set an interval timer to alert you via alarm or vibration at the end of each interval.
5. Record occurrences and non-occurrences:
    a. Record a behavior occurrence if the behavior occurs *at the moment the interval ends.*
    b. Record a behavior non-occurrence if the behavior is not occurring at the moment the interval ends, even if the behavior has occurred at other times during the interval.
6. Following session completion, summarize the data:
    a. If you are interested in *number*, count the number of intervals in which the behavior occurs. Report this number as an *estimated count.*
    b. If you are interested in *time*, count the number of intervals in which the behavior occurs and divide that number by the total number of intervals to get a percentage of intervals in which the behavior occurred. Report this percentage as an *estimated duration.*

Accuracy of Interval-Based Measurement Systems

There are numerous research studies regarding the inaccuracies of interval-based systems (Ary & Suen, 1983; Harrop & Daniels, 1986; Ledford et al., 2015; Powell et al., 1975; Rapp et al., 2007; Yoder et al., 2017). Despite these studies, interval-based recording procedures continue to be used in the applied behavioral literature, especially for measuring prosocial behaviors, communicative responses, or challenging behaviors. Moreover, common recommendations have been provided, including the use of intervals that are approximately the same length as or smaller than the average behavior duration per occurrence (Kazdin, 2010; Cooper, Heron, & Heward, 2007), although these recommendations do not always result in accurate measurement. It is also commonly reported that PIR overestimates behavior occurrence, WIR underestimates behavior occurrence, and MTS both overestimates and underestimates behavior occurrence. However, the behavior of all interval-based systems is more complicated than simple under or overestimation. For example, the extent to which each under or overestimates behavior is reliant on (a) whether it is an estimation of count or duration, (b) size of interval relative to the average duration per occurrence, (c) whether the estimate is for a

short duration or long duration behavior, and (d) number of occurrences per session. We illustrate these issues with Figures 5.1 and 5.2; additional and more complex analyses of issues with interval-based systems can be found in a number of peer-reviewed publications, cited above.

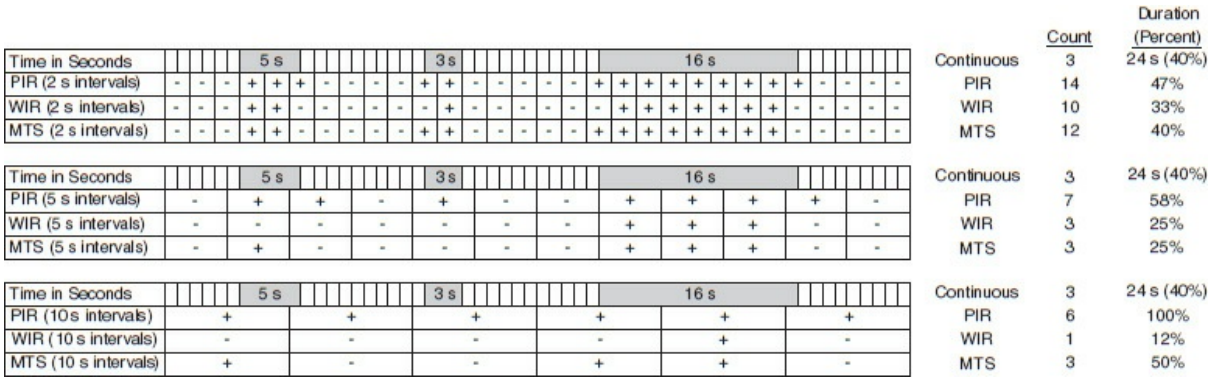


Figure 5.1 Sample data depicting three occurrences of a long-duration behavior (depicted by gray fill), and the estimates of count and duration when partial interval recording, whole interval recording, and momentary sampling are used with 2-, 5-, and 10-second intervals.

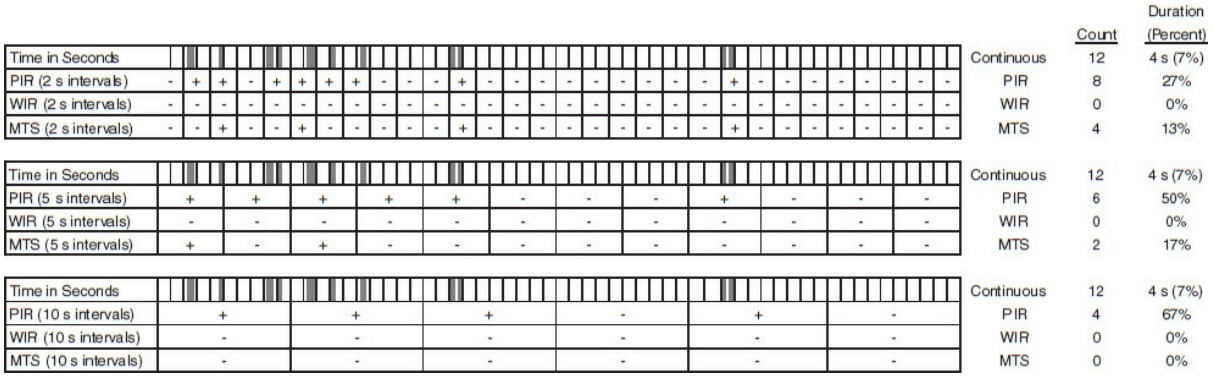


Figure 5.2 Sample data depicting 12 occurrences of a short-duration behavior (depicted by gray fill), and the estimates of count and duration when partial interval recording, whole interval recording, and momentary sampling are used with 2-, 5-, and 10-second intervals.

Illustration of Accuracy for Behaviors with Non-Trivial Durations

Figure 5.1 depicts a one-minute "session"; this is not a typical session length but results from this brief illustration hold for session lengths common in SCD research (Ledford et al., 2015; Yoder et al., 2017). Each cell in the top row corresponds to 1 of 60 seconds in that minute; shaded cells represent a behavior "occurring" during that portion of the session. Thus, you can see that the minute-long session included three behavior occurrences, totaling 24 seconds. The second through fourth rows depict the time period divided into thirty 2-second intervals. In each of these cells is a "+", denoting that a behavior occurrence was coded, or a "-", indicating that a behavior occurrence was not coded, according to each interval system. The remaining two charts show the same

behavior occurrence, with behavior occurrences marked for 5 seconds (middle) and 10 seconds (bottom) intervals. This figure includes data that would be consistent with a behavior that occurs for at least a few seconds at a time (i.e., long duration; non-trivial duration), such as crying, being on-task, or engaging in parallel play with peers. For these behaviors, we present the accuracy of interval systems for estimating number and time, although duration (percentage of session in which the behavior occurred) is most often of interest when behaviors of non-trivial durations are measured.

For all comparisons in Figure 5.1, the accurate count of behavior occurrence is 3 (i.e., within the 1-minute interval, the behavior occurred three different times). As is reported in the data on the right side of the figure, PIR never resulted in an accurate count—for all three interval sizes, PIR resulted in an estimated count of 6–14, at least double and up to almost 5 times the actual count. WIR resulted in an accurate estimate for one of three interval sizes, overestimated for one, and underestimated for one. MTS was accurate for two of three interval sizes. Thus, when 2-second intervals were used, all three systems resulted in overestimates of count; when 5-second intervals were used, PIR resulted in overestimates; and when 10-second intervals were used, PIR resulted in overestimates and WIR resulted in underestimates of behavior counts.

For all comparisons in Figure 5.1, the accurate duration of behavior occurrence is 24 seconds, or 40% of the session. As is reported in the data on the right side of the figure, all three interval systems resulted in somewhat accurate estimates with very small intervals (33–47%), but with larger intervals, PIR overestimated and WIR underestimated duration of behavior occurrence and MTS under (5 seconds) **or** overestimated (10 seconds) occurrence. These patterns occur because WIR will "miss" occurrences that do not span an entire interval (e.g., any occurrence less than 10 seconds in duration, if intervals are 10 seconds), while PIR will over-count any occurrence that lasts for longer than an interval length (e.g., a 3-second occurrence will always be estimated as two 2-second occurrences when 2-second intervals are used). MTS, on the other hand, includes random error—that is, behavior occurrence is likely to be somewhat accurate, with increased accuracy when the interval size is shorter.

Illustration of Accuracy for Behaviors With Trivial Durations

Figure 5.2 depicts a one-minute "session," with cells depicting occurrences and interval system data similar to Figure 5.1. However, in Figure 5.2, behavior occurrences are depicted which are trivial in duration (1/3 of a second, for the purposes of this illustration). These types of behaviors are often measured in SCD research—for example, utterances made by a toddler, hits to the head by a child with autism and self-injurious behavior, and number of times an adult imitates a child's play behavior. Count is most often of interest when behaviors of trivial duration are measured (e.g., a child can hit himself 50 times during a 10-minute session, and still a relatively short duration of total hits would be measured).

For all comparisons in Figure 5.2, the accurate count of behavior occurrence is 12. As

is shown in the data on the left side of the figure, none of the interval-based systems resulted in accurate counts; all were underestimates. For behaviors with trivial durations, neither MTS nor WIR is appropriate, even when very small intervals are used. PIR resulted in underestimates, with greater underestimates for bigger intervals and when more behaviors occur (e.g., are closer in time to each other). This predictable and lawful behavior by PIR allows us to use a statistical Poisson correction to improve the accuracy for estimating counts (Yoder et al., 2017). This correction can only be used when the number of intervals *in which a behavior onset occurs* is recorded. The formula involves a natural log transformation of the quotient of the number of "non-occurrence" intervals divided by the total number of intervals; that number is multiplied by the quotient of session duration divided by interval duration (in seconds) to obtain the final, corrected count estimate. A spreadsheet that performs the necessary calculations is available at: http://tinyurl.com/Poisson-Correction (Yoder et al., 2017); the formula is:

$$-\ln(\#\,\text{nonoccurrence intervals} \div \text{total}\,\#\,\text{intervals}) \times (\text{session duration} \div \text{interval duration})$$

Use of the Poisson transformation considerably increases accuracy of count estimations of behaviors of trivial duration (Yoder et al., 2017); thus we suggest its use when count of these behaviors is of interest. Even when the correction is used, more accurate results are obtained by using small intervals (Yoder et al., 2017). For example, in this example provided in Figure 5.2, all estimates of count are increased by 1–2 instances, making the estimates closer to the continuous count (with the most accurate correction resulting from the most accurate beginning estimate, with 2-second intervals).

We also present duration data for Figure 5.2; it is almost never of interest to estimate duration of these types of behaviors. No interval-based systems allow us to do so accurately, although MTS with *very small intervals* results in somewhat accurate estimates. We suggest interval-based systems not be used to estimate duration of behaviors with trivial durations; suggestions for the use of measurement systems by dimension (count, time) and type (continuous, non- continuous) are shown in Table 5.4.

Reporting Use of Interval Systems

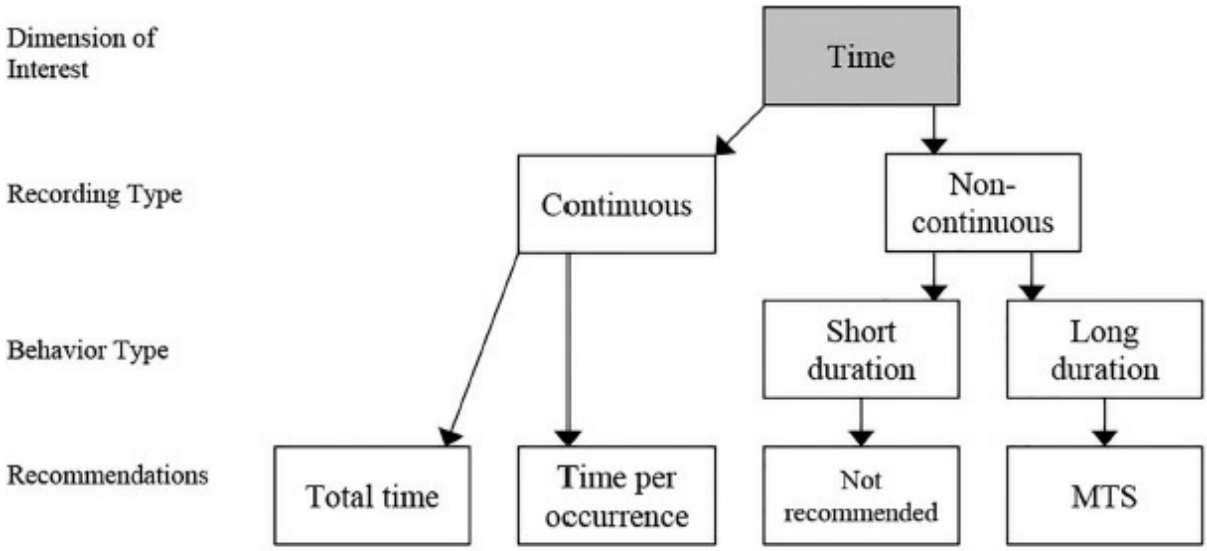
When interval-based systems are used, researchers should take care to report all parameters (system type, duration of intervals, number of intervals per session), explicitly identify the system as an *estimate* of behavior occurrence, name what dimension of behavior is being estimated (e.g., number, time), and discuss the likelihood of error. If time is being estimated (i.e., duration, latency, inter-response time), provide results as a percentage of intervals in which the behavior occurred as an estimated percentage of duration of the session. If number is the dimension of interest, report the number of intervals in which behavior occurred as an estimated count, using the Poisson correction for PIR previously described. In Figure 5.3, we provide a flow chart that can be

used for selecting a measurement system based on whether you will use continuous or non-continuous recording systems, the dimension of interest (time, number), and the type of behavior (long duration, short duration).

Table 5.4 Suggestions for Measurement Based on Dimension of Interest and Type.

	Number	**Time**
Continuous	Event recording Timed event recording	Total duration recording Duration per occurrence recording
Non-Continuous	PIR, using a Poisson Correction, for behaviors of trivial duration (e.g., hits, imitation, utterances). Report **number** of intervals as **count** estimate.	MTS, using small intervals, for behaviors of non-trivial duration (e.g., engagement, parallel play, tantrum behavior). Report **percentage** of interval as **duration** estimate.

Note: We do not suggest the use of PIR for estimating time, MTS for estimating count, or WIR for estimating either.



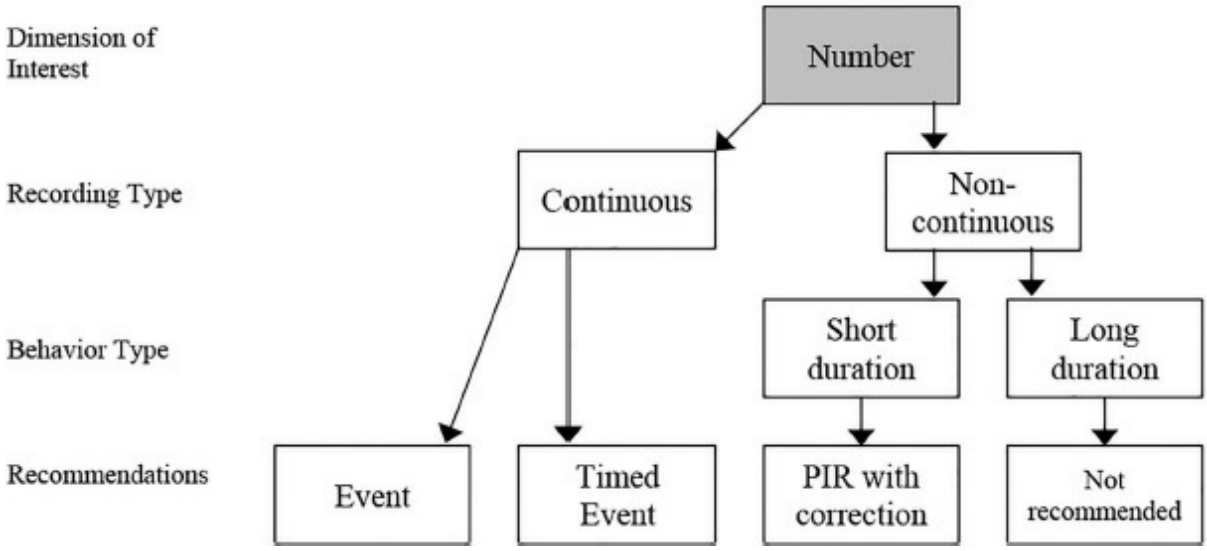


Figure 5.3 Flow charts for determining measurement system use when time is the dimension of interest (top chart) or number is the dimension of interest (bottom chart).

Collecting Data on More Than One Behavior

As SCD researchers, we are often interested in changes in more than one behavior. For purposes of experimental decisions, you must always specify a primary dependent variable. It is the analysis of this behavior that will drive decisions about condition changes (read more about condition changes for specific designs in Chapters 9–12). However, additional behaviors are often measured in the context of SCD research. For example, you might measure both duration of engagement and number of social interactions for child participants (Ledford et al., 2015) or measure adult fidelity to procedures (percentage correct) for adult participants as well as duration of engagement

for a child participant (Ledford, Zimmerman, Harbin, & Ward, 2017). Whether the variables are for the same or different participants, one should be named explicitly as the primary variable.

Sometimes, we are interested in coding a group of variables that are related to each other. For example, in a study designed to assess the effects of an intervention on classroom engagement for a young child, we might be interested in coding whether he or she was engaged with materials or people, unengaged, engaged in stereotypy, or appropriately waiting. Given video records, we could separately code for each behavior using duration per occurrence or MTS recording. However, especially in the case of MTS, we could also code all behaviors simultaneously if they are *exhaustive* (i.e., inclusive of all potential behaviors) and *mutually exclusive* (i.e., cannot occur at the same time). That is, at the end of each interval, rather than recording "occurrence" or "non-occurrence," we would record *engaged, unengaged, stereotypy*, or *waiting*. Although these behaviors will co-vary (e.g., if engagement improves, one of the other behaviors must decrease), a single behavior should still be named as the primary behavior of interest and that behavior should be used to make experimental decisions. Use of an exhaustive and mutually exclusive code (including simple occurrence/non-occurrence codes) allows for more flexibility in the analysis of reliability data (see Calculating Agreement section, below).

Data Collection

DSOR is a hallmark of SCD research. This aligns well with the type of data collection that occurs (or should occur) in practice. Generally, SCD researchers use measures that align with proximal and context-bound outcomes (e.g., directly measure change in the behavior we targeted in the context in which it was taught; Yoder & Symons, 2010). This is in contrast to measures that are distal and generalized. For example, teaching a child to name math facts in a small group in his classroom and measuring his progress in acquiring those facts during the small group session involves measuring a proximal and context-bound outcome. Teaching a child to name math facts and then measuring growth on a standardized measure of math achievement in a clinical setting is a distal and generalized outcome. These concepts are not truly dichotomous and SCD research includes outcomes measurement that can involve dependent variables that are more or less proximal and more or less context-bound. In any case, almost all SCR data are collected via researcher-developed measures, in part due to the lack of appropriate standardized measures for repeated used over time, but also because researcher-developed measures can be designed to be sensitive to small but meaningful changes in participant behavior. Below, we describe the type of information that should be collected and the use of technology to improve data collection and analysis.

Planning and Conducting Data Collection

Data collection not only involves gathering information about the specific behavior of interest (performance information), but also other information critical to interpretation and organization (situational data; i.e., participant identification numbers, implementer initials, date, time; McCormack & Chalmers, 1978). In addition, study-specific information such as instructional phase or modifications should be recorded so that you have a historical record of decisions made during the study. Finally, summary information should be recorded, including summary statistics (e.g., percentage correct, total number of intervals) and whether inter- observer agreement (IOA) and procedural fidelity (PF) data were collected and if so, the scores. If you use the exact same form for primary and secondary (IOA) data collection, it is important to have a section on the form to designate whether you are the primary or secondary observer. In Appendices 5–1, 5–2, 5–3, and 5–4, you will see example data collection forms for trial-based event recording, free operant event recording, interval recording, and duration recording.

Using Technology

Although the critical nature of data collection and essential components of measurement

have remained more or less unchanged over time, technological advances have resulted in changes in the processes of data collection. Most of these changes are beneficial (e.g., increased feasibility, improved analysis, automatic calculations). Some potentially troublesome issues with technology are potential increased risk for confidentiality violations due to information stored on electronic devices and increased risk for data loss due to electronics failure. However, overall, the use of technology for data collection has moved the field forward and increased the feasibility of measuring increasingly complex behaviors. Although technology changes at a rate faster than book publication changes, two important technological advances seem relevant to discuss: use of video recording and use of electronic applications.

Video recording experimental sessions is not a new idea; however, the relative ease and widespread social use of recording via portable electronics devices has increased the feasibility and social acceptability of using these devices in applied settings. Video recording sessions has several notable benefits; it allows for: (a) a researcher to implement a condition as intended, while collecting data at a later time; (b) researchers to have more detailed discrepancy discussions (see below); and (c) blind measurement (i.e., for someone who is unaware of condition assignment to collect data). Despite these considerable positive attributes, video recording may pose additional concerns for participants, including those related to privacy and confidentiality (i.e., it may increase the chance that a non-researcher may see research activities). When video is used, participants (or their legal guardians who provide consent) should be notified of potential drawbacks of the use of technology (see Chapter 2). When video is used, the same information described above should be collected, via paper/pencil data collection or electronically.

The use of electronic applications for data collection fall into two primary categories: computer-based programs that can be used to code data from video (e.g., ProCoderDV, Tapp, & Walden, 1993) and mobile applications on phones or other portable electronic devices. When codes are used for participant information (e.g., pseudonyms or participant numbers rather than names), use of these products does not necessarily increase the likelihood of privacy or confidentiality concerns. Moreover, they allow for more precise measurement (e.g., timed event recording) and often perform basic calculations (e.g., percentage of intervals). These applications are often free or low-cost (see Table 5.5; note that application utility, availability, and pricing change frequently); some high-cost options are available and widely used in practice. When determining whether an electronic application is the right fit for an SCD study, you should consider whether: (a) use of the device is permitted and feasible, and whether connectivity is required and likely to be an issue; (b) the device provides or allows you to input all of the relevant information needed; (c) all data collectors have easy access to a device compatible with the data collection software or application; and (d) you can adequately manage, analyze, and store data given the constraints of the product.

Table 5.5 Low or No-Cost Data Collection Applications.

Name	Author/Developer	Cost
Behavior Tracker	NexTechnologies	$0.99
Countee	Peic, D., & Herandez, V.	Free
Intervals	elocinSoft	$4.99

All applications retrieved from http://itunes.apple.com in 2017.

Potential Problems Related to Dependent Variable Measurement

In SCD research, data are collected repeatedly over time, and almost always via observational recording. Thus, humans observe and record behavior (usually based on researcher-devised systems), and we make decisions based on those observations. Mark Wolery, an SCD researcher who considerably influenced the field of early childhood special education, has said, "humans are the worst data collectors but are often superior to all other options" (2011). Although problems, such as invalidity, inaccuracy, and unreliability are not specific to SCD or repeated observational measurement, the nature of measurement in SCD does pose some different problems than those generally faced by group design researchers.

Invalidity

There are multiple types of validity; thus far we have discussed *internal validity* (believability that results are due to independent variable) and *external validity* (generality); in the next chapter we will discuss social validity. Now, we discuss the type of validity most relevant to repeated measurement of dependent variables in the context of SCD, **construct validity**—which refers to whether your measurement procedures accurately reflect the concept you are interested in measuring (Crano & Brewer, 2002). Although we measure specific, observable behaviors in SCD research, we do so because they represent an important construct such as social or academic competence (Shadish, Cook, & Campbell, 2002). However, the match between well-defined and reliably measured behaviors and broadly-defined, socially important constructs can be difficult to achieve. For example, assume your definitions for problem behavior include touching others without permission. Given that definition, pats on the back and inadvertent touching in line count as problem behavior—thus, your construct validity might be low if those behaviors are not problematic. While specific and observable operational definitions might result in high reliability, it does not necessarily ensure that the definitions are sufficient for allowing the measurement of the behavior you are interested in. Especially when measuring broader social constructs like "interactions" or "engagement," you should ensure that your specific and observable defined behaviors are well aligned with the concepts from which they were derived (Barlow & Hersen, 1984).

Inaccuracy

Inaccuracy refers to the failure of the measurement system to perfectly reflect behaviors

that actually occurred: (a) behaviors that occurred were not coded or (b) behaviors that did not occur were coded. Reasons for inaccuracy include human error and nonspecific definitions that omit or provide limited information regarding examples and non-examples in context. Unfortunately, accuracy is not a construct that is easily measured; that is, a "true" value of behaviors is dependent on a human observer (or sometimes computerized or other mechanized counts), but these transformations can never be considered "true" values. Instead, we increase confidence in the *accuracy* of measurement via assessment of reliability (Kazdin, 2010).

Unreliability

To increase the likelihood of accurate measurement, we rely on measuring the *reliability* of measurement, or the extent to which two observers will record behavior occurrence the same way. When observers disagree on behavioral occurrences, one of three common problems may be present: bias, drift, or error.

Observer bias

Bias refers to the likelihood that a data collector has conscious or unconscious beliefs which impact their data collection in a predictable direction. Bias generally occurs when a researcher believes his or her intervention will "work" to change behavior (cf. Chazin, Ledford, Barton, & Osborne, 2017), although it can also occur such that a researcher believes the intervention is unlikely to work. For example, if a behavioral researcher compares a behavioral intervention to a sensory-based intervention, he or she may likely be biased *against* the sensory intervention and be biased *in favor of* the behavioral intervention. It is important to note that bias does not necessarily include conscious decision-making or malevolent or unethical intent. Bias can be detected and prevented by collecting interobserver agreement data, frequently graphing and analyzing data, and using blind observers.

Observer drift

Observer drift refers to the tendency of a data collector to depart from accurate use of definitions over time. This is especially problematic in SCD research because of the repeated and extended nature of data collection for a single participant. Observer drift can be detected and prevented by collecting interobserver agreement data, frequently graphing and analyzing data, encouraging consistent referencing of coding definitions, and having discrepancy discussions.

Error

Bias and drift are specific inaccuracies that lead to predictable errors. However, some mistakes are simply unsystematic inaccuracies that result from observers incorrectly applying definitions. These can include (a) observer inattention, generally leading to underestimates of behavior occurrence; (b) difficulty adjusting coding given new conditions (e.g., onset of a new condition dramatically changes number of behaviors that occur, increasing complexity of data collection task), (c) misinterpretation of definitions, and (d) unexpected ambiguous occurrences. Error can be reduced by training observers to a set criterion before beginning data collection, and training in a range of contexts (e.g., situations likely to be contacted during the study, across experimental conditions); limiting the amount of data collection done in a short period of time; being familiar with your research participants and their likely behaviors; and having discrepancy discussions.

Ensuring Reliability and Validity of Data Collection

When planning and conducting SCD research, it is of paramount importance to ensure that you collect valid and reliable data on your dependent variables of interest. Doing so improves the internal validity of your study by improving the confidence that any changes between conditions indicated by your data are indicative of actual changes in participant behavior and not unplanned or unrelated factors.

Operationalize Behaviors

When writing non-examples, clearly exemplify behaviors that are similar to those of interest, but do not represent the construct of interest (e.g., if you are interested in vocal social interactions between peers, ensure that non-socially directed labeling of items is *not* counted as an occurrence) to minimize the likelihood of ambiguous occurrences. Non-examples should not simply be a list of opposites of the examples provided; rather they should serve to identify behaviors that are close but not counted as the target behavior. Individualizing operational definitions based on child observations prior to study onset is desirable, if possible.

Pilot Data Collection Procedures

When conducting research, it is important to ensure that definitions and measurement procedures that you carefully devised are accurate and appropriate for gathering information about the dependent variable of interest. It is prudent to ensure this is the case *prior* to beginning data collection for the study. Thus, when possible, researchers should consider piloting their data collection systems before beginning the first condition in a study. This pilot can be conducted with the intended participants, individuals who are similar to the intended participants, or confederates. Benefits and drawbacks of piloting with each group are shown in Table 5.6. Note that these data will *not* be reported in research reports and generally do not require IRB approval; however, you generally do need client or parent/guardian permission to collect data, especially if individuals are identifiable (e.g., via video). During pilot data collection, ask yourself whether, using your definitions, all observers (a) captured all relevant behaviors that matched your construct of interest, and (b) did not capture similar behaviors that did not match your construct. Following piloting procedures, you should assess reliability and validity, and revise definitions, examples, and non-examples accordingly.

Table 5.6 Benefits, Weaknesses, and Examples of Use of Varying Participants in Pilot Activities.

	Benefit	Weakness	Example

Intended participants	If participants are easily accessible, identify idiosyncratic behaviors not considered in initial development	Participants may have similar levels of behavior to likely baseline levels, but data system may not work well when behavior changes during intervention conditions	While planning a study intended to improve toy engagement, Jen practiced using her data collection system by observing the young child she intended to recruit for study participation in her typical classroom activities.
Individuals similar to intended participants	If participants are difficult to access, similar individuals can result in identification of likely issues during data collection such as potentially ambiguous behaviors	Same as above; choosing several different individuals with different levels of behavior can help to remediate this problem	While planning a study designed to improve reading rates in a public school, David practices his data collection system with several young children in a lab school he visits frequently
Confederates	Confederates can devise a variety of situations with multiple levels and types of behavior occurrence	Confederates may not engage in behaviors that are similar to participants	While planning a study designed to improve social interactions among peers, Justin recruits several undergraduate students to set up pretend play scenarios among themselves, with some sessions

	including high rates of interactions and some including low rates.

Train Observers

Following development and testing of your data collection system (definitions, examples, non-examples, measurement procedures), train all primary and secondary (i.e., IOA) data collectors. To do so, we suggest (a) providing definitions, examples, non-examples, and procedures in writing; (b) practicing coding alongside the data collector, answering questions, and resolving conflicts; (c) discussing any discrepancies and revising written guidelines as appropriate; (d) both observers independently coding a second session (e.g., at the same time or from the same video) and calculating the extent to which you agree; (e) discussing any discrepancies and revising written guidelines as appropriate; and (f) repeating until the trainee reaches your criterion level. Generally, an acceptable criterion level for training is 90% agreement between the primary investigator and all other observers. See below for specifics regarding calculating agreement.

Use Blind Observers

Blind observers refer to data collectors who do not know the condition in effect for the data they are collecting, which can be costly and logistically difficult (Wolery & Garfinkle, 2002). For example, Chazin and colleagues (2017) conducted three different types of sessions to determine whether physical activity had an impact on subsequent behavior during large group activities— seated activities, activities designed to evoke moderate-to-vigorous physical activity, and typical classroom activities. Following implementation of one of three conditions each day, the classroom large group activity was recorded. In the video recording, there was no indication of which condition had preceded the large group activity, so observers could code data without being potentially biased regarding outcome measurement (Chazin et al., 2017). Although blind observers are rarely used in SCD research (Tate et al., 2016), they are critical for reducing the possibility of observer bias in instances where a condition is not apparent (an assumption that has received recent attention but was acknowledged years ago; cf. Bushell, Wrobel, & Michaelis, 1968). For some research questions, blind observers are less feasible (e.g., in a study regarding the use of visual supports, it will be apparent to observers whether these supports are present or absent). However, observers can be recruited who are blind to study purpose and hypotheses (e.g., observers are trained on dependent variable data collection but are given no information about changes between condition and how that may impact measurement).

Collect and Present IOA Data

SCD researchers most often asses reliability between two observers who have observed and recorded behavior simultaneously but independently and reported the extent to which they agree as a percentage; this is often referred to as **interobserver agreement** (IOA) but can also be called inter-rater reliability or inter-assessor agreement. To collect IOA data, two independent data collectors observe and record behavior during a single measurement occasion on identical but separate data collection forms. When data are collected in situ, rather than via video, observers should take care to truly be independent; this may require consideration of physical positions and data collection forms. For example, the two observers may need to position themselves on opposite sides of the room so that they are less likely to view each other's data collection forms or devices. In addition, when interval-based systems or trial-based even recording is used, observers should collect data for occurrences and non-occurrences so decisions are not apparent (i.e., one observer will not be able to ascertain whether the other observer is marking an occurrence or a non-occurrence). When interval-based recording is used, take care to synchronize your recording devices so that uncoordinated timing does not result in discrepant outcome measurement.

Formative Analysis of IOA Data

Following data collection, researchers should analyze IOA data. Immediate analysis should be formative in nature, and should occur following each IOA measurement occasion. Formative analysis should be used to inform researchers regarding the extent to which their definitions and procedures are adequate and to alert researchers when additional training is needed. For formative analysis purposes, researchers should graph data from both observers on a single graph (Artman, Wolery, & Yoder, 2010; Chazin et al., 2017; Ledford, Artman, Wolery, & Wehby, 2012; Ledford & Wolery, 2013). This allows for the visual analysis of differences between observers and allows researchers to identify potential observer drift or bias. For example, in the top panel for Figure 5.4, the average percent agreement is within acceptable ranges (see below), and you can see that the blinded observer sometimes counted more occurrences of initiations than the primary observer, and sometimes counted fewer initiations. This suggests that systematic bias is *not* present and drift is unlikely. The middle panel of Figure 5.4 shows data with identical agreement percentages (81.7%) but all errors in baseline were such that the blind observer identified more positive outcomes (a greater number of initiations) in baseline and fewer positive outcomes during intervention (fewer initiations)—this suggests the likelihood of observer bias is high. In the bottom panel of Figure 5.4, you can see that the blind observer's data slowly drifts farther from the primary observer's data. This suggests drift is present, although it is not possible to determine which observer (if not both) is becoming less accurate in applying the definitions.

After data are plotted, any differences between observers should be discussed and a

consensus should be agreed for each instance (e.g., What is the correct response?); this is referred to above as a **discrepancy discussion**. Following consensus coding, primary data can be altered to more accurately match coding definitions and reduce errors, but original calculations should be reported (i.e., data can be re-plotted so that error is not shared with eventual consumers, but *recorded IOA percentages should never be altered*; see below). If bias is likely, a blinded observer should code *all* remaining sessions (and previous sessions, if they are video recorded). If drift is likely, observers should be re-trained.

Calculate Agreement

In addition to plotting IOA data for formative analysis, researchers should calculate agreement using either percentage agreement or the Kappa coefficient for the purposes of formative and summative evaluation. Formatively, researchers should analyze disagreements and determine whether additional training is needed. Summatively, researchers should report agreement to support reliability of data collection.

Percentage Agreement

Percentage agreement is a simple calculation that is intuitive and widely used. Despite these benefits, percentage agreement is a calculation that is influenced by chance agreement, behavior rates, and measurement system used (cf., Kratochwill & Wetzel, 1977). Generally speaking, percentage agreement refers to the number of opportunities in which two observers agree, considering the total number of opportunities for agreement. Percentage agreement is calculated and interpreted differently depending on the measurement system used.

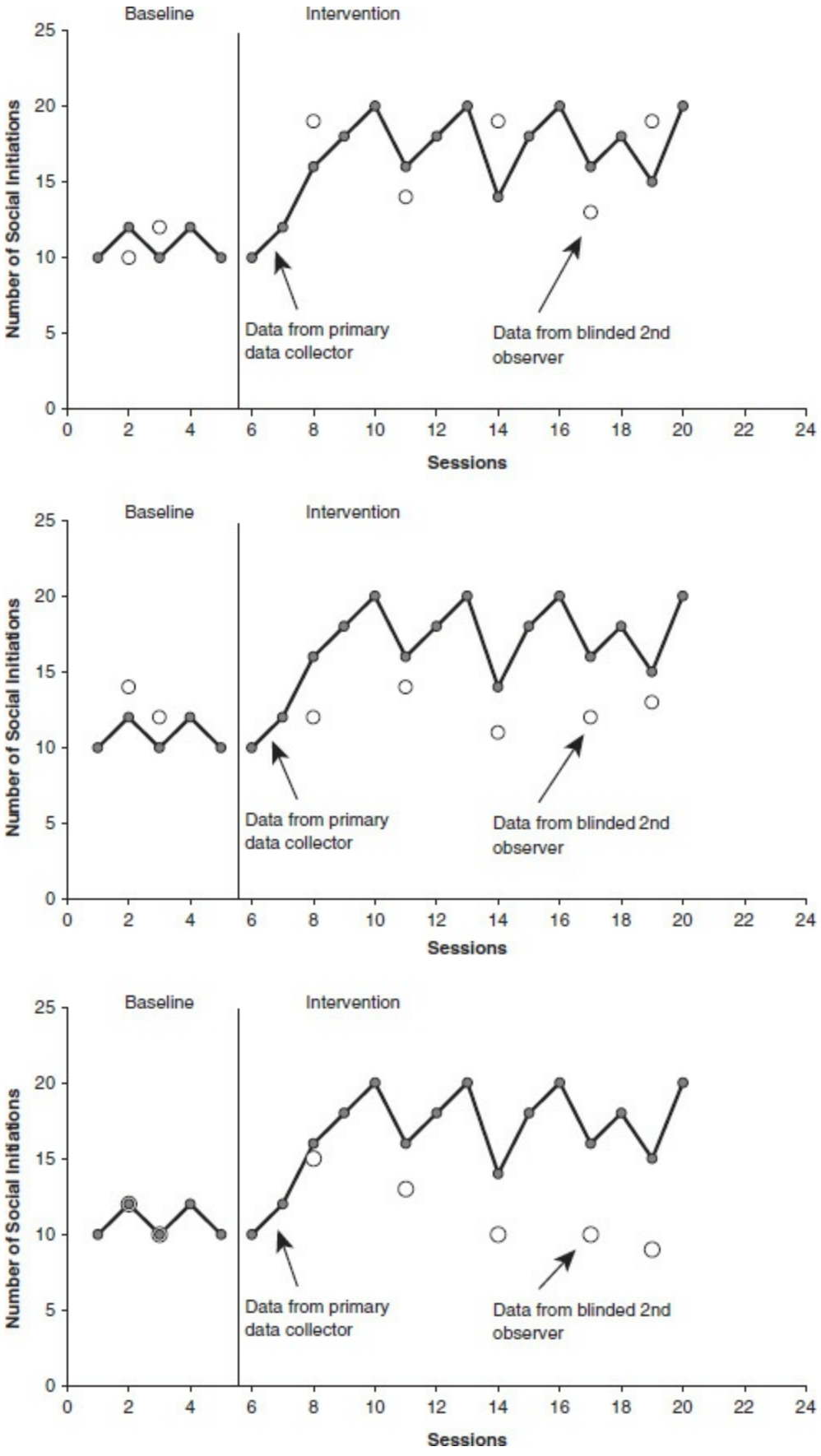


Figure 5.4 Three graphs depicting data from a primary observer alongside data from a secondary observer. The top

panel depicts adequate agreement with disagreements occurring in both directions. The middle panel depicts adequate agreement with potential observer bias. The bottom panel depicts adequate agreement with evidence of observer drift.

POINT-BY-POINT AGREEMENT FOR TRIAL-BASED BEHAVIORS AND INTERVAL-BASED RECORDING SYSTEMS

When trial-based or interval-based measurement is conducted, agreement can be calculated using trial-by-trial (or interval-by-interval) comparisons (**point-by-point agreement**). To conduct agreement in this way, compare the code for each interval (or trial) for one observer with the code for the corresponding interval (or trial) for the second observer. Note whether the codes are the same (agreement) or different (disagreement). After determining the number of intervals coded as agreement or disagreement, calculate total percent agreement (Tawney & Gast, 1984):

$$\left(\frac{\#\ of\ agreements}{\#\ of\ agreements + \#\ of\ disagreements}\right) \times 100$$

Historically, average agreement of 80% or better has been considered acceptable (Kazdin, 2010) and it is a common threshold for determining sufficiency (What Works Clearinghouse, 2014). The extent to which this is true depends on several factors, including the complexity of the behavior and context and the degree of behavior change between conditions. For example, 80% average agreement regarding whether a child named sight words correctly is likely to be viewed as too low because coding correct and incorrect responding to a simple task is generally quite straightforward; 80% average agreement for social interactions in a typical classroom free play context is more reasonable due to the complexity of the code and the context. Also, as illustrated in Figure 5.4, 80% agreement when changes between conditions are small results in decreased confidence that the change was due to the intervention rather than bias. Any sessions in which IOA was lower than 80% should be explained in text. See the top panel of Figure 5.5 for an example calculation of point-by-point agreement when an interval system is used.

OCCURRENCE AND NON-OCCURRENCE AGREEMENT FOR TRIAL-BASED BEHAVIORS AND INTERVAL-BASED RECORDING SYSTEMS

Because chance agreement is likely when rates of behaviors are low (e.g., if almost all trials or intervals are non-occurrences), some researchers have suggested the use of **occurrence agreement** (Tawney & Gast, 1984). To calculate occurrence agreement, you code agreements and disagreements (as described above) *only for intervals in which at least one observer noted an occurrence* (which we have abbreviated as "occurrence trials" or OT). Thus, the calculation is:

$$\left(\frac{\#\ of\ agreements\ for\ occurrence\ trials}{\#\ of\ agreements\ for\ OT + \#\ of\ disagreements\ for\ OT}\right) \times 100$$

Similarly, **non-occurrence agreement** can be calculated, using only trials in which at least one observer noted that a behavior did not occur (which we have abbreviated as "non-occurrence trials", or "NOT").

$$(\#\,of\,agreements\,for\,non-occurrence\,trials \#\,of\,agreements\,for\,NOT + \#\,of\,disagreements\,for\,NOT) \times 100$$

Non-occurrence agreement and occurrence agreement may be greater than or less than total agreement, depending on the types of disagreements that occurred in the session. We recommend evaluating occurrence and non-occurrence agreements for formative use, but these are rarely reported in published manuscripts.

POINT-BY-POINT AGREEMENT FOR FREE OPERANT BEHAVIORS MEASURED WITH TIMED EVENT RECORDING

Before collecting data for behaviors you measured with timed event recording, you should establish a time frame within which you will record an agreement if both observers mark an occurrence. For example, in Figure 5.5, the middle panel depicts that Observer #1 marked an occurrence at 1:28 and Observer #2 marked an occurrence at 1:26. Although the time stamp is not exactly the same, it seems unlikely that Observer #1 would count a true occurrence and miss another true occurrence, and vice versa for Observer #2. What is more likely is that one observer had a slightly quicker response time. Generally, a window of a few seconds (e.g., a maximum of 2–5 seconds, depending on the complexity of the code and context) is acceptable. With timed event recording, there are no non-occurrences, so total agreement is calculated similarly to occurrence agreement for trial-based behaviors or interval-based measurement. First, you will line up occurrences to determine how many agreements you have within your given time window. Then you will count disagreements that occurred *outside* the time window (e.g., if your time window was 2 seconds, and one observer marked an occurrence at 3:32 and the other marked an occurrence at 3:36) and the instances in which one observer marked an occurrence and the other observer marked nothing. Thus, agreements include instances where both observers marked an occurrence at exactly the same time and instances where both observers noted an occurrence within the given time window. Disagreements include instances in which both observers marked an occurrence outside of the time window and occurrences that were only marked by one observer. Then, agreement is calculated as:

$$(\#\,of\,agreements\,within\,given\,time\,window \#\,of\,agreements + \#\,disagreements) \times 100$$

When the number of occurrences is very low (which is often true in either baseline conditions or intervention conditions, depending on intervention goals), even one disagreement can result in very low agreement. In these cases, the reason for low disagreement should be reported and additional IOA data should be collected (above and

beyond the usual minimum levels).

AGREEMENT FOR FREE OPERANT BEHAVIORS MEASURED WITH EVENT OR TOTAL DURATION RECORDING

Unlike the previous examples above, it is not possible to measure point-by-point when using event or total duration recording. Instead, **gross agreement** (total agreement) is calculated as such:

$$(\mathrm{smaller\ measurement} \div \mathrm{larger\ measurement}) \times 100$$

For example, if one observer recorded a total duration of 220 seconds and the second observer recorded a total duration of 242, the agreement would be 90.9% ([220/242]×100). This type of agreement is inferior because it prevents identifying discrepancies (e.g., no information is available about at what point the disagreements occurred) and because no evidence is available that all of the "agreement" (e.g., 220 seconds in the example above) actually referred to time in which both observers marked the same code. Thus, this is the least conservative and least preferred agreement method. When duration per occurrence is measured, point-by-point agreement can be measured based on onset (e.g., agreement on number of events) and gross agreement can be measured based on duration (e.g., agreement on duration), as shown in the bottom panel of Figure 5.5.

PARTIAL INTERVAL RECORDING (5 s intervals)

Time in Seconds		5 s			3 s			16 s				
Occurrences for Primary Observer	-	+	+	-	+	-	-	+	+	+	+	-
Occurrences for Secondary Observer	-	-	+	-	-	-	-	+	+	+	-	-
Agreement or Disagreement	A	D	A	A	D	A	A	A	A	A	D	A

Calculation for Point-by-Point Agreement:
Agreements: 9
Disagreements: 3
(9/[9+3])*100=75%

Calculation for Occurrence Agreement:
Agreements: 4
Disagreements: 3
(4/[4+3])*100=57%

Calculation for Non-Occurrence Agreement:
Agreements: 5
Disagreements: 3
(5/[5+3])*100=63%

TIMED EVENT RECORDING (2 s agreement window)

Behavior Occurrence Time Stamps
01:23
01:28
01:59
03:32
03:49
04:59

Behavior Occurrence Time Stamps
01:23
01:26
03:36
03:49

Exact Agreements: 2
Agreements within given time window: 1
Disagreement outside of time window: 1
Disagreemnts regarding occurrence: 2
([2+1]/[2+1+1+2])*100=50%

DURATION PER OCCURRENCE RECORDING

Occurrences in seconds
23
92
46
3
10
43

Total duration: 217 seconds

Occurrences in seconds
21
84
52
29

Total duration: 186 seconds

Point-by-point count agreement:
Agreements=4 Disagreements=2
(4/[4+2])*100=67%

Gross agreement for duration:
(186/217)*100=86%

Figure 5.5 Sample agreement calculations for trial or interval-based data (top panel), timed event recording (middle panel), and duration recording (bottom panel).

Kappa

As mentioned previously, chance agreement is likely when rates of behavior are very low and very high (i.e., a non-observer could score all intervals as occurrences or non-occurrences and have adequate agreement with an accurate observer). Although we don't expect many data collectors to purposefully falsify data, this chance agreement is still somewhat troubling, since it indicates that high agreement may not be highly associated with accuracy. The Kappa coefficient is superior relative to percentage agreement, because it mathematically corrects for chance agreement (Cohen, 1960); over time many researchers have argued that Kappa should be used instead of the more common percent agreement (Hartmann, 1977; Kratochwill & Wetzel, 1977; Watkins & Pacheco, 2000), despite different methodological issues related to base rates (for a more comprehensive review, see Yoder & Symons, 2010). Kappa can be calculated when interval based systems or duration recording is used, but cannot be calculated for event recording or timed event recording because to use Kappa, you must have information on occurrences *and* non-occurrences. To calculate Kappa, you divide percentage agreement minus chance agreement by one minus chance agreement; this leaves the proportion of agreement that is not related to chance. The calculation for chance agreement is:

$$\frac{(\#\,of\,OT\,for\,O1)\times(\#\,of\,OT\,for\,O2)}{(Total\,\#\,of\,trials)^2}+\frac{(\#\,of\,NOT\,for\,O1)\times(\#\,of\,NOT\,for\,O2)}{(Total\,\#\,of\,trials)^2}$$

Note: O1 = observer 1. O2 = observer 2. OT = occurrence trials. NOT = non-occurrence trials.

Note we use "trials" for the example, but it could also refer to intervals (for interval-based systems) or time (for duration measures). Several online calculators are available for the calculation of Kappa since the calculations are somewhat complex. Because Kappa removes the portion of agreement attributable to chance, acceptable Kappa values are somewhat lower than percentage agreement values (generally, 0.60 rather than 0.80 as the minimum acceptable value).

Summary

After choosing a behavior of interest, researchers follow systematic steps to ensure meaningful assessment of the outcome of an SCD study. These steps include: carefully defining the behavior and identifying examples and non-examples, characterizing the behavior based on reversibility and duration type, determining the dimension of interest, choosing a measurement system, piloting use of the system, training observers, and making modifications if needed. Following the initiation of data collection, additional steps are needed to ensure the reliability of data collection, including collection and formative and summative assessment of interobserver agreement data. Following the steps outlined in this chapter will ensure that the dependent variable assessment in your study results in meaningful conclusions about actual behavior occurrence and change.

Appendix 5.1

Trial-Based Event Recording

Data Collector Initials:_____ Date:_____ Condition Description:_____

Circle One: PRIMARY Observer **SECONDARY (IOA) Observer**

Participant Number: _____

Trial	Stimuli/Participant	Response
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		

Possible Responses and Definitions:

Notes

Percent correct: _____(correct)/_____(total trials)=_____

IOA for this session (attach completed form): _____

Procedural fidelity for this session (attach completed form): _____

Appendix 5.2

Free Operant Timed Event Recording

Data Collector Initials:_____ Date:_____ Condition Description:_____

Circle One: PRIMARY Observer SECONDARY (IOA) Observer

Participant Number: _____

Event	Time Stamp	Behavior
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		

Possible Behaviors and Definitions:

Notes

Number of occurrences of Behavior A (_____): _____

Number of occurrences of Behavior B (_____): _____

IOA for this session (attach completed form): _____

Procedural fidelity for this session (attach completed form): _____

Appendix 5.3

Interval Recording

Data Collector Initials: _____ **Date:** _____ **Condition Description:** _____

Circle One: PRIMARY Observer **SECONDARY (IOA) Observer**

Participant Number: _____ **Interval Recording Type:** MTS PIR

Interval	Behavior	Cumulative Time
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		
21		
22		
23		
24		
25		

Interval	Behavior	Cumulative Time
26		
27		
28		
29		
30		
31		
32		
33		
34		
35		
36		
37		
38		
39		
40		
41		
42		
43		
44		
45		
46		
47		
48		
49		
50		

Possible Codes and Definitions:

Notes

Appendix 5.4

Duration per Occurrence Recording

Data Collector Initials: _____ Date: _____ Condition Description: _____

Circle One: PRIMARY Observer SECONDARY (IOA) Observer

Participant Number: _____

Event	Start	Stop
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		
11		
12		
13		
14		
15		
16		
17		
18		
19		
20		

Behavioral Definitions and Onset/ Offset Rules:

Notes

Number of occurrences _____ Total duration: _____ (add number of seconds)

Duration per occurrence = Total duration/Number of occurrences = _____

IOA for this session (attach completed form): _____

Procedural fidelity for this session (attach completed form): _____

References

Artman, K., Wolery, M., & Yoder, P. (2010). Embracing our visual inspection and analysis tradition: Graphing interobserver agreement data. *Remedial and Special Education*, *33*, 71–77.

Ary, D., & Suen, H. K. (1983). The use of momentary time sampling to assess both frequency and duration of behavior. *Journal of Behavioral Assessment*, *5*, 143–150.

Au, A., Ho, G. S., Choi, E. W., Leung, P., Waye, M. M., Kang, K., & Au, K. Y. (2014). Does it help to train attention in dyslexic children: pilot case studies with a ten-session neurofeedback program. *International Journal on Disability and Human Development*, *13*, 45–54.

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

Barton, E. E., Pribble, L., & Chen, C. (2013). The use of e-mail to deliver performance-based feedback to early childhood practitioners. *Journal of Early Intervention*, *35*, 270–297.

Bushell, D., Wrobel, P. A., & Michaelis, M. L. (1968). Applying "group" contingencies to the classroom study behavior of preschool children. *Journal of Applied Behavior Analysis*, *1*, 55–61.

Chazin, K. T., Bartelmay, D. N., Lambert, J. M., & Houchins-Juarez, N. (2017). Brief report: Clustered forward chaining with embedded mastery probes to teach recipe following. *Journal of Autism and Developmental Disorders*, *47*, 1249–1255.

Chazin, K. T., Ledford, J. R., Barton, E. E., & Osborne, K. O. (2017). The effects of antecedent exercise on engagement during large group activities for young children. *Remedial and Special Education*.doi: 10.1177/0741932517716899

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*, 37–46.

Cooper, J. O. (1981). *Measuring behavior* (2nd ed.). Columbus, OH: Charles E. Merrill.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis*. New York, NY: Pearson.

Crano, W. D., & Brewer, M. B. (2002). *Principles and methods of social research* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Deltaworks Limited. (2016). *Interval Timer* (Version 3.0.3). [Mobile application software]. Retrieved from http://itunes.apple.com

Doke, L. A., & Risley, T. R. (1972). The organization of day-care environments: Required vs. optional activities. *Journal of Applied Behavior Analysis*, *5*, 405–420.

Gentry, D., & Haring, N. (1976). Essentials of performance measurement. In N. G. Haring & L. Brown (Eds.), *Teaching the severely handicapped Volume 1*. New York, NY: Grune and Stratton.

Green, V. A., Drysdale, H., Boelema, T., Smart, E., van der Meer, L., Achmadi, D., …

Lancioni, G. (2013). Use of video modeling to increase positive peer interactions of four preschool children with social skills difficulties. *Education and Treatment of Children*, *36*, 59–85.

Harrop, A., & Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. *Journal of Applied Behavior Analysis*, *19*, 73–77.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis*, *10*, 103–116.

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

Kamps, D., Conklin, C., & Wills, H. (2015). Use of self-management with the CW-FIT group contingency program. *Education and Treatment of Children*, *38*, 1–32.

Kamps, D., Mason, R., Thiemann-Bourque, K., Feldmiller, S., Turcotte, A., & Miller, T. (2014). The use of peer networks to increase communicative acts of students with autism spectrum disorders. *Focus on Autism and Other Developmental Disabilities*, *29*, 230–245.

Kazaroza, A. (2017). *Simple interval timer* (Version 2.0.1) [Mobile application software]. Retrieved from http://itunes.apple.com

Kazdin, A. E. (2010). *Single-case research designs. Methods for clinical and applied settings* (2nd ed). New York, NY: Oxford Press.

Kratochwill, T. R., & Wetzel, R. J. (1977). Interobserver agreement, credibility, and judgment: Some considerations in presenting observer agreement data. *Journal of Applied Behavior Analysis*, *10*, 133–139.

Lane, J. D., & Ledford, J. R. (2014). Using interval-based systems to measure behavior in early childhood special education and early intervention. *Topics in Early Childhood Special Education*, *34*, 83–93.

Leatherby, J. G., Gast, D. L., Wolery, M., & Collins, B. C. (1992). Assessment of reinforcer preferences in multi-handicapped students. *Journal of Developmental and Physical Disabilities*, *4*, 15–36.

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

Ledford, J. R., Ayres, K. A., Lane, J. D., & Lam, M. F. (2015). Accuracy of interval-based measurement systems in single case research. *Journal of Special Education*, *49*, 104–117.

Ledford, J. R., Lane, J. D., Shepley, C., & Kroll, S. (2016). Using teacher-implemented playground interventions to increase engagement, social behaviors, and physical activity for young children with autism. *Focus on Autism and Other Developmental Disabilities*, *31*, 163–173.

Ledford, J. R., & Wolery, M. (2013). The effects of graphing a second observer's data on judgments of functional relations when observer bias may be present. *Journal of Behavioral Education*, *22*, 312–324.

Ledford, J. R., Zimmerman, K. N., Harbin, E. R., & Ward, S. R. (2017). Improving the use of evidence-based instructional practices by paraprofessionals. *Focus on Autism and Other Developmental Disabilities*. doi: 10.1177/1088357617699178

Lloyd, B. P., Weaver, E. S., & Staubitz, J. L. (2016). A review of functional analysis methods conducted in public school classroom settings. *Journal of Behavioral Education*, *25*, 324–356.

Luke, S., Vail, C. O., & Ayres, K. M. (2014). Using antecedent physical activity to increase on-task behavior in young children. *Exceptional Children*, *80*, 489–503.

McCormack, J., & Chalmers, A. (1978). *Early cognitive instruction for the moderately and severely handicapped*. Champaign, IL: Research Press.

Mudford, O. C., Taylor, S. A., & Martin, N. T. (2009). Continuous recording and interobserver agreement algorithms reported in the *Journal of Applied Behavior Analysis* (1995–2005). *Journal of Applied Behavior Analysis*, *42*, 165–169.

Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis*, *8*, 463–469.

Rapp, J. T., Colby, A. M., Vollmer, T. R., Roane, H. S., Lomas, J., & Britton, L. N. (2007). Interval recording for duration events: A re-evaluation. *Behavioral Interventions*, *22*, 319–345.

Reichow, B., Barton, E. E., Good, L., & Wolery, M. (2009). Brief report: Effects of pressure vest usage on engagement and problem behaviors of a young child with developmental delays. *Journal of Autism and Developmental Disorders*, *39*, 1218–1221.

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

Shepley, C., Lane, J. D., & Shepley, S. B. (2016). Teaching young children with social-communication delays to label actions using videos and language expansion models. *Focus on Autism and Other Developmental Disabilities*, *31*, 243–253.

Sutherland, K. S., Alder, N., & Gunter, P. L. (2003). The effect of varying rates of opportunities to respond to academic requests on the classroom behavior of students with EBD. *Journal of Emotional and Behavioral Disorders*, *11*, 240–248.

Tapp, J., & Walden, T. (1993). PROCODER: A professional tape control, coding, and analysis system for behavioral research using videotape. *Behavior Research Methods, Instruments, & Computers*, *25*, 53–56.

Tate, R. L., Rosenkoetter, U., Vohra, S., Horner, R., Kratochwill, T., Sampson, M., … Wilson, B. (2016). Single case reporting guidelines in behavioral interventions (SCRIBE) 2016 statement. *Archives of Scientific Psychology*, *4*, 1–9. doi:10.1037/arc0000026

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

Watkins, M. W., & Pacheco, M. (2000). Interobserver agreement in behavioral research: Importance and calculation. *Journal of Behavioral Education*, *10*, 205–212.

Wehby, J. H., & Hollahan, M. S. (2000). Effects of high-probability requests on the latency to initiate academic tasks. *Journal of Applied Behavior Analysis*, *33*, 259–262.

What Works Clearinghouse. (2014). *Procedures and standards handbook* (Version 3.0). Retrieved from https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_procedures_v3_0_standar

Wolery, M. (February, 2011). Data collection and display. Unpublished presentation. SPED 3013, Vanderbilt University.

Wolery, M., & Garfinkle, A. N. (2002). Measures in intervention research with young children who have autism. *Journal of Autism and Developmental Disorders*, *32*, 463–478.

Wolery, M., & Ledford, J. R. (2013). Monitoring child progress. In M. E. McLean, M. L. Hemmeter, & P. Snyder (Eds.), *Essential elements for assessing infants and preschoolers with special needs*. Boston, MA: Pearson.

Yoder, P. J., Ledford, J. R., Harbison, A., & Tapp, J. (2017). Partial-interval estimation of count. *Journal of Early Intervention.*

Yoder, P., & Symons, F. (2010). *Observational measurement of behavior*. New York, NY: Springer Publishing.

Zimmerman, K. N., & Ledford, J. R. (2017). Evidence for the effectiveness of social narratives: Children without ASD. *Journal of Early Intervention, 39,* 199–217.

Zimmerman, K. N., Ledford, J. R., & Barton, E. E. (2017). Using visual activity schedules for young children with challenging behavior. *Journal of Early Intervention*, *39*, 339–358.

6
Independent Variables, Fidelity, and Social Validity

Erin E. Barton, Hedda Meadan-Kaplansky, and Jennifer R. Ledford

Important Terms

theory of change, procedural fidelity, dosage, control variables, independent variables, implementation fidelity, treatment integrity, checklists, self-reports, direct systematic observation, social validity, blind observers, normative comparisons, maintenance, sustained used, participant preference

Planning and Implementing Study Conditions

Single case design (SCD) research allows for assessment of causal relations between one or more researcher-manipulated independent variables (intervention or treatment) and one or more dependent variables (behaviors). Conducting SCD research involves the process of systematically asking questions, designing and defining an independent variable, implementing conditions as planned, and repeatedly measuring a dependent variable; this process allows experimental control to be established and functional relations to be detected (Horner et al., 2005). Experimental control is a specific relation between participants, dependent variables, and independent variables. In SCD research, establishing experimental control means demonstrating functional relations, which occur when: (a) research is carried out with sufficient experimental rigor, with increased confidence in results if the experiment meets contemporary SCD research standards (Chapters 9–12; Horner et al., 2005; Kratochwill et al., 2013); (b) there is clear documentation that implementation of the independent variable resulted in changes in the dependent variable(s) replicated at three different points in time; and (c) the context is well defined and threats to internal validity are identified and minimized.

Identifying causal relations in SCD studies includes careful planning and implementation of data collection (dependent variables) and implementation (independent variables). Once dependent variables are defined with precision, the independent variable can be planned and designed to produce the hypothesized change (Kennedy, 2005). You should use existing experimental research to drive planning of independent variables; new studies can be designed to answer new questions about or examine adaptations to independent variables that have been studied previously, or test entirely new independent variables that are based on a well-informed theory of change. A **theory of change** refers to a conceptual framework that describes why an intervention should result in changes in a given target behavior. You can use the hypothesized change to formulate research questions and clarify goals of the study (i.e., testing hypotheses). Then, you can use your research questions to select an appropriate SCD (see Chapters 9–12 for descriptions of SCDs) and operationalize (define), plan, order, and implement all conditions such that a functional relation can be detected (Kennedy, 2005).

The consistency of behavior change across planned replications is critical and a primary feature of the analysis of functional relations. When planning SCD research studies, you should carefully plan the order of conditions such that you have the opportunity to demonstrate and replicate behavior change. Depending on the design, study conditions might include baseline, intervention(s), generalization, and maintenance conditions (see Chapters 9–12 for descriptions of specific SCDs). *All* experimental procedures—defined by specific study conditions—are operationalized to allow for valid interpretations of results and facilitate future replications. Given that

study conditions are repeatedly manipulated and implemented by humans, are implemented repeatedly over time, and might change rapidly, appropriately documenting that study conditions and all experimental procedures were implemented as planned and operationalized is required (Horner et al., 2005; Ledford & Wolery, 2013a; Wolery, 2013).

Measurement of Fidelity

Procedural fidelity is the degree to which procedures of all experimental conditions are implemented as intended (Ledford & Wolery, 2013a; Wolery, 2011). Measurement of procedural fidelity has been suggested as a necessary component of SCD research to identify implementation errors made by researchers for both formative and summative purposes (e.g., Billingsley, White, & Munson, 1980; Wolery, 2011). When procedural fidelity data are carefully collected and reported, it is possible to use these data to (a) make decisions regarding the likelihood of adequate implementation in regular environments, (b) determine sufficiency of interventions implemented with low fidelity (e.g., If a practitioner implements an intervention correctly only 70% of the time, will it still be effective?), and (c) explain variability in results when this variability is related to inconsistency in intervention implementation (Fettig, Schultz, & Sreckovic, 2015; Wood, Ferro, Umbreit, & Liaupsin, 2011; Wood, Umbreit, Liaupsin, & Gresham, 2007). Conversely, absence of fidelity data limits conclusions regarding whether experimental procedures were implemented properly and at sufficient **dosage** levels (i.e., Were procedures implemented each time they should have been implemented, for as long as they should have been implemented?), which is important for *both* group and SCD research (Kratochwill et al., 2013; Ludemann, Power, & Hoffman, 2017). Adequate measurement of procedural fidelity *and* documentation that the procedures were implemented as intended across experimental conditions is required to document causal relations. Procedural *in*fidelity is a major risk of bias in both group and SCD research (Ludemann et al., 2017; Wolery, 2013). Risk of bias refers to "believability" of a research study or the extent to which threats to internal validity have been controlled for or minimized (Reichow, Barton, & Maggin, 2017; see Chapter 1 for a discussion of internal validity in SCD research and Chapter 13 for additional information regarding risk of bias). Studies that adequately measure and report procedural fidelity have low risk of bias on this dimension; other study components (e.g., dependent variable reliability) should be evaluated separately.

Defining Experimental Conditions

Operationalizing experimental conditions requires defining the exact procedures, parameters, and processes of the experimental conditions. Procedures are the components (what is done), parameters are related to dosage (quantification of components), and processes can be identified as trainings required for implementation. Some procedures are the same across conditions (**control variables**); others are differentially implemented across conditions (**independent variable**). Control variables, unlike independent variables, are either always present or absent to the same degree in

every condition (including baseline), and will not change when the independent variable is manipulated. They typically include contextual, set-up, and session completion variables. Table 6.1 lists examples of independent and control variables that should be measured for two common interventions. Independent and control variables have to be planned, measured, and reported with replicable precision. Reliability of independent and control variables should be measured and documented in all experimental conditions (Ledford & Wolery, 2013a). Clear documentation that control variables *did not change* and independent variables *did change*, and corresponded to changes in dependent variables is required to establish experimental control. You should conduct ongoing measurement of control and independent variables to ensure that changes across conditions occurred for the planned independent variable and *only* for the independent variable.

Table 6.1 Control and Independent Variables for Two Common Interventions.

Planned Step	Variable Type	Implementation Condition(s)
Time delay procedures (CTD, PTD) to teach discrete academic skills		
Present stimulus	Control	Baseline, Intervention
Give task direction	Control	Baseline, Intervention
Provide prompt	*Independent*	Intervention
Wait interval	Control	Baseline, Intervention
Give reinforcement for correct response	Control	Baseline, Intervention
Provide 10 trials	Control	Baseline, Intervention
Differential reinforcement for other behaviors (DRO) to decrease aggression during free play		
Provide 10 preferred play materials	Control	Baseline, Intervention
Provide 5 equally-spaced task demands	Control	Baseline, Intervention
Provide reinforcement for engaging in behaviors other than aggression	*Independent*	Intervention
End session after 10 minutes	Control	Baseline, Intervention

Implementation Fidelity

When a non-researcher implements procedures in the context of SCD studies, there are two potential "levels" of fidelity. First, we must, as usual, document that conditions (e.g., baseline, intervention) were implemented as intended—procedural fidelity. A second

important characteristic is whether the *researcher* implemented training as intended. For example, if you intend to train a teacher to implement a systematic prompting procedure, you would want to plan (a) a systematic training for the teacher, and (b) systematic procedures for use during baseline and intervention sessions. Then, you would measure whether *both* the training and experimental conditions were implemented as intended. **Implementation fidelity**, the extent to which experimenters trained implementers as planned, has been under-reported in SCD research (Dunst, Trivette, & Raab, 2013; Fettig & Barton, 2014). However, these training data provide important information regarding feasibility and replicability of experimental procedures. Thus, the fidelity of any didactic training (e.g., workshop) and ongoing coaching (e.g., live feedback during each session) provided to implementers should be planned, measured, and reported as implementation fidelity. You should also provide experience, training, and demographic characteristics of indigenous implementers to ensure the study can be replicated (Dunst et al., 2013; Sutherland, McLeod, Conroy, & Cox, 2013). Appendix 6.1 provides an example of a form that might be used to measure implementation fidelity of commonly used types of didactic training; Table 6.2 describes differences among training, independent, and control variables.

Table 6.2 Fidelity Measurement.

Type	Description	Measurement
Implementation Fidelity	Measurement of training variables. (Was training of implementers conducted as intended?)	Should be measured when researchers train implementers; not generally measured when expert researchers implement conditions without training.
Procedural Fidelity	Measurement of independent and control variables in baseline and intervention conditions. (Were all experimental conditions conducted as intended?)	Should be measured in all studies.
Treatment Integrity (also referred to as Treatment Fidelity)	Measurement of independent variables in intervention conditions only. (Was the intervention conducted as intended?)	Should not be measured because limits analysis of differentiation.

Adherence and Differentiation

In group research design, procedural fidelity is critical to determine both that participants in the intervention group received the intervention as intended *and* that participants in the control group did not receive the intervention (Wolery, 2011). Likewise, in SCD research, procedural fidelity provides evidence that the independent variable was implemented as intended *and* not present (or present at low levels) during baseline or control conditions. Procedural fidelity should provide two types of evidence: *adherence* (you implemented the intervention as planned) and *differentiation* (you implemented different steps in each condition; Ledford & Wolery, 2013a; Sutherland et al., 2013). Adherence to the protocol provides evidence that the intervention was delivered as planned, and refers to how closely implementer behavior mirrored prescribed procedures. Differentiation refers to differences between experimental conditions (typically baseline and intervention), and provides evidence that procedures between conditions were implemented differently from one another. Or more specifically, provides evidence that control variables were implemented or present in the same manner, and independent variables were only implemented during experimental conditions. Measurement of variables across all experimental conditions is necessary to determine both that independent variables were used correctly during intervention *and* that no other changes occurred between baseline and intervention conditions. Both conclusions are essential for increased confidence that results are due to planned and controlled changes between conditions. Example 6–1 describes a study reporting adherence but not differentiation (Barton, 2015); Example 6–2 describes a study measuring both adherence and differentiation across all study conditions (Ledford & Wolery, 2013b).

Applied Example 6–1

Barton, E. E. (2015). Teaching generalized pretend play and related behaviors to young children with disabilities. *Exceptional Children*, *81*, 489–506.

In this study, Barton examined the relation between preschool teachers' use of the system of least prompts and contingent imitation and acquisition, maintenance, and generalization of pretend play and related behaviors by four children with disabilities. She measured and reported implementation fidelity of didactic teacher training and the coach's use of feedback before, during, and after intervention sessions with teacher implementers. She used direct systematic observation to record teachers' use of intervention procedures. Although she reported that teachers implemented the intervention with fidelity during intervention conditions, she did not provide information regarding teacher behaviors during probe conditions. Thus, adherence data are provided, but differentiation with probe conditions is not clear. This reduces confidence in functional relations and increases risk of bias.

Types and Measurement of Fidelity

Measurement of fidelity has increased in recent years, but the percentage of articles that report fidelity data varies widely (Barton & Fettig, 2013; Ledford & Wolery, 2013a). For example, the degree and timing of measurement varies across studies. A similar term, ***treatment integrity***, is used when data are collected (on independent or control variables) during intervention conditions *only*, such that assessment of adherence is possible, but not differentiation. Thus, we recommend measurement of fidelity across all conditions, including implementer training (i.e., procedural fidelity and implementation fidelity). Procedural fidelity data should be collected frequently (20%–33% of sessions) in all conditions by an independent observer.

Applied Example 6–2

Ledford, J. R., & Wolery, M. (2013b). Peer modeling of academic and social behaviors during small-group direct instruction. *Exceptional Children*, *79*, 439–458.

In this study, Ledford and Wolery (2013b) examined the relation between use of progressive time delay within a small group setting on academic and social behaviors of young children with and without disabilities. Researchers used direct systematic observation to measure and report procedural fidelity. They measured implementation of study procedures during all conditions—probe, instruction, and generalization. This allowed for analysis of both adherence and differentiation data; confidence in functional relation is high and risk of bias is low.

Formative Analysis

Formative analysis of procedural fidelity can be used to evaluate ongoing needs and identify when to provide additional training to the implementer. Formative analysis also allows you to detect and minimize threats to internal validity resulting from inaccurate or inconsistent implementation. However, the type of measurement used for procedural fidelity is critical to ensure this evaluation can occur; common types include direct systematic observation, checklists, and self-reports. Although they are often used, **checklists** (i.e., dichotomous yes/no measurement for behaviors that may occur multiple times per session) may not be sensitive to intermittent errors made by implementers. Therefore, use of checklists should be limited to binary variables or procedures/behaviors that are expected to occur once per session. **Self-reports** (i.e., implementers measuring their own implementation) have been shown to have low validity for measuring fidelity because implementers typically overestimate accuracy of their own behaviors (Lane, Kalberg, Bruhn, Mahoney, & Driscoll, 2008; Martino, Ball, Nich, Frankforter, & Carroll, 2009). Thus, **direct systematic observation** of implementer behaviors is preferable and recommended (i.e., counting whether an implementer used the behavior in the manner in which it was intended, as often as intended). Checklists, self-reports, and direct systematic observation also can be used in combination. For example, the implementation fidelity form shown in Appendix 6.1 uses both a checklist and direct systematic observation. Examples 6–3 and 6–4 describe studies that used direct systematic observation of implementer behaviors across study conditions to measure procedural fidelity. In a review of procedural fidelity features, Ledford and Wolery (2013a) found that only 40% of studies use this type of measurement and that use of direct counts has actually decreased over time. Example 6–4 describes use of direct systematic observation to record procedural fidelity in an SCD study (Barton, Pokorski,

Sweeney, & Velez, 2017); Appendix 6.2 is the data collection form used in this study. Appendix 6.3 is an example of a data collection form to measure procedural fidelity for a constant time delay procedure. Appendices 6–2 and 6–3 include set up, contextual, and repeated use variables with both checklist and direct systematic observation. Each form facilitates an analysis of adherence and differentiation when completed across experimental conditions. Regardless of which measurement system is used, you should carefully consider which variables should be measured (e.g., which independent variables will differ across conditions, which other variables with potential to influence dependent variables should remain constant across conditions). Further, if procedural infidelity occurs, you should systematically re-train implementers and closely monitor fidelity.

Applied Example 6–3

Pennington, R. C., Stenhoff, D. M., Gibson, J., & Ballou, K. (2012). Using simultaneous prompting to teach computer-based story writing to a student with autism. *Education and Treatment of Children*, *35*, 389–406.

Pennington, Stenhoff, Gibson, and Ballou (2012) examined use of simultaneous prompting via a computer to teach story writing to a child with autism. They operationalized all study conditions (i.e., baseline, intervention, maintenance) and measured all implementer procedures and behaviors. The data collection included binary (yes/no) responses and direct systematic observation of implementer behaviors across conditions. Although they referred to this measurement as "treatment integrity" (p. 398), they measured what we refer to as "procedural fidelity." They measure and report fidelity such that adherence to study procedures and differentiation between conditions can be documented. However, it is unclear if procedural fidelity was measured across more than 20% of sessions for all conditions, which limits confidence in results.

Applied Example 6–4

Barton, E. E., Pokorski, E. A., Sweeney, E. M., & Velez, M. (2017). The use of the system of least prompts to teach board game play within small groups of young children. *Journal of Positive Behavior Interventions.*

Barton et al. (2017) examined use of the system of least prompts to teach preschool children with and without disabilities to play board games. They operationalized all study conditions (i.e., probe, intervention, generalization) and measured all implementer procedures and behaviors. Appendix 6.2 represents the data collection form they used across study conditions. This allowed for measurement of adherence to study procedures and differentiation between conditions. Further, they measured fidelity for a minimum of 25% of sessions per condition for all four participants. Confidence in the documentation of a functional relation is high and risk of bias is low.

Summative Analysis

Summative analysis of procedural fidelity increases internal validity (reduces risk of bias) of the study and can be used to describe variability in the dependent variable (Wood et al., 2011). Summative analysis of procedural fidelity should occur for each participant in a study to confirm implementation accuracy did not vary among participants (Moncher & Prinz, 1991) or to determine whether child outcomes are related to differential implementation (e.g., Was implementation more accurate for a child with optimal outcomes when compared with a child with more variable or less accurate implementation?). Summative analysis allows you to document that the intervention was implemented as planned and precisely describe conditions under which it was effective. This provides a foundation for recommendations about circumstances under which an independent variable is likely to work and promotes experimental replications. For example, procedural fidelity measurement might form the basis for parametric comparisons (i.e., studies examining high and low procedural fidelity), component analyses (e.g., to construct and deconstruct multicomponent interventions), or feasibility studies (e.g., identifying behaviors indigenous implementers are most likely to use with accuracy).

Reporting Fidelity

Even when all experimental procedures and variables are adequately assessed, researchers might not report sufficient information to allow readers to analyze data. You

should separately report each procedural step (implementer behavior) for which data are collected. These data can be presented in a table, or, in the case of consistently high fidelity, you can identify each behavior and report that each behavior was implemented with adequate fidelity. In addition, you should explicitly report during which conditions and for which participants' fidelity data were collected and to what extent (e.g., during how many sessions) data were collected during each condition and for each participant. Implementer behaviors designed to change (independent variables) should be measured, as well as behaviors designed to remain constant across conditions (control variables; Ledford & Wolery, 2013a). Authors also should measure and report physical and social conditions in each experimental condition. At minimum, condition descriptions should include procedural steps and rules, length and frequency of measurement occasions, and environmental characteristics (location, physical size and arrangement, social context). Description of implementers should include role (classroom teacher, researcher), education and experience, specific intervention training, and demographic data. Example 6–5 describes a study that measured and reported procedural fidelity and adequately described implementers.

The following steps can be used when designing a procedural fidelity measurement system:

1. Define procedures of all experimental conditions (consider the specific procedural components [actions] and parameters [frequency]). Ensure procedures or behaviors that are expected to occur more than once per session are counted (rather than simply noted as present or absent, binary measurement). Identify the following variables:
    1. Setup and completion variables (might be binary)
    2. Contextual variables
    3. Repeated use control variables
    4. Repeated use independent variables
    5. Implementer training (when there is an indigenous implementer)
2. Identify the intended frequency for each behavior (might be a range or contingent on the participants behavior)
3. Develop a systematic and practical measurement system
    1. Use **checklist** for binary behaviors or procedures
    2. Use a **direct systematic observational** system for repeated-use behaviors
    3. Measure independent (different between conditions) and control (the same across conditions) variables
    4. Measure implementer training variables when there is an indigenous implementer
4. Identify a plan for formative evaluation of procedural fidelity data across conditions and participants (and tiers); collect procedural fidelity data
    1. Collect and monitor implementation fidelity data when using an indigenous implementer (e.g., collect data on whether you trained the implementer according to your protocol).
    2. Regularly collect data on the occurrence of procedures—for all conditions and all participants (20-33% of sessions for each condition and each participant and more often if possible).
5. Compare to planned components and parameters; calculate percentages
    1. Calculate the percentage of correct use for each variable and each participant
    2. Calculate percentage of correct implementation by variable and condition
6. Use data in formative evaluation
    1. If low, usually re-train implementers
    2. Identify areas of incorrect implementation
    3. Monitor imprecise implementation
    4. Reinforce correct implementation
7. Use data in summative evaluation
    1. If low, compare to variability in DV
    2. Document levels of use of the IV to identify conditions under which effects were observed
    3. Make recommendations for future study or for application of IV in practice.

Figure 6.1 Task analysis for designing a comprehensive procedural fidelity measurement system.

In sum, we recommend the following: (a) measuring all experimental variables, conditions, participants, and levels of implementation (i.e., procedural fidelity); (b) using direct systematic observations (counts derived from direct observation); and (c) reporting explicitly (e.g., naming variables, conditions, and participants for which data were collected). Figure 6.1 provides a task analysis for designing a comprehensive procedural fidelity measurement system.

Applied Example 6–5

Ledford, J. R., Zimmerman, K. N., Chazin, K. T., Patel, N. M., Morales, V. A., & Bennett, B. P. (2017). Coaching paraprofessionals to promote engagement and social interactions during small group activities. *Journal of Behavioral Education*, *26*, 410–432. doi: 10.1007/s10864-017-9273-8

Ledford and colleagues (2017) examined use of in situ coaching and performance-based feedback on use of environmental arrangement, prompting, and praise by three paraprofessionals in preschool classrooms. They operationalized all study conditions (i.e., baseline, intervention, maintenance, enhanced maintenance, and generalization) and measured all implementer (i.e., coaching) procedures and behaviors. Appendix 6.4 represents the data collection form used across study conditions. This allowed for measurement of adherence to study procedures and differentiation between conditions. They measured adherence to coaching during intervention sessions, and absence of coaching during baseline and generalization sessions. Further, they measured fidelity for a minimum of 40% of sessions per condition for all participants. They also described experience and demographics of coaches (researchers). Confidence in documentation of a functional relation is high and risk of bias is low.

Social Validity

In their seminal paper, Baer, Wolf, and Risley (1968) discussed important dimensions of behavioral research. One important quality was that such work should be *applied*—dependent variables targeted for change should be socially important. In their update on behavioral research, published nearly 20 years later (Baer, Wolf, & Risley, 1987), they argued that social validity, "the extent to which all the consumers of an intervention like it" (p. 322), was a secondary measure of effectiveness in behavioral sciences. Kazdin (1977) defined social validity as the presence of "changes in behavior that are clinically significant or actually make a difference in the client's life" (p. 427) and Wolf (1978) argued that subjective feedback data have a place in applied research. More recently, Horner et al. (2005) included social validity as one of the quality indicators for SCD research and stated that the social validity of SCD research could be enhanced by (a) selecting dependent variables that are socially important; (b) demonstrating that independent variables can be applied with fidelity by indigenous implementers in typical contexts; (c) demonstrating indigenous implementers report the intervention is feasible, effective, and will be maintained; and (d) demonstrating the intervention is effective.

The evaluation of social significance should be completed by a variety of stakeholders. Schwartz and Baer (1991) described four groups of stakeholders that could be involved in evaluation of the social validity of an intervention or program. These groups were: (a) direct consumers—recipients of the intervention (e.g., children, teachers, parents, administrators); (b) indirect consumers—people who could be affected by the intervention, but are not direct recipients (e.g., parents and peers of direct participants); (c) members of the immediate community—people who interact regularly with direct and indirect consumers (e.g., neighbors of participants); and (d) members of the extended community—people who may not know direct recipients, but live in the same community (e.g., librarian at the local library).

We recommend that you collect data from multiple stakeholders to understand the social validity of the intervention from different perspectives. For example, suppose you develop an intervention to increase the reading fluency of Max, a middle-school student with specific learning disabilities. You could collect social validity data from Max (the direct consumer); his parents, teachers, and peers (indirect consumers); other teachers and parents of middle-school students (members of immediate community); and taxpayers who make recommendations for additional funding for school programs (members of extended community).

Dimensions of Social Validity

Wolf (1978) recommended that researchers address three levels of social validation:

goals, procedures, and outcomes. A study in which all three dimensions were measured is described in Example 6–6.

Goals

Were the goals socially important? In other words, are we teaching participants to engage in behaviors that are valued by society? Is increasing or decreasing a specific behavior important for quality of life of participants and people who interact with them? Some goals have wide support for importance (e.g., a number of prosocial behaviors; Hurley, Wehby, & Feurer, 2010). The social validity of other goals, like teaching academic skills to youth with severe disabilities, has been recently contested, separate from arguments of effectiveness (e.g., Ayres, Lowery, Douglas, & Sievers, 2011, 2012; Courtade, Spooner, Browder, & Jimenez, 2012). We note that although it is generally preferable to teach socially valuable skills, it is sometimes necessary to use less immediately useful behaviors to adequately control for threats to internal validity (e.g., if a researcher was aware that a child was struggling to learn letter names, but also knew that skill was being explicitly targeted in his classroom, she might decide to teach him letter *sounds* instead, to avoid potential history effects). For example, answering the question of whether one teaching procedure is more effective than another procedure is potentially highly important for a child (i.e., has potential to result in beneficial outcomes). However, to compare these procedures, identifying behaviors the child is unlikely to be exposed to elsewhere (e.g., school, home) is necessary to prevent history effects. Thus, sometimes skills taught are somewhat irrelevant or extraneous, in order to better test the question of interest.

Procedures

Were the procedures socially acceptable? This facet of social validity is often referred to as "treatment acceptability." In other words, is it feasible to implement the intervention the way it was designed? Is it acceptable in terms of cost, time, efforts, ethics, and appearance? If procedures are not feasible for use in the intended context, required too much time, or were perceived as unethical, they may be unlikely to be initiated or used by indigenous implementers. Further, some have argued that socially accepted procedures are more likely to be correctly implemented by indigenous implementers (e.g., Baer et al., 1987; Perpletchikova & Kazdin, 2005).

Outcomes

Are the outcomes socially significant? This dimension of social validity relates to perceived effectiveness of the intervention and satisfaction of consumers with results. In other words, are outcomes of the intervention meaningful and important to consumers?

For example, a consistent change from 20 to 15 talk-outs per class session for a high school student may be experimentally significant, but 15 talk-outs may still be far too high to allow for effective instruction (e.g., socially insignificant; not socially valid).

Applied Example 6–6

Weng, P. L., & Bouck, E. C. (2014). Using video prompting via iPads to teach price comparison to adolescents with autism. *Research in Autism Spectrum Disorders*, *8*, 1405–1415.

Weng and Bouck (2014) examined effectiveness of video prompting to teach price comparison to three secondary students with autism. To evaluate social validity of the intervention, researchers conducted **interviews** *before* and *after* the study with the students with autism (**direct consumers**) and their teachers (**indirect consumers**). Two students responded to interview questions verbally and by pointing to pictures and the third student used his AAC device to answer questions. The interview questions were focused on social validity of the **goals** (the social importance of teaching price comparison), **procedures** (acceptability of video prompting), and **outcomes** (the effectiveness of the intervention in increasing students' skills). Although this provided a comprehensive evaluation of social validity of the intervention, the subjective nature of the measurement might have led to biased responses. Additional measures that are less subject to bias might increase confidence in social validity of the intervention. These are discussed in the next sections.

Assessment of Social Validity

Researchers have used different methods and tools to assess social validity. Each method focuses on a different aspect of social validity and we recommend use of more than one method or tool when evaluating social validity of an intervention; Table 6.3 provides information regarding different ways to assess each dimension of social validity.

Typical Subjective Measures

Subjective measures are used to gather information from different stakeholders related to their perspectives on social importance of goals, procedures, and outcomes of an intervention. You can use *interviews*, *questionnaires*, and *rating scales* to collect subjective evaluation data on social validity. We recommend a person who is not directly engaged in implementation of the intervention collect subjective measures (e.g., conduct interviews) to reduce social desirability bias. Recognition that this type of social validity measurement could be potentially problematic and subjective has been discussed for many years; however, researchers also recognized that this did not render social validity unimportant (e.g., Wolf, 1978). There are some data suggesting social validity

ratings are related to fidelity of implementation and self-reported use of interventions; thus, these data may have long-term impacts on whether implementers continue interventions after a research study is completed (e.g., Carter & Pesko, 2008; Perpletchikova & Kazdin, 2005; Wehby, Maggin, Partin, & Robertson, 2012). Strain, Barton, and Dunlap (2012) reviewed results from several unrelated studies and suggested that social validity results were not necessarily predictable or related to change in outcome measures. However, measurement of social validity may provide additional information for a better understanding of the intervention as a whole. Chung, Snodgrass, Meadan, Akamoglu, and Halle (2016) described differences between behavioral observation data and social validity data from interviews and emphasized the importance of valuing both graphed observation data and subjective measures of social validity in intervention research.

Table 6.3 Descriptions and Types of Social Validity Measurement in Single Case Research.

	Social Importance of the Goals	Acceptability of the Procedures	Social Importance of the Outcomes
Participant or Stakeholder Judgments (potentially subject to bias)			
Purpose	Gather opinions of stakeholders regarding skills they value	Assess stakeholder opinions regarding the feasibility and appropriateness of procedures	Assess stakeholder opinions regarding whether behavior changes were important
Assessment	In-depth interviews, questionnaires, rating scales	In-depth interviews, questionnaires, rating scales	In-depth interviews, questionnaires, rating scales
Schedule	Pre-intervention	Pre-intervention, post-intervention	Post-intervention
Normative Comparison (less subject to bias)			
Purpose	Identify behaviors of target children distinguishing them from demographically similar peers	Identify procedures used with children with similar characteristics, targeting the same or similar behaviors	Evaluate intervention outcomes by comparing behaviors of target participants to those of peers
Assessment	Formal assessments, behavioral observations	Literature review	Formal assessments, behavioral observations
Schedule	Pre-intervention	Pre-intervention	Post-intervention

Blind Ratings (less subject to bias)			
Purpose	Identify behaviors of target children distinguishing them from demographically similar peers, using raters who do not know the study purpose	Survey the opinions of raters who do not know the study purpose regarding the acceptability and feasibility of the procedures	Assess opinions of raters who do not know the study purpose regarding whether behavior changes were apparent
Assessment	Pre-intervention video ratings	Intervention video ratings	Pre- and post-intervention video ratings
Schedule	Post-intervention	Post-intervention	Post-intervention
Maintenance or Sustained Use (less subject to bias)			
Purpose	Evaluate the continued use of target behaviors after the intervention is completed	Evaluate the use of procedures after the intervention is completed	Evaluate the continued use of target behaviors after the intervention is completed
Assessment	Behavioral observation	Behavioral observation	Behavioral observation
Schedule	Post-intervention	Post-intervention	Post-intervention
Participant Preference (less subject to bias)			
Purpose	NA	Assess which condition the participant prefers	NA
Assessment	NA	Interview, questionnaire, choice-making/preference assessment	NA
Schedule	NA	Pre-intervention, during the course of the study (some designs)	NA

Although objective quantification of social validity data is possible, we continue to analyze social validity primarily using subjective measures not based on observational data (e.g., Kamps et al., 1998, Snodgrass, Chung, Meadan & Halle, 2017). However, measurement of objective data related to social validity is preferred because "subjective data may not have any relationship to actual events" (Wolf, 1978, p. 212). Although considered theoretically important, social validity analysis in applied research may not

be prevalent (Carr, Austin, Britton, Kellum, & Bailey, 1999; Snodgrass et al., 2017) because subjective measures can be insensitive to effects (Kennedy, 2002). For example, if you conduct daily intervention sessions with a child in his home for six weeks, a parent might report favorable opinions regarding goals, procedures, and outcomes of the treatment because of the desire to please you (for discussions about this problem with subjective reports, see Hurley, 2012 or Garfinkle & Schwartz, 2002). Despite this weakness, one review found social validity was primarily assessed via self-reported satisfaction and was rarely assessed objectively (in only 1 of 90 reviewed studies; Hurley, 2012).

Measures Less Subject to Bias

There are at least four types of social validity measurement that are less subject to bias and we recommend their use: (a) normative comparisons (Rapoff, 2010; Houten, 1979), (b) blind ratings (Meadan, Angell, Stoner, & Daczewitz, 2014), (c) measurement of maintenance or sustained use (Kennedy, 2005), and (d) participant preference measurement (Hanley, 2010). Although these suggestions are not new, we believe expanded use and feasibility of video recording as well as emphasis on research in typical environments makes these procedures increasingly relevant.

NORMATIVE COMPARISONS

When **normative comparisons** are used, the participants' targeted behavior (i.e., dependent variable) is compared to a normative or 'typical' group whose behavior is considered acceptable. Data for both the target participants and the normative group are collected and compared. Normative comparisons can be helpful to determine: (a) what intervention goals are socially important and (b) whether participants reached typical or acceptable levels of the target behavior following intervention. For example, Smith and Van Houten (1996) compared behavior of children with developmental delays to behavior of typically-developing children. All children exhibited some stereotypy; these data could be used in subsequent studies to determine a socially acceptable level of stereotypy. Another example of the use of normative comparisons, in addition to other social validity measures, is reported in Example 6–7.

Applied Example 6–7

Hochman, J. M., Carter, E. W., Bottema-Beutel, K., Harvey, M. N., & Gustafson, J. R. (2015). Efficacy of peer networks to increase social connections among high school students with and without autism spectrum disorder. *Exceptional Children*, *82*, 96–116.

Hochman and her colleagues (2015) examined effects of a lunchtime peer network intervention on social engagement and peer interactions of four adolescent students with autism. To evaluate social validity of the intervention, researchers asked the students with autism (direct consumers) and their parents (indirect consumers) to complete a survey that included both Likert-type and open-ended questions (subjective measures). Adults who served as network facilitators were also asked to complete a survey with Likert-type and open-ended questions. All surveys were completed at the *end* of the study and focused on social validity of goals, procedures, and outcomes. In addition, researchers selected three different male peers without disabilities for each focus student and observed them during an entire lunch period. Direct observations of peers were used to establish a range of typical social interaction (normative comparison).

MAINTENANCE OR SUSTAINED USE DATA

Maintenance or **sustained use** data are measures used to evaluate if procedures and outcomes of an intervention continue after the research is completed (Kennedy, 2005). Although rarely used, this important measure is related to likelihood of maintained and generalized behaviors, especially when indigenous implementers are trained to use intervention procedures. If an intervention is effective in changing participant behavior, but indigenous implementers do not continue its use once a study is completed, likelihood of maintained behavior change is low. Thus, measurement of continued use by practitioners and caregivers in typical environments is an important measure of social validity that answers the question of how acceptable (and feasible) stakeholders find procedures. Although measures of treatment acceptability (e.g., rating scales) are often used in place of direct measurement, at least one study has shown that these two measures do not necessarily agree (Farmer, Wolery, Gast, & Page, 1988).

BLIND RATINGS

Blind ratings can be used to less subjectively determine whether participants' behavior is rated as "different" before and after intervention or during baseline versus intervention conditions (socially important outcomes) by people who are unaware of the condition in effect for the session(s) they watch (e.g., pre- or post-intervention, baseline

or treatment conditions) and/or the purpose of the study. These ratings can also be used to determine whether procedures in one condition are more acceptable than those in another (socially acceptable procedures). In one study, music and no-music conditions were compared when used to teach signed vocabulary words to toddlers with disabilities. Graduate students who were blind to condition type rated muted videos of both types of sessions (ones with and without music) regarding whether participants appeared happy. Although acquisition results were similar, one participant appeared happier during music conditions; this finding may suggest that musical interventions may be both equally effective and more socially acceptable for some young children (Koutsavalis, 2011). Another example of use of blind raters, in addition to other social validity measures, is reported in Examples 6–8 and 6–9.

Applied Example 6–8

Meadan, H., Stoner, J. B., Angell, M. E., Daczewitz, M., Cheema, J., & Rugutt, J. K. (2014). Do you see a difference? Evaluating outcomes of a parent-implemented intervention. *Journal of Developmental and Physical Disabilities*, *26*, 415–430.

Meadan and her colleagues (2014) developed a parent-implemented communication strategies (PiCS) intervention and examined effectiveness of the intervention on parents' (fidelity of implementation) and children's (responding and initiating) behavior. To evaluate social validity of goals, procedures, and outcomes, parents completed a researcher- developed Likert-type questionnaire *before* and *after* the intervention and participated in an interview conducted by an external evaluator of the project at the *end* of the project (subjective measures). In addition, researchers randomly selected 2-minute video clips of parent-child interaction from pre- and post-intervention conditions. Three group of adult raters, who were unaware of participants and intervention status (blind ratings), participated: (a) parents of young children with disabilities, (b) special education teachers who work with young children with disabilities, and (c) speech language pathologists who work with young children. Each rater watched pre- and post-intervention video clips, in random order, and evaluated parent and child behavior.

Applied Example 6–9

Bailey, K. M., & Blair, K. S. (2015). Feasibility and potential efficacy of the family-centered Prevent-Teach-Reinforce model with families of children with developmental disorders. *Research in Developmental Disabilities*, *47*, 218–233.

Bailey and Blair (2015) examined feasibility and efficacy of a family-centered Prevent- Teach-Reinforce (PTR) model with three families of young children with disabilities. Researchers used three methods to evaluate social validity of goals, procedures, and outcomes. Parents (direct consumers) were asked to complete a self-rating form to measure perceived effectiveness and acceptability of the intervention (a modified version of the PTR Self Evaluation: Social Validity form). Parents also participated in an interview at the *end* of the intervention to examine their satisfaction with procedures and outcomes and their plan to continue using the model (subjective measures). In addition, two participants unaware of study purpose and goals reviewed the behavior intervention plans and viewed three random videos of baseline and intervention conditions, to rate parent and child behavior (blind rating).

PARTICIPANT PREFERENCE

Participant preference for interventions has typically been measured using rating scales or post-intervention questionnaires. Objective measurement of participant preference during intervention implementation is both possible and preferable, even for young children or those who have significant language or cognitive impairments (Hanley, 2010). This is perhaps the most important measure of whether intervention procedures are acceptable to the primary consumer (i.e., the recipient of the intervention). For example, this type of measurement can be assessed experimentally in the context of a simultaneous treatments design. Simultaneous treatments designs compare participant choice or preferences for two or more intervention conditions, which are concurrently available across sessions (see Chapter 12 for additional information regarding simultaneous treatment designs). Although participant ratings of acceptability as a social validity measure are far less common than other stakeholders' (e.g., parents, teachers; Hurley, 2012) ratings, when intervention strategies are similarly effective for a participant, preference for intervention is crucial information for interventionists (Ledford, Chazin, Harbin, & Ward, 2017; State & Kern, 2012). For example, Heal and Hanley (2011) measured participant preference for three play-based interventions by allowing participants to choose which intervention they wanted to receive for each session: (a) instruction embedded in play, (b) pre-session modeling then play, or (c) pre-session direct instruction then play. Pre-session direct instruction not only led to greater acquisition of targeted information, but also was chosen most often by participants as

the most highly preferred intervention. Objective preference procedures such as this one may result in more valid results, especially for participants who are young or those who have cognitive or language impairments.

Summary

This chapter described two major components of SCD research: procedural fidelity and social validity. Both of these components provide information regarding use and feasibility of independent variable(s). Procedural fidelity is also critical for decreasing risk of bias. Although procedural fidelity data have historically been inadequately reported, comprehensive procedural fidelity assessment using appropriate measures is critical for documenting functional relations. Social validity measures can provide interesting and important information; however, using measures that are less subject to bias (e.g., normative comparisons) may provide more valid evidence of the extent to which goals, procedures, and outcomes are socially acceptable. In addition, social validity data should be analyzed using rigorous methods (e.g., qualitative analysis of interview data), clearly presented, and included in discussion of the study. Additional research is needed on the extent to which different types of social validity measures result in similar conclusions and the degree to which each correlates with outcome measures, procedural fidelity, and maintained use of procedures.

Appendix 6.1

Implementation Fidelity—Teacher Training

Date:_____ Data Collector:_____ Trainer_____ Location:_____

Total # Trainees at Start		Total # Trainees at End	
Start time		End time	
Agenda presented	Yes No	Trainer introduced self	Yes No
Training purpose identified	Yes No	Trainer introduced trainees	Yes No
Handouts provided	Yes No	Trainer describes handouts	Yes No

Training Objectives/Skills	**Skill/Obj. described**	**Handout shown**	**Skill modeled**	**Prompt to practice skill**	**Feedback provided**	**Skill mastery evaluated**
Objective/skill 1:	Yes No	Yes No	Yes No	Yes No	Yes No	Yes No
Objective/skill 2:	Yes No	Yes No	Yes No	Yes No	Yes No	Yes No
Objective/skill 3:	Yes No	Yes No	Yes No	Yes No	Yes No	Yes No
Objective/skill 4:	Yes No	Yes No	Yes No	Yes No	Yes No	Yes No
Objective/skill 5:	Yes No	Yes No	Yes No	Yes No	Yes No	Yes No
Objective/skill 6:	Yes No	Yes No	Yes No	Yes No	Yes No	Yes No
Objective/skill 7:	Yes No	Yes No	Yes No	Yes No	Yes No	Yes No
Objective/skill 8:	Yes No	Yes No	Yes No	Yes No	Yes No	Yes No

# of questions asked by trainees		**# questions answered correctly**	
		# questions answered incorrectly	
		# questions redirected/delayed	

Trainer provided a break	Yes No	Trainer reviewed purpose of training	Yes No
Trainer asked if trainees had additional questions	Yes No	Trainer identified next steps	Yes No
Trainer prompted trainees to create action plans	Yes No	Total # of trainee individual action plans created:	

Implementation Fidelity Score	
(a) Total correct steps:	
(b) Total incorrect steps:	
Total a/(a+b)*100:	

Implementation fidelity form used to measure trainer behaviors during the initial training of indigenous implementers. **Note:** *This didactic training used a combination of a behavior skills training approach* (Miltenberger, 2012) *and practice based coaching* (Snyder, Hemmeter, & Fox, 2015).

Appendix 6.2

Board Game Study Procedural Fidelity

Date:	Session:	Peer:
Game:	Implementer:	PF:

Greet students and provide directive	Yes	No	NA	*Model a turn (game play priming)*	Yes	No	NA
Game set up correctly with all materials	Yes	No	NA	*End session after game over / timer goes off*	Yes	No	NA
Start timer after saying "Let's play!"	Yes	No	NA	*Session between 5 and 15 minutes*	Yes	No	NA
Review game-specific rules	Yes	No	NA	*Thank children for playing and return them to class*	Yes	No	NA
Review 4 steps on visual schedule	Yes	No	NA				

	SLP				Teacher Behaviors			
Turn	**Step 1***	**Step 2***	**Step 3***	**Step 4***	**Turn Praise[1]**	**Prosocial Praise[2]**	***Narration*[3]**	**Edible Reinforcement[4]**
1								
2								
3								
4								
5								
6								
7								
8								
9								
10								
11								

*mark **C** (correct) or **IN** (incorrect) for implementation of SLP for each turn for target child (mark correct if child completed independently and no SLP was required)

[1]mark **C** (correct) or **IN** (incorrect) for delivering praise *1-2 times* during or after completion of each turn **for both children**

[2] mark **C** (correct), **IN** (incorrect), or **NA** (no opportunity) for delivering praise following a peer-directed prosocial behavior (correct if praise provided between 1-7s following turn/behavior) within that turn

[3]mark **C** (correct), **IN** (incorrect), or **NA** (insufficient opportunities) for narrating game play behaviors *1-5 times* during each minute (**for both children combined**)

[4]tally – each should occur 1x/2-3 minutes to count as **C** during scoring

NOTE: 1 turn is equivalent to each child taking his/her turn

SCORING

Pre/Post-Session: (Y/[Y + N]) x 100
SLP—score by condition: C/([C + IN] x 100)

Baseline/Generalization/Follow-up:

- Pre/Post-Session: **NA** for reviewing game-specific rules, visual schedule, and modeling a turn (priming)
- SLP is **C** if not provided during *any* step of *any* turn
- SLP is **IN** if provided during *any* step of *any* turn
- Turn praise is **IN** if provided during *any* turn

Intervention:

- SLP is **C** for a step if: (a) the child does not begin/inappropriately attempts the step, the implementer provides the initial prompt within 2–5s, and (b) if the child does not begin/inappropriately attempts the step following the first prompt, the implementer provides the controlling prompt within 2–5s
- SLP is **IN** for a step if: (a) the procedure is not implemented as above, or (b) the implementer implements the procedure when the child is independently completing the step

Turn Praise: C/([C + IN] x 100)
Prosocial Praise: C/([C + IN] x 100)
Edible Reinforcement: C/([C + IN] x 100)
Narration: C/([C + IN] x 100)
Overall: (total # corrects/[total C + IN]) x 100

Procedural Fidelity C/(C + IN) x 100				
	C	IN	Total	%
Pre-/Post-Session				
SLP Step 1				
SLP Step 2				
SLP Step 3				
SLP Step 4				
Turn Praise				
Prosocial Praise				
Edible Reinforcement				
Narration				
Overall				

Appendix 6.3

Procedural Fidelity (Expressive Task) 4s CTD

Student: _____ Instructor: _____ Date: ____/____/____

Session: _____ Start Time: _____ Stop Time: _____ Total Session Time: _____
Delay Interval: *4 s*

Condition/Phase: _____ Behavior: _____ Observer: _____

Directions: *While observing teacher, please record whether teacher emitted behavior during instructional procedure for each trial.*

Teacher greets student	Yes No
Appropriate materials present	Yes No
Teacher reviews expectations	Yes No

Key: (+) = occurrence; (-) = nonoccurrence

Trial	Stimulus	T presents stimulus	T gives attending cue	T ensures participant attends	T gives instructional cue	T waits 4 s	Student Responding			T consequates correctly	T waits intertrial interval
							Correct	Incorrect	No Response		
1											
2											
3											
4											
5											
6											
7											
8											
9											
10											
11											
12											
13											
14											
15											
Fidelity Percentage											

Comments:

Appendix 6.4

Procedural Fidelity Teaching Coaching

Procedural Fidelity Teaching Coaching

Data Collection Forms for Implementers-Procedural Fidelity for Coaching Sessions and Feedback

Date: _____ Session # ___ Session Type: ____ Adult Participant: _____ Child Participant: _____Observer: _____

Did coaching session occur same day as observation: **Yes No** Length of coaching session: before: _____ after: _____

How did coaching occur: **in person via phone n/a** Coaching form provided to participant on day of session: **Yes No**

Correct materials present: **Yes No**

Amount of implementation time		**Just coach**
Coach + Participant	**Just participant**	

Ask if modeling is requested: **Yes No**
Provide direction or modeling *when requested*:

YES (tally)	**NO (tally)**

Coach Delivers Praise at least once per minute (each box is a minute)									
YES	NO	YES	NO	YES	NO	YES	NO	YES	NO

Coach/Participant	Review goals	Review data	Ask if they have questions	Answer questions	Give @ least 2 positive statements	Record responses	Request teacher input-success	Record responses	Request teacher input-challenges	Record responses	Ask if they have questions	Answer questions	Give @ least 1-2 goals	Request teacher input	Identify goal(s) for next session	Record responses	Ask if they have any questions	Answer questions
NOTES:																		

References

Ayres, K. M., Lowery, A., Douglas, K., & Sievers, C. (2011). I can identify Saturn but I can't brush my teeth: What happens when curricular focus for students with severe disabilities shifts. *Education and Training in Autism and Developmental Disabilities*, *45*, 11–21.

Ayres, K. M., Lowery, A., Douglas, K., & Sievers, C. (2012). The question remains: What happens when curricular focus for students with severe disabilities shifts. Rejoinder to Courtade et al. (2012). *Education and Training in Autism and Developmental Disabilities*, *47*, 14–22.

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.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1987). Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *4*, 313–327.

Bailey, K. M., & Blair, K. S. (2015). Feasibility and potential efficacy of the family-centered prevent-teach- reinforce model with families of children with developmental disorders. *Research in Developmental Disabilities*, *47*, 218–233.

Barton, E. E. (2015). Teaching generalized pretend play and related behaviors to young children with disabilities. *Exceptional Children*, *81*, 489–506.

Barton, E. E., & Fettig, A. (2013). Parent-implemented interventions for young children with disabilities: A review of fidelity features. *Journal of Early Intervention*, *35*, 194–219.

Barton, E. E., Pokorski, E. A., Sweeney, E. M., & Velez, M. (2017). The use of the system of least prompts to teach board game play within small groups of young children. *Journal of Positive Behavior Interventions*.

Billingsley, F. F., White, O. R., & Munson, R. (1980). Procedural reliability: A rationale and an example. *Behavioral Assessment*, *2*, 229–241.

Carr, J. E., Austin, J. L., Britton, L. N., Kellum, K. K., & Bailey, J. S. (1999). An assessment of social validity trends in applied behavior analysis. *Behavioral Interventions*, *14*, 223–231.

Carter, E. W., & Pesko, M. J. (2008). Social validity of peer interaction intervention strategies in high school classrooms: Effectiveness, feasibility, and actual use. *Exceptionality*, *16*, 156–173.

Chung, M. Y., Snodgrass, M. R., Meadan, H., Akamoglu, Y., & Halle, J. W. (2016). Understanding communication intervention for young children with autism and their parents: Exploring measurement decisions and confirmation bias. *Journal of Developmental and Physical Disabilities*, *28*, 113–134.

Courtade, G., Spooner, F., Browder, D., & Jimenez, B. (2012). Seven reasons to promote standards-based instruction for students with severe disabilities: A reply to Ayres, Lowery, Douglas, & Sievers (2011). *Education and Training in Autism and*

Developmental Disabilities, *47*, 3–13.

Dunst, C. J., Trivette, C. M., & Raab, M. (2013). An implementation science framework for conceptualizing and operationalizing fidelity in early childhood intervention studies. *Journal of Early Intervention*, *35*, 85–101.

Farmer, R., Wolery, M., Gast, D. L., & Page, J. L. (1988). Individual staff training to increase the frequency of data collection in an integrated preschool program. *Education and Treatment of Children*, *11*, 127–142.

Fettig, A., & Barton, E. E. (2014). Parent implementation of function-based intervention to reduce children's challenging behavior: A literature review. *Topics in Early Childhood Special Education*, *34*, 49–61.

Fettig, A., Schultz, T. R., & Sreckovic, M. A. (2015). Effects of coaching on the implementation of functional assessment—based parent intervention in reducing challenging behaviors. *Journal of Positive Behavior Interventions*, *17*, 170–180.

Garfinkle, A. N., & Schwartz, I. S. (2002). Peer imitation: Increasing social interactions in children with autism and other developmental disabilities in inclusive preschool classrooms. *Topics in Early Childhood Special Education*, *22*, 26–39.

Hanley, G. P. (2010). Toward effective and preferred programming: A case of the objective measurement of social validity with the recipients of behavior-change programs. *Behavior Analysis in Practice*, *3*, 13–21.

Heal, N. A., & Hanley, G. P. (2011). Embedded prompting may function as embedded punishment: Detection of unexplained behavioral process within a typical preschool teaching strategy. *Journal of Applied Behavior Analysis*, *44*, 127–131.

Hochman, J. M., Carter, E. W., Bottema-Beutel, K., Harvey, M. N., & Gustafson, J. R. (2015). Efficacy of peer networks to increase social connections among high school students with and without autism spectrum disorder. *Exceptional Children*, *82*, 96–116. doi:10.1177/0014402915585482

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.

Houten, R. V. (1979). Social validation: The evolution of standards of competency for target behaviors. *Journal of Applied Behavior Analysis*, *12*, 581–591.

Hurley, J. J. (2012). Social validity assessment in social competence interventions for preschool children: A review. *Topics in Early Childhood Special Education*, *32*, 164–174.

Hurley, J. J., Wehby, J. H., & Feurer, I. D. (2010). The social validity assessment of social competence intervention behavior goals. *Topics in Early Childhood Special Education*, *30*, 112–124.

Kaiser, A. P., & Hemmeter, M. L. (2013). Treatment fidelity in early childhood special education research: Introduction to the special issue. *Journal of Early Intervention*, *35*, 79–84.

Kamps, D. M., Kravits, T., Lopez, A. G., Kemmerer, K., Potucek, J., & Garrison, L. (1998). What do peers think? Social validity of peer-mediated programs. *Education and*

Treatment of Children, *21*, 107–134.

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

Kennedy, C. H. (2002). The maintenance of behavior change as an indicator of social validity. *Behavior Modification*, *26*, 594–604.

Kennedy, C. H. (2005). *Single case designs for educational research*. Boston, MA: Allyn & Bacon.

Koutsavalis, M. A. (2011). *The effects of sung versus spoken word on the sign acquisition and generalization of preschool children*. Unpublished thesis. Vanderbilt University, Nashville, TN.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, *34*, 26–38.

Lane, K. L., Kalberg, J. R., Bruhn, A. L., Mahoney, M. E., & Driscoll, S. A. (2008). Primary prevention programs at the elementary level: Issues of treatment integrity, systematic screening, and reinforcement. *Education and Treatment of Children*, *31*, 465–494.

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., & Wolery, M. (2013a). Procedural fidelity: An analysis of measurement and reporting practices. *Journal of Early Intervention*, *35*, 173–193.

Ledford, J. R., & Wolery, M. (2013b). Peer modeling of academic and social behaviors during small-group direct instruction. *Exceptional Children*, *79*, 439–458.

Ledford, J. R., Zimmerman, K. N., Chazin, K. T., Patel, N. M., Morales, V. A., & Bennett, B. P. (2017). Coaching paraprofessionals to promote engagement and social interactions during small group activities. *Journal of Behavioral Education*, *26*, 410–432. doi: 10.1007/s10864-017-9273-8

Ludemann, A., Power, E., & Hoffmann, T. C. (2017). Investigating the adequacy of intervention descriptions in recent speech-language pathology literature: Is evidence from randomized trials useable? *American Journal of Speech-Language Pathology*, *26*, 443–455.

Martino, S., Ball, S., Nich, C., Frankforter, T. L., & Carroll, K. M. (2009). Correspondence of motivational enhancement treatment integrity ratings among therapists, supervisors, and observers. *Psychotherapy Research*, *19*, 181–193.

Meadan, H., Angell, M. E., Stoner, J. B., & Daczewitz, M. (2014). Parent-implemented social-pragmatic communication intervention: A pilot study. *Focus on Autism and Other Developmental Disabilities*, *29*, 95–110.

Meadan, H., Stoner, J. B., Angell, M. E., Daczewitz, M., Cheema, J., & Rugutt, J. K. (2014). Do you see a difference? Evaluating outcomes of a parent-implemented intervention. *Journal of Developmental and Physical Disabilities*, *26*, 415–430.

Miltenberger, R. G. (2012). Behavior skills training procedures. In R. G. Miltenberger (Ed.), *Behavior modification: Principles and procedures* (5th ed., pp. 217–235).

Belmont, CA: Wadsworth.

Moncher, F. J., & Prinz, R. J. (1991). Treatment fidelity in outcome studies. *Clinical Psychology Review*, *11*, 247–266.

Pennington, R. C., Stenhoff, D. M., Gibson, J., & Ballou, K. (2012). Using simultaneous prompting to teach computer-based story writing to a student with autism. *Education and Treatment of Children*, *35*, 389–406.

Perepletchikova, F., & Kazdin, A. E. (2005). Treatment integrity and therapeutic change: Issues and research recommendations. *Clinical Psychology: Science and Practice*, *12*, 365–383.

Rapoff, M. A. (2010). Editorial: Assessing and enhancing clinical significance/social validity of intervention research in pediatric psychology. *Journal of Pediatric Psychology*, *35*, 114–119.

Reichow, B., Barton, E. E., & Maggin, D. (2017). *Risk of bias assessment for single case designs*. Unpublished manuscript, Anita Zucker Center for Excellence in Early Childhood Studies, University of Florida, Gainesville, FL.

Schwartz, I. S., & Baer, D. M. (1991). Social validity assessments: Is current practice state of the art? *Journal of Applied Behavior Analysis*, *24*, 189–204.

Smith, E. A., & Van Houten, R. (1996). A comparison of the characteristics of self-stimulatory behaviors in "normal" children and children with developmental delays. *Research in Developmental Disabilities*, *17*, 253–268.

Snodgrass, M. R., Chung, M. Y., Meadan, H., & Halle, J. W. (2017). Social validity in single-case research: A systematic literature review of prevalence and application. *Under Review*.

Snyder, P. A., Hemmeter, M. L., & Fox, L. (2015). Supporting implementation of evidence-based practices through practice-based coaching. *Topics in Early Childhood Special Education*, *35*, 133–143.

State, T. M., & Kern, L. (2012). A comparison of video feedback and in vivo self-monitoring on the social skills of an adolescent with Asperger Syndrome. *Journal of Behavioral Education*, *21*,18–33.

Strain, P. S., Barton, E. E., & Dunlap, G. (2012). The utility of social validity. *Education and Treatment of Children*, *35*, 183–200.

Sutherland, K. S., McLeod, B. D., Conroy, M. A., & Cox, J. R. (2013). Measuring implementation of evidence- based programs targeting young children at risk for emotional/behavioral disorders: Conceptual issues and recommendations. *Journal of Early Intervention*, *35*, 129–149.

Wehby, J. H., Maggin, D. M., Partin, T. C. M., & Robertson, R. (2012). The impact of working alliance, social validity, and teacher burnout on implementation fidelity of the good behavior game. *School Mental Health*, *4*, 22–33.

Weng, P. L., & Bouck, E. C. (2014). Using video prompting via iPads to teach price comparison to adolescents with autism. *Research in Autism Spectrum Disorders*, *8*, 1405–1415.

Wolery, M. (2011). Intervention research: The importance of fidelity measurement.

Topics in Early Childhood Special Education, *31*, 155–157.

Wolery, M. (2013). A commentary: Single-case design technical document of the what works clearinghouse. *Remedial and Special Education*, *34*, 39–43.

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

Wood, B. K., Umbreit, J., Liaupsin, C. J., & Gresham, F. M. (2007). A treatment integrity analysis of function- based intervention. *Education and Treatment of Children*, *30*, 105–120.

Wood, B. K., Ferro, J. B., Umbreit, J., & Liaupsin, C. J. (2011). Addressing the challenging behavior of young children through systematic function-based intervention. *Topics in Early Childhood Special Education*, *30*, 221–232.

7
Visual Representation of Data

Amy D. Spriggs, Justin D. Lane, and David L. Gast

Important Terms

graphic display, abscissa, ordinate, origin, tic marks, axis labels, condition, phase, condition labels, figure caption, line graph, bar graph, cumulative graph, semi-logarithmic chart, scale break, blocking

Graphs should represent complex information without distortion, and should serve a clear purpose (Tufte, 2001). They should "induce the reader to think about the substance rather than about methodology, graphic design, the technology of graphic production, or something else" (Tufte, 2001, p. 1). Maximizing the impact of your data while minimizing consumer focus on "something else" can be done by following guidelines for graphing data that come from professional organizations (e.g., American Psychological Association [APA]), historical precedent, and empirical knowledge (i.e., research). In single case design (SCD) research, graphic displays are not only a way to share your outcomes with consumers of your research (as is also common in between-groups studies), but also to enable you to make formative decisions throughout the process of the study. Thus, well-designed graphics are essential in good SCD research.

Graphic displays (e.g., line graphs, bar graphs, cumulative graphs) and tables serve two basic purposes. First, they assist in organizing data during the data collection process, which facilitates formative evaluation. Second, they provide a detailed summary and description of behavior over time, which allows readers to analyze the relation between independent and dependent variables. The underlying purpose or function of the graphic display is communication. For the person collecting data, the graph is a vehicle for efficiently organizing and summarizing a participant's behavior over time. It

allows the researcher to analyze, point by point, the effect a particular event has on a participant's behavior. In SCD research, visual analysis is the primary method of data evaluation; thus, appropriate graphing is critical. In addition to reliance on graphically displayed data for communication and analysis, practitioners may find graphing economical in terms of time saved by not having to review daily data forms prior to making program decisions and by not maintaining ineffective intervention programs.

Independent analysis of relations between variables is one of many strengths characteristic of SCD research. By reporting all data, readers can determine for themselves whether a particular intervention has a reliable and "significant" effect on a participant's behavior. The graph, as a compact and detailed data-reporting format, permits independent analysis by not only the researchers, but also by consumers. Although data could be reported in written narrative, such a format would be cumbersome and difficult to reliably analyze.

Graphic representation of data provides researchers and consumers with an efficient, compact, and detailed summary of participant performance. A well-constructed graph communicates to readers (a) sequence of experimental conditions and phases, (b) time spent in each condition, (c) independent and dependent variables, (d) experimental design, and (e) relations between variables. Therefore, it is not surprising that applied researchers rely heavily on graphic displays.

Graphic Displays of Data

Four basic principles help graphs communicate information to readers: clarity, simplicity, explicitness, and good design (Parsonson & Baer, 1978). Moreover, graphs should allow the consumer to get "the greatest number of ideas in the shortest time with the least ink in the smallest space" (Tufte, 2001). A well-constructed graph will (a) use easily discriminable data points and data paths, (b) clearly separate experimental conditions, (c) avoid clutter by keeping the number of behaviors plotted on one graph to a minimum, (d) provide brief descriptive labels, and (e) use appropriate proportions. In addition, it is your responsibility to select an appropriate graphic display for presenting data. The type of display will depend upon type of data collected and intended communication. Generally, SCD researchers present all data (e.g., baseline, intervention, probe, and review data) via one or more types of graphs. By presenting all data, you enable readers to independently analyze data patterns.

Applied researchers use three basic types of graphic displays: (a) line graphs, (b) bar graphs, and (c) cumulative graphs (also called cumulative records). Although this chapter discusses only the simplest figures within each of the three categories, you should be aware that there are numerous variations.

Before discussing each type of graph, you should be familiar with basic components and symbols used in graphic representations. Figure 7.1 presents major components of a simple line graph and simple bar graph. As shown, there are several common components across the two types of figures. These include:

- ***Abscissa***: Horizontal line (x axis) that typically identifies the time variable (e.g., sessions, days, dates). Typically, SCD data are presented ordinally on the abscissa (e.g., Session 1 comes before Session 2 in time, which comes before Session 3; but the time between Sessions 1 and 2 and Sessions 2 and 3 is not necessarily the same)
- ***Ordinate***: Vertical line (y axis) that typically identifies the dependent variable (e.g., percentage, number, duration, responses per minute)
- ***Origin***: Common point of intersection of the abscissa and ordinate
- ***Tic Marks***: Points along both the abscissa and ordinate representing values (e.g., 0%, 10%, 20%; Sessions 1, 2, 3)
- ***Axis Labels***: Numeric value corresponding to a tic mark.
- ***Condition***: Procedurally similar sessions (e.g., Baseline, Intervention). Conditions should be separated on graphs with solid lines (condition change lines)
- ***Phase***: Within-condition variations (e.g., procedural modifications within an intervention condition). Phases should be separated on graphs with dotted lines (phase change lines). Refer to Figure 7.1 for an example of a phase change

within an experimental condition (i.e., moving from a CRF to a VR-3 schedule of reinforcement)

- ***Condition Labels***: One or two descriptive words or common abbreviations that identify each experimental condition (e.g., Baseline, Social Reinforcement)
- Figure Number and ***Figure Caption***: The figure number is used in the narrative to direct a reader's attention to the appropriate graph, and the figure caption provides a brief and explicit description of dependent and independent variables and any other relevant information, including defining any abbreviations used in the figure

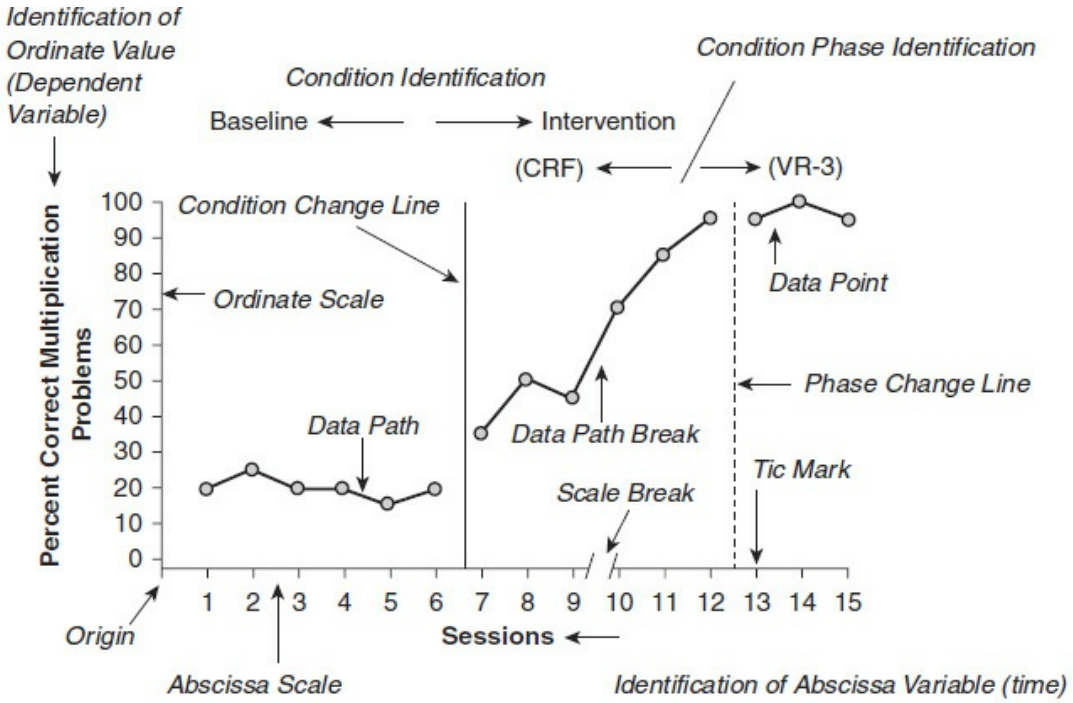


Figure Number and Caption →

Figure 5. Percentage of multiplication problems completed correctly by Jen during baseling baseline and intervention (CRF and VR-3) conditions.

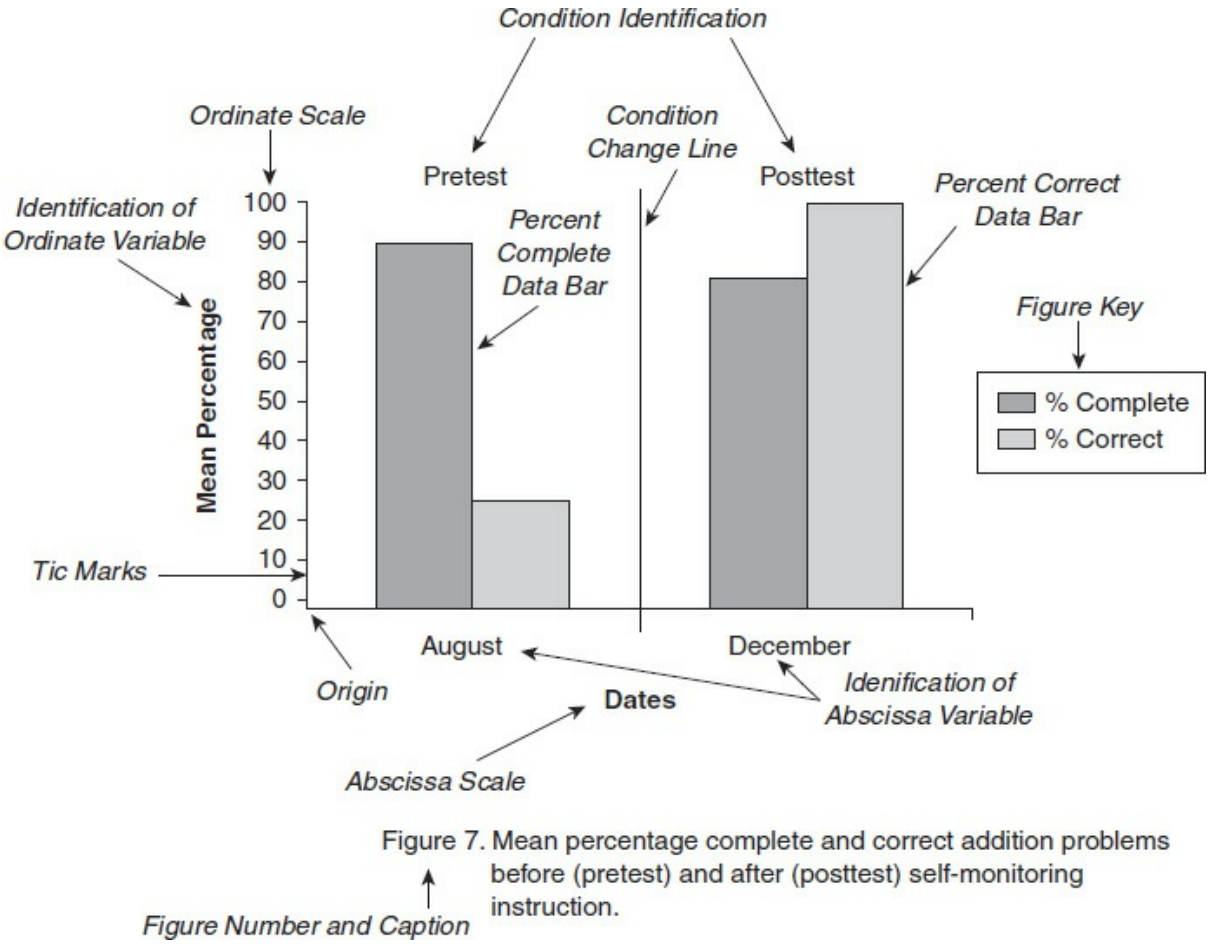


Figure 7.1 The basic components (italicized) of a simple line graph and simple bar graph.

Line Graphs

Line graphs represent the most commonly used graphic display, both in SCD research and more broadly (Tufte, 2001); they represent data over time. Figure 7.2 shows a simple line graph on which percentage of trials with compliance is measured at varying integrity levels (100%, 50%, and 0%) for one participant. In the interest of simplicity and clarity, seldom are more than three data paths plotted on a single graph. If several secondary behaviors are being monitored, as in the case of monitoring the effect of an intervention on non-target behaviors (i.e., response generalization), additional graphs can be used. Figure 7.3 shows one way to present data for several non-target behaviors that are being monitored concurrently.

The line graph has several advantages, the most important of which being that it is familiar to most readers and, thus, is easily read and understood. In addition, it is easy to construct and permits the researcher or practitioner to continuously evaluate the effect an intervention has on dependent variable(s), thus facilitating formative evaluation and the decision to maintain or modify the intervention. Line graphs should generally be used to display primary variables that are measured over time in the context of an SCD.

Bar Graphs

Applied researches have traditionally used **bar graphs** to display discrete data and comparative information; the height of the bar indicates the magnitude of the data. The great versatility of a bar graph is indicated by its numerous variations. Generally, bar graphs should be used to present summative level data rather than changes over time (which should be presented via line graph; for a review of line graph versus bar graph interpretation, see Shah & Hoeffner, 2002). Bar graphs are also helpful for displaying categorical data related to narrative reviews of SCD data (e.g., the number of participants of given age ranges included in reviewed studies). As shown in Figure 7.4, a bar graph can be used to summarize a student's performance or behavior on a pre- and post-test measure of a generalized tendency, before and after introducing an intervention for improving social interactions among peers, within the context of an SCD.

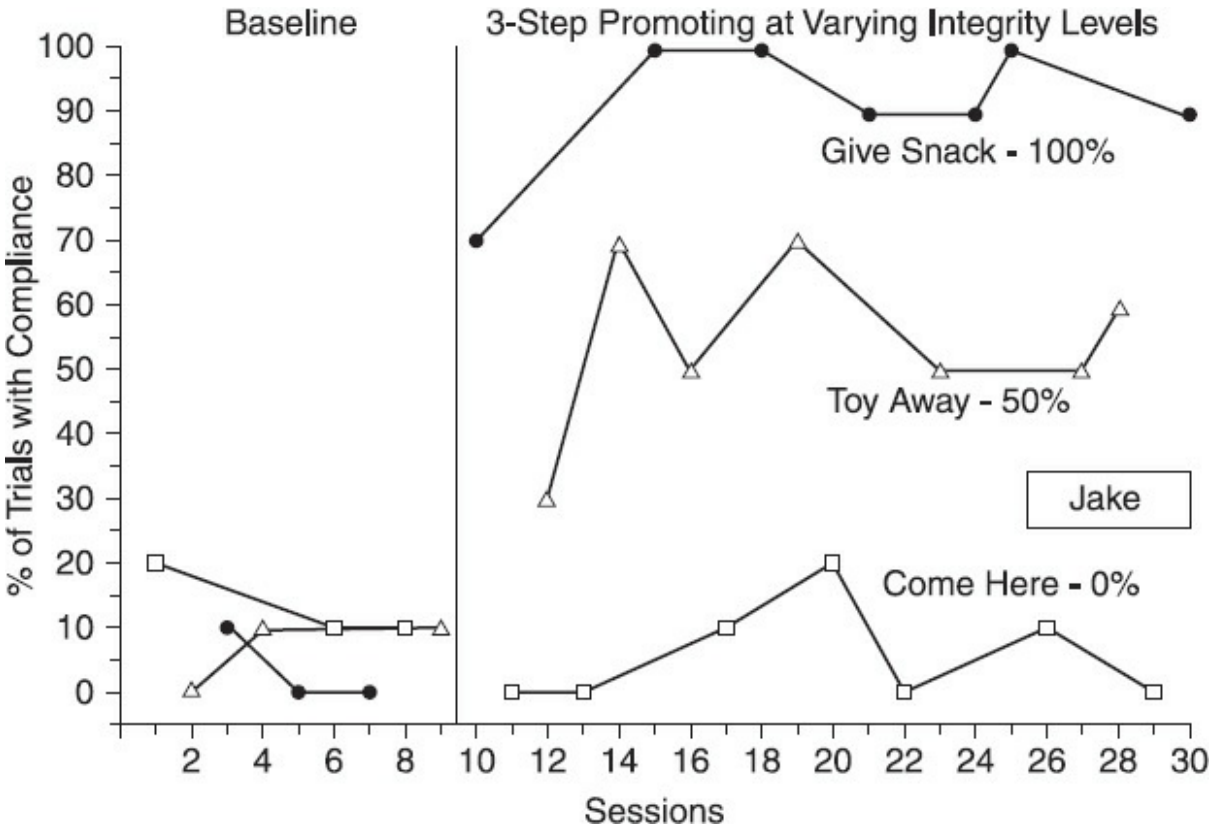


Figure 7.2 Graph showing one simple line graph on which the dependent variable (% compliance) is measured at varying integrity levels across one student.

Source: Wilder, D. A., Atwell, J., &Wine, B. (2006). The effects of varying levels of treatment integrity on child compliance during treatment with a three-step prompting procedure. *Journal of Applied Behavior Analysis, 39*, 369–373.

A variation of a simple bar graph is the subdivided bar graph. Figure 7.5 uses the subdivided bar graph format to indicate the mean percentages of sedentary, light, moderate, and vigorous physical activity of a child with autism spectrum disorder on the playground. This method of plotting summarizes the magnitude of target behavior with

a single bar, which permits a quick and easy comparison of data, in addition to conserving space.

Bar graphs provide a simple and straightforward summary of data that are easily understood and analyzed. Though they are not recommended for displaying continuous data, they represent an excellent format for displaying and communicating important comparisons in a final research report or literature review. In their construction, it is important to remember to keep the width of each bar identical and thus not perceptually mislead readers. In addition to simplicity and clarity indicative of a well-designed bar graph, it is easy to construct. Bar graphs may prove useful when communicating a summary of progress to stakeholders (e.g., participants, clients, parents).

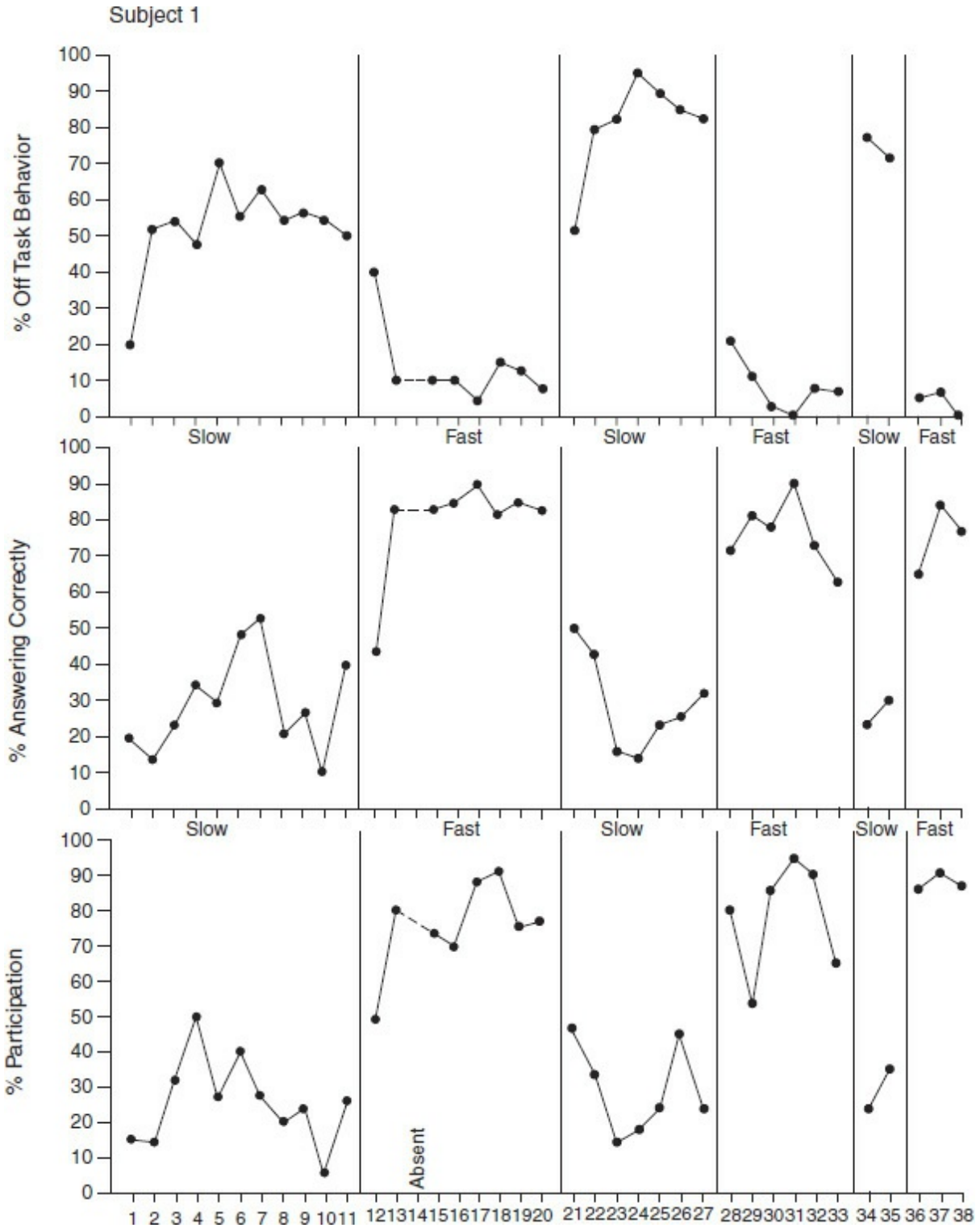


Figure 7.3 Simple line graph showing three behaviors being monitored concurrently.

Source: Carnine, D. W. (1976). Effects of two teacher presentation rates on off-task behavior, answering correctly, and participation. *Journal of Applied Behavior Analysis*, *9*, 199–206.

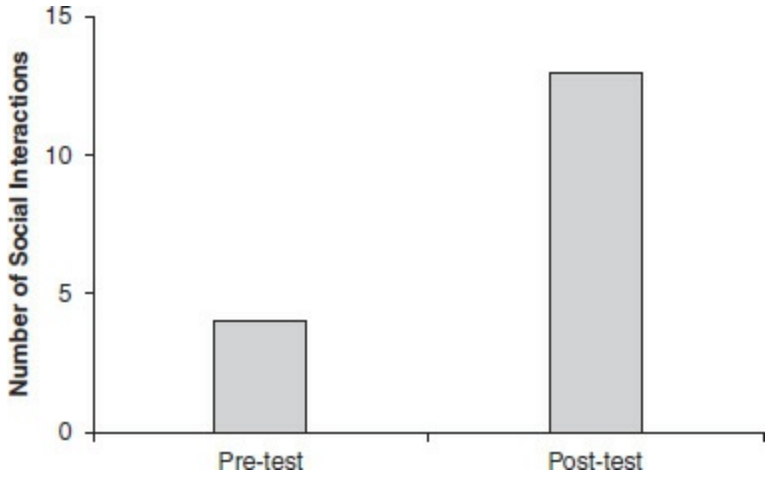


Figure 7.4 Single bar graph showing a student's behavior (number of social interactions) during a pre- and post-test of generalized tendencies during a study on promoting peer-to-peer interactions at school.

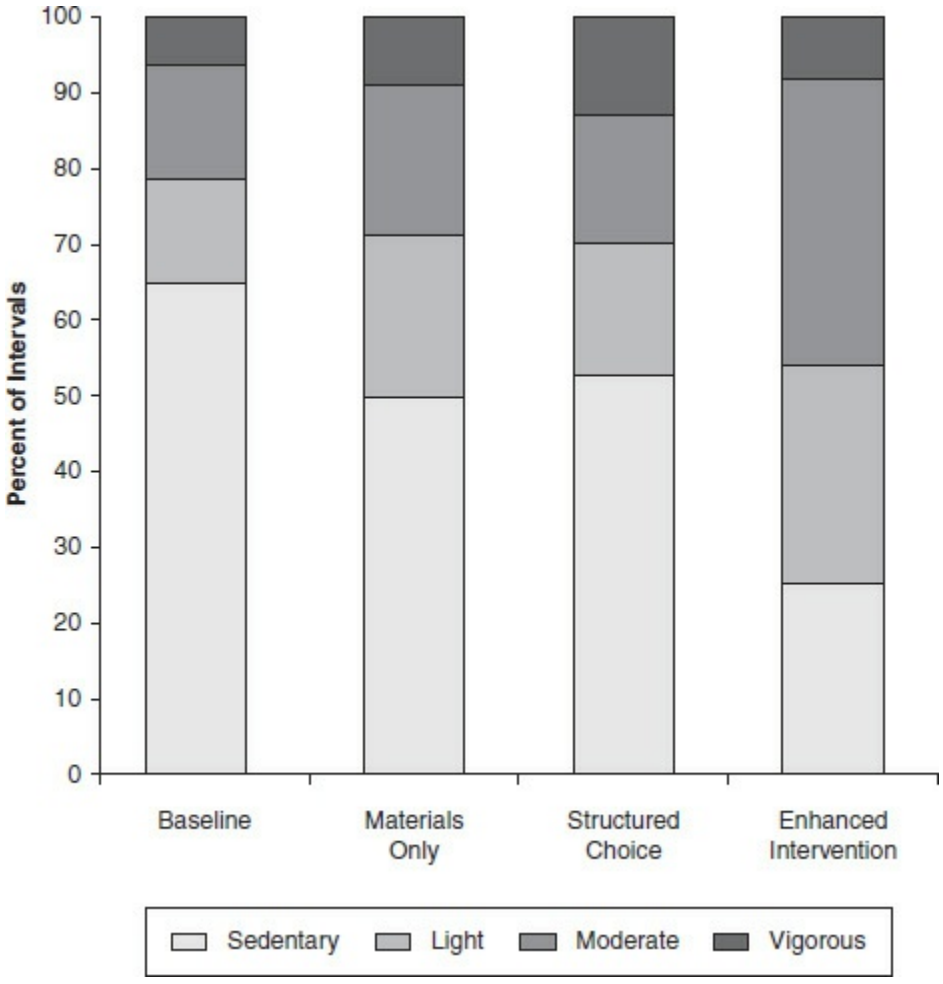


Figure 7.5 Subdivided bar graph used to show time spent in various levels of physical activity in a student with autism spectrum disorder.

Source: Ledford, J. R., Lane, J. D., Shepley, C., & Kroll, S. M. (2016). Using teacher-implemented playground interventions to increase engagement, social behaviors, and physical activity for young children with autism. *Focus on Autism and Other Developmental Disabilities*, *31*, 163–173.

Cumulative Graphs

Cumulative graphs (also called cumulative records) have been used less frequently by applied researchers than line graphs and bar graphs. Few examples are found in the empirical applied research literature, although several research teams have used cumulative graphs to display data regarding participant preference (cf. Heal & Hanley, 2007; Ledford, Chazin, Harbin, & Ward, 2017). Cumulative graphs provide an excellent visual summary of participant progress toward goal mastery or choice for condition (see discussion regarding simultaneous treatment designs in Chapter 12). "When cumulative records are plotted … the number of responses recorded during each observation period is added (thus the term *cumulative*) to the total number of responses recorded during all previous observation periods. In a cumulative record the y axis value of any data point represents the total number of responses recorded since the beginning of data collection" (Cooper, Heron, & Heward, 2007, pp. 135–136). Thus, a flat line in a cumulative record indicates no responding occurring across sessions and an increasing data path represents some level of responding. Because responses are cumulative, or additive, it is not possible to have decreases over time on a cumulative graph.

When plotting data on a cumulative graph and *a priori* criteria are met, the cumulative number returns to zero (Figure 7.6 shows a cumulative graph where criteria have been reached). The number of correct responses for the second step begins accumulating from zero. This is also the case if the cumulative number reaches the upper limit on the y-axis (Cooper et al., 2007). Plotting data in this manner allows readers to see, as shown in Figure 7.6, that after Step 1, the participant met criteria for subsequent steps in much fewer trials. In the case that a participant's cumulative graph is reset to zero due to reaching limits on the y axis, the numbers would simply be added together to figure cumulative number, rate, etc. Figure 7.7 illustrates using a cumulative graph within an A-B-A-B withdrawal design. In this example, eye-goggles nearly stopped eye-poking while used. Using a cumulative graph in this manner allows a clear demonstration of intervention effect on the behavior.

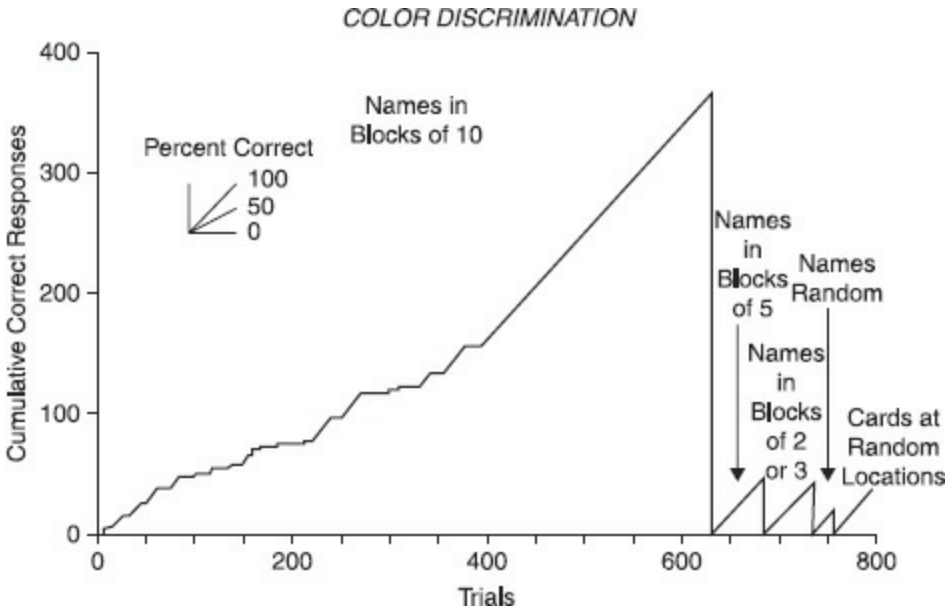


Figure 7.6 Cumulative graph showing how to continue recording data after criteria are reached for a step.

Source: Williams, G., Perez-Gonzalez, L. A., & Queiroz, A. B. (2005). Using a combined blocking procedure to teach color discrimination to a child with autism. *Journal of Applied Behavior Analysis*, *38*, 555–558.

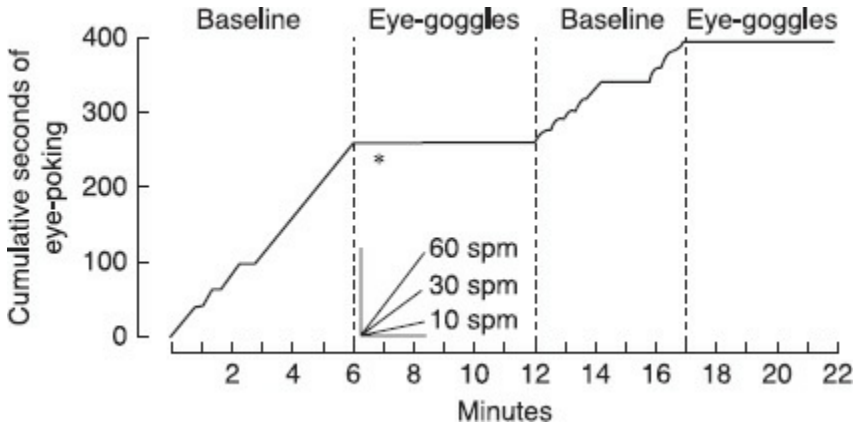


Figure 7.7 Cumulative graph used within an ABAB withdrawal design.

Source: Kennedy, C. H., & Souza, G. (1995). Functional analysis and treatment of eye poking. *Journal of Applied Behavior Analysis*, *28*, 27–37.

According to Cooper et al., (2007), cumulative graphs should be chosen over simple line graphs or bar graphs when: (a) total number is important for reaching a specific goal, (b) giving feedback to participants, (c) opportunities to respond are consistent, and (d) behavior change patterns would be more accurately reflected by using a cumulative graph. Cumulative records are often used for secondary measures when a participant has one opportunity to respond during each measurement occasion or session (e.g., when a participant can choose which intervention condition is in effect for the day; Ledford et al., 2017).

Semi-logarithmic Charts

Semi-logarithmic charts are used when absolute changes in behavior (which are what we have discussed to this point) are not the focus of research. Absolute behavior changes are documented using equal-interval recording, where amounts are "equal" between tic marks on the graph. In contrast, relative behavior changes can be captured when the distance between tic marks are proportionally equal. "For example, a doubling of response rate from 4 to 8 per minute would appear on a semi-logarithmic chart as the same amount of change as a doubling of 50 to 100 responses per min" (Cooper et al., 2007, p. 139). Figure 7.8 uses a semi-logarithmic chart to graphically display a student's performance (count per minute) on a discrimination task for purposes of exemplifying behavioral methods of instruction and Precision Teaching as complementary practices for students with autism spectrum disorder. Lindsley (1992) provides a review of Precision Teaching, as well as use of semi-logarithmic charts for demonstration of changes in behavioral programs, for readers interested in a brief history of Precision Teaching and detailed descriptions of each component of a semi-logarithmic chart. Graphing responses on semi-log charts is similar to transforming outcome data using natural logs, a common practice in between-groups research.

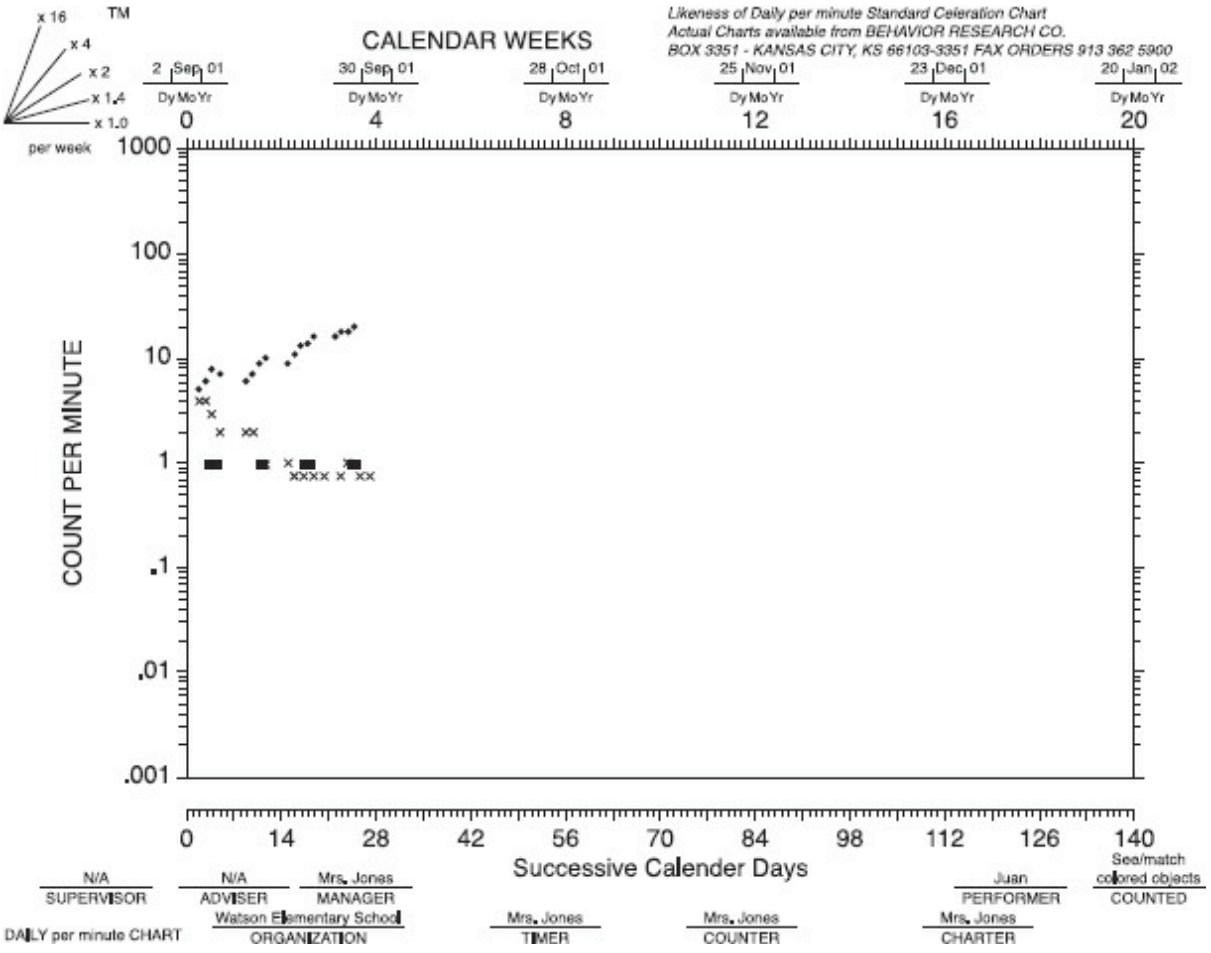


Figure 7.8 Semi-logarithmic chart to demonstrate learning progress.

Guidelines for Selecting and Constructing Graphic Displays

Before analyzing graphically displayed data, it is important to evaluate the appropriateness of the format to display your data. The primary function of a graph is to communicate without assistance from the accompanying text. This requires that you (a) select the appropriate graphic display (line graph, bar graph, or cumulative graph) and (b) present the data as clearly, completely, and concisely as possible. How data are presented and how figures are constructed directly influences a reader's ability to evaluate functional relations between independent and dependent variables. Though there are few hard and fast rules that govern figure selection, graph construction, or data presentation, there are recommended guidelines for preparing graphic displays (APA, 2009; Parsonson & Baer, 1978; Sanders, 1978). Following these guidelines should facilitate objective evaluations of graphically displayed data.

Figure Selection

When plotting time series data, you should generally use a line graph, and when plotting summative data, you should generally use a bar graph. Cumulative records are helpful when sessions represent a single opportunity to respond, or when reaching a cumulative number is critical (often true in experimental analyses with non-human subjects). Combination bar and line graphs are sometimes used when two or more variables are measured to simplify display (cf. Shepley, Spriggs, Samudre, & Elliot, 2017), even if all data are collected over time (e.g., when we would generally recommend using a line graph). For example, if two data paths are likely to have similar values throughout the study, a researcher might decide to present one as a bar graph and another as a superimposed line graph. Although this goes against advice above regarding representing time series data in bar graphs, in some situations, it can improve accessibility and decrease confusion. As previously mentioned, avoiding clutter by keeping the number of behaviors plotted on one graph to a minimum is a key component to a well constructed graph; with more than three data paths on a single graph, "the benefits of making additional comparisons may be outweighed by the distraction of too much visual 'noise' " (Cooper et al., 2007, p. 132).

Graph Construction

The historically preferred proportion of ordinate (y axis) to abscissa (x axis) has been reported to be a ratio of 2:3, 3:5, or 3:4 (Kubina, Kostewicz, Brennan, & King, 2017). This has been viewed by researchers as limiting the degree of perceptual distortion. The same

data are graphed in Figure 7.9 using the 2:3 ratio (Figure 7.9a) and larger and smaller ratios (Figure 7.9b and c). The 2:3 convention may be appropriate when there are relatively few data points on the graph. When there are a large number of data points, using a 1:3 ratio (Figure 7.10a) may be more appropriate. The data in Figure 7.10 are also distorted using larger and smaller ratios (Figure 7.10b and c). It is clear that data appear drastically different based on the ratio of height to width; however, it is unclear in what situations the historically-suggested ratios are appropriate. For example, studies suggest that the density of data (e.g., the number of data points *per cm* on the x-axis) impacts data analysis decisions (Shah & Hoeffner, 2002). This outcome is not specific to SCD design graphs; additional research is needed to guide the construction of graphs with time series data. A recent review suggests the average ratio for most design types is approximately 4/10 and expert SCD researchers prefer a ratio of 0.25 as compared to larger (0.55, 0.65, 0.75) ratios (Ledford, Barton, Severini, Zimmerman, & Pokorski, 2017). Our suggestions are to:

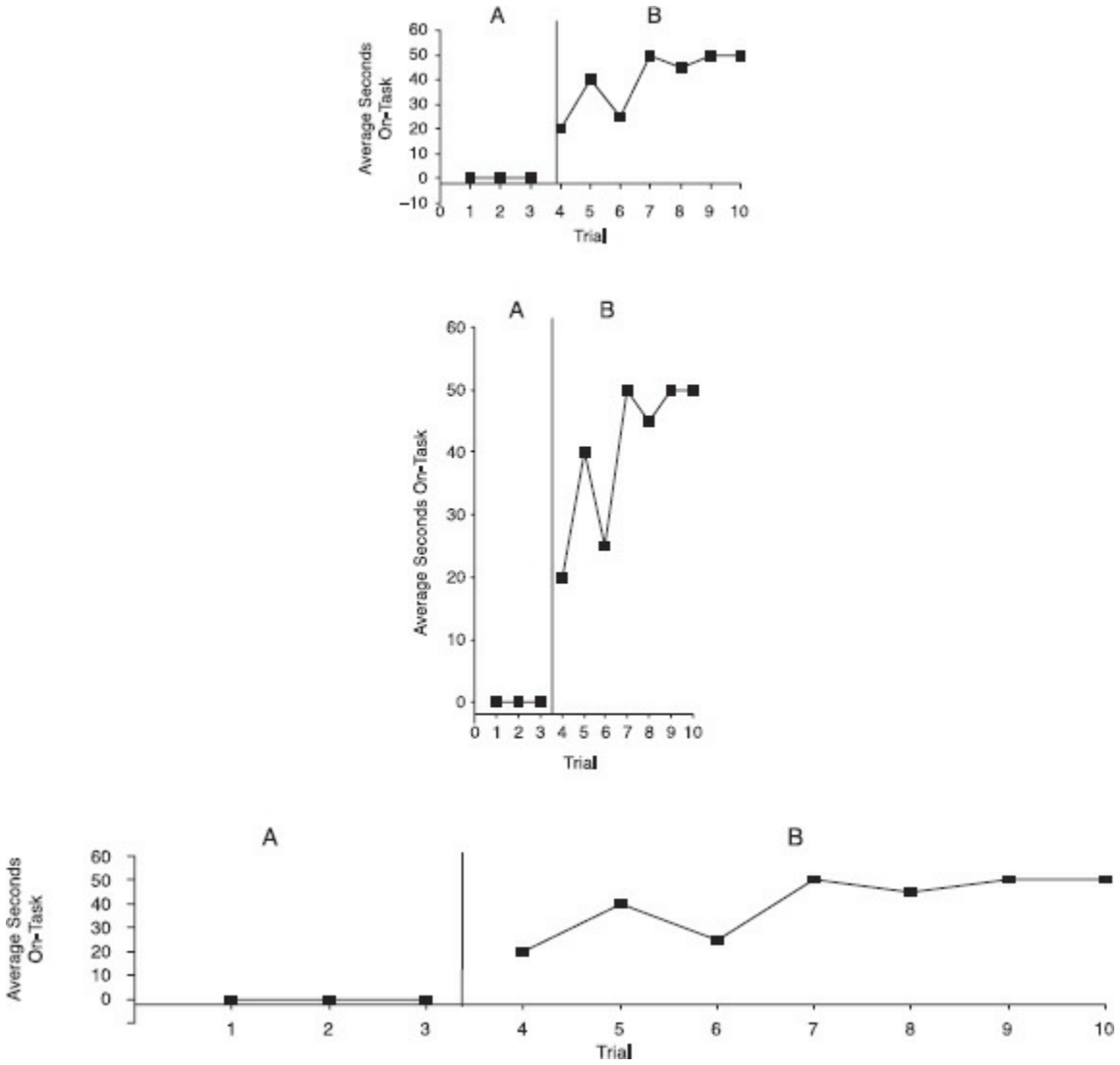
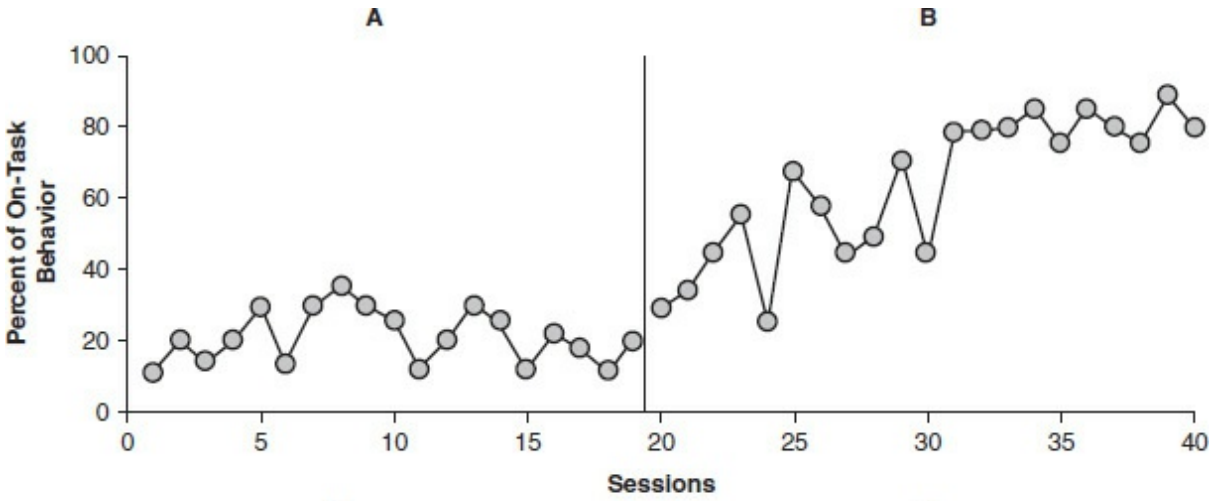
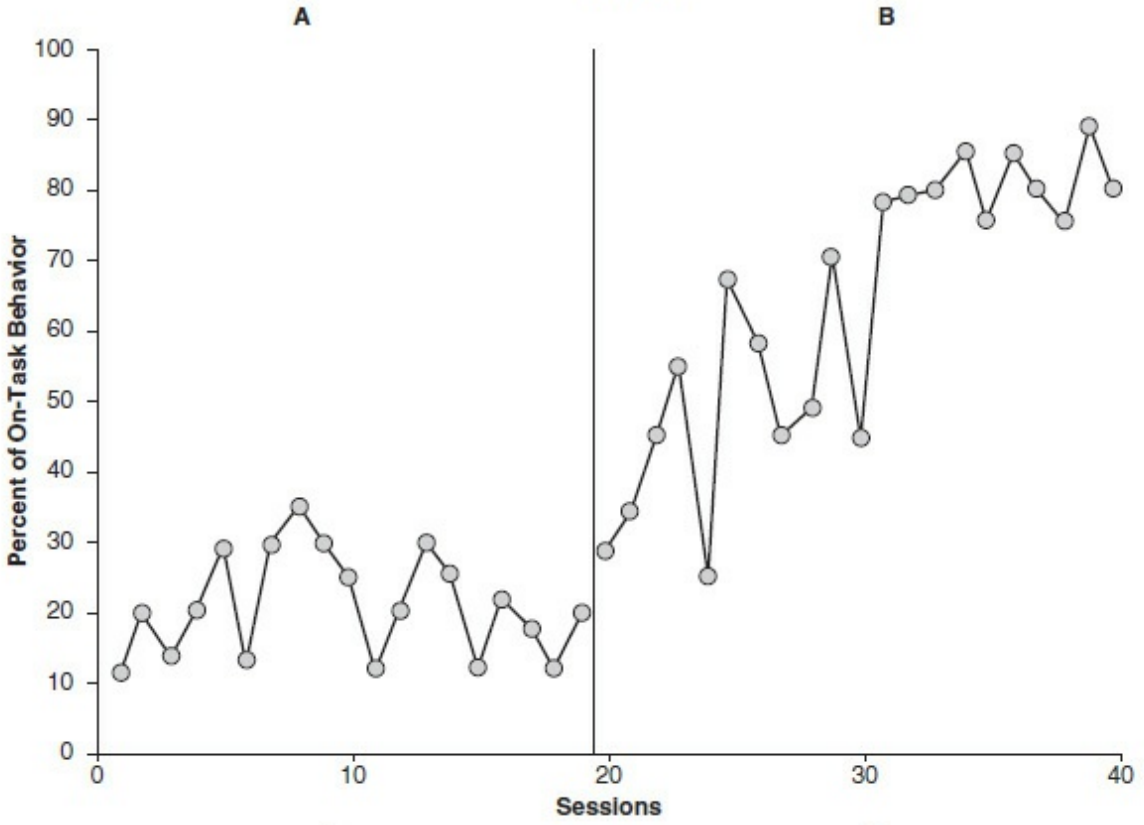


Figure 7.9 Various graphing proportions when there are relatively few data points: On the above graphs, identical data are displayed using various abscissa/ordinate ratios. Graph (a) illustrates a 2:3 ratio, graph, (b) uses a 3:2 ratio

which creates a steeper slope, exaggerating the change along the data path, and (c) uses a ratio of approximately 1:5 which creates a more shallow data path, reducing appearance of variability of data and change over time.





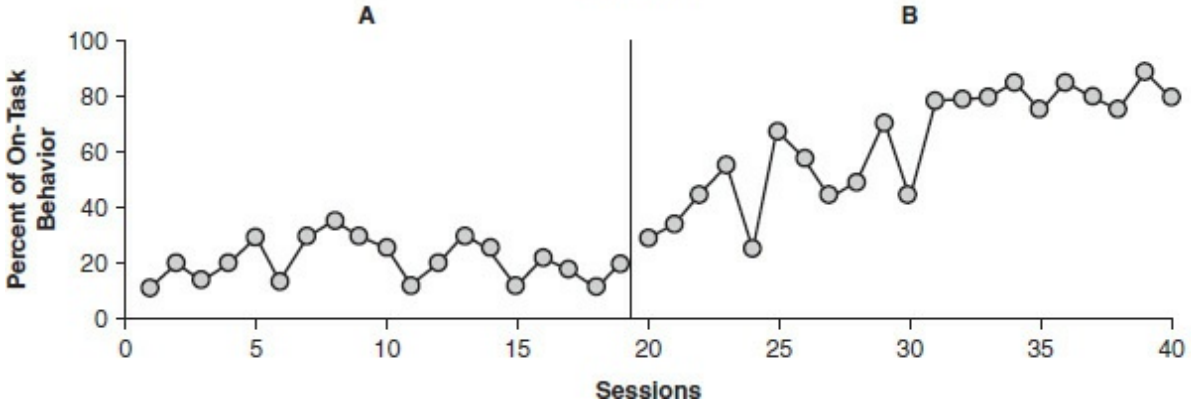


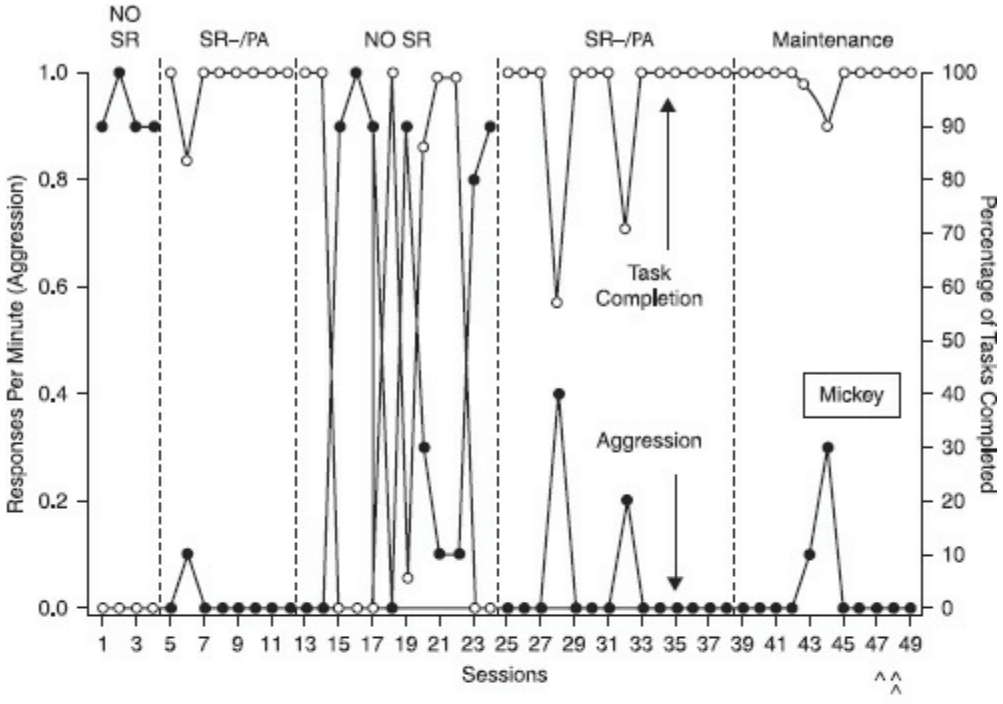
Figure 7.10 Various graphing proportions when there are a large number of data points: On the above graphs, identical data are displayed using various abscissa/ordinate ratios. Graph (a) illustrates a 1:3 ratio graph, (b) uses a 3:2 ratio which creates a steeper slope, exaggerating the change along the data path, and (c) uses a ratio of approximately 1:5 which creates a more shallow data path, reducing appearance of variability of data and change over time.

1. Use a ratio that does not distort data and allows for discrimination between data points (e.g., 2/3 ratio for graphs with relatively few data points, 1/3 for graphs with a large number of data points).
2. Use a font consistent with the font used in your narrative text (usually Times New Roman) for all text on the graph, including figure labels, condition labels, and axis labels.
3. Ensure that numbers presented as x-axis labels are easy to read and that tic marks between axis labels are used to assist the reader in identifying midpoints. For example, if you label every other session (e.g., 2, 4, 6) you should put a tic mark at each session; if you label every 10th session (e.g., 10, 20, 30) on a graph with many sessions, use tic marks at every 5th session.
4. Ensure that numbers presented as y-axis labels are easy to read and separated in space. For example, use data labels 0, 20, 40… for graphs with a percentage dependent variable (e.g., graphs with a maximum y-value of 100).
5. If multiple data paths appear on the same graph, use different marker shapes (e.g., triangles, circles, squares). Use filled (black) markers for one data path and unfilled (white) markers for the second. If a third data path is used, you can use gray fill, but ensure that the markers are big enough for these to be discriminated from black-filled markers.
6. Use thin lines for data paths so as not to obscure marker position (e.g., 1.0 point lines in Microsoft Excel) and markers that are large enough to be differentiated from each other but small enough so that readers can accurately detect the y-value.
7. Use the same ordinate size and maximum y-value on all graphs reporting the same measurement units in the same research report (Kennedy, 1989).
8. Label data paths using text boxes and arrows (see Figure 7.6).
9. Do *not* use color, gridlines, keys, or titles.

A **scale break** is sometimes used when the entire abscissa or ordinate scale is not presented. The abscissa scale should be divided into equal interval sessions, days, time, etc. When data are not collected continuously, a scale break should be inserted on the abscissa between the two non-consecutive data points (See Figure 7.1 and Mayfield & Vollmer, 2007 for examples). Although some articles are published showing a scale break on the ordinate (cf. Maglieri, DeLeon, Rodriguez-Catter, & Sevin, 2000), we caution researchers against it as it can inadvertently distort data (Dart & Radley, 2017).

The zero origin tic mark along the ordinate ideally should be placed slightly above the abscissa when any data point value is zero (referred to as "floating" the zero). When

constructing line or bar graphs, it is particularly important not to mistake a zero level for the absence of plotted data. If there are no zero level data points to be plotted on a line graph, the zero origin tic mark need not be raised above the abscissa. Barton and Reichow (2012) have described procedures for floating the zero and other procedures in common graphing programs and Vanselow and Bourret (2012) have developed online video tutorials for the same purposes. We will note here that in common Microsoft programs, it is preferable to use what are identified as scatterplots (marked scatter) rather than line graphs due to the relative flexibility (e.g., alignment with data values, ability to add a precise condition change line; Vanselow & Bourret, 2012).



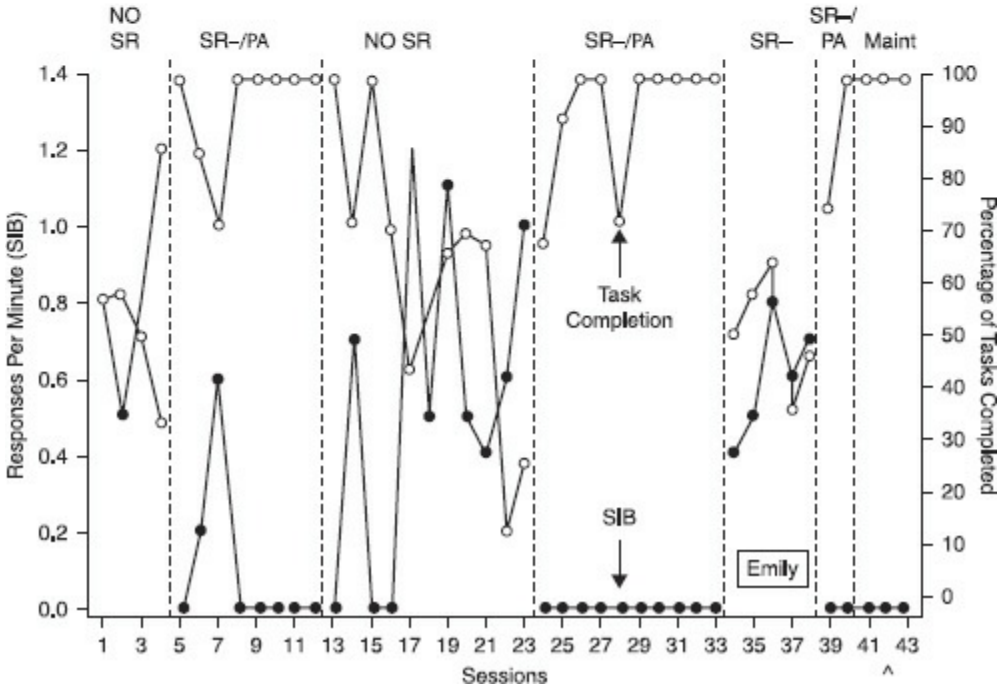


Figure 7.11 Figure captions, experimental condition labels, and data path labels should be concise but explanatory. They should provide sufficient information to allow readers to identify dependent and independent variables as well as experimental design. Any abbreviations used for graph labels should be explained in the figure caption (e.g., BL=baseline).

Source: Hoch, H., McComas, J. J., Thompson, A. L., & Paone, D. (2002). Concurrent reinforcement schedules: Behavior change and maintenance without extinction. *Journal of Applied Behavior Analysis*, *35*, 155–169.

The dependent measure should be clearly and concisely labeled along the ordinate. Most often, a single dependent measure is labeled along the left ordinate. When more than one dependent measure is graphed, the right ordinate may also be used. Figure 7.11 exemplifies using the left and right ordinate to graph two different measures on the same graphic display; rate is shown on the left ordinate and percentage is shown on the right ordinate. Abbreviations and symbols (e.g., %, #) are discouraged in favor of descriptive labels. The frequency with which data are collected (e.g., sessions, days, weeks) should be noted along the abscissa on line graphs.

Data Presentation

A data path using a solid line to connect two points implies that there is continuity in the data collection process. Dashed or omitted data path lines are used to identify discontinuous data. Dashed lines have, on occasion, been used to connect two points between which no data have been collected (such as connecting data points when the participant has been absent). It is inappropriate to connect data points of two different experimental conditions or condition phases (i.e., data paths should not cross experimental condition and phase lines).

When graphing similar behaviors on multiple graphs, it is important to maintain ordinate size consistency. Figure 7.12 illustrates how effects can be distorted when ordinate sizes are not consistent (Dart & Radley, 2017; Kennedy, 1989).

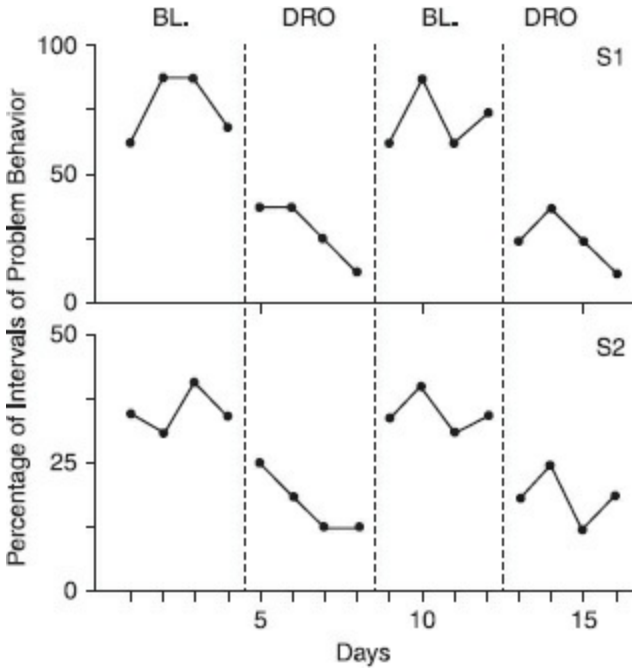


Figure 7.12a Graphs showing the effects of using inconsistent ordinate scales vs. consistent ordinate scales.

Source: Kennedy, C. H. (1989). Selecting consistent vertical axis scales. *Journal of Applied Behavior Analysis, 22,* 338–339.

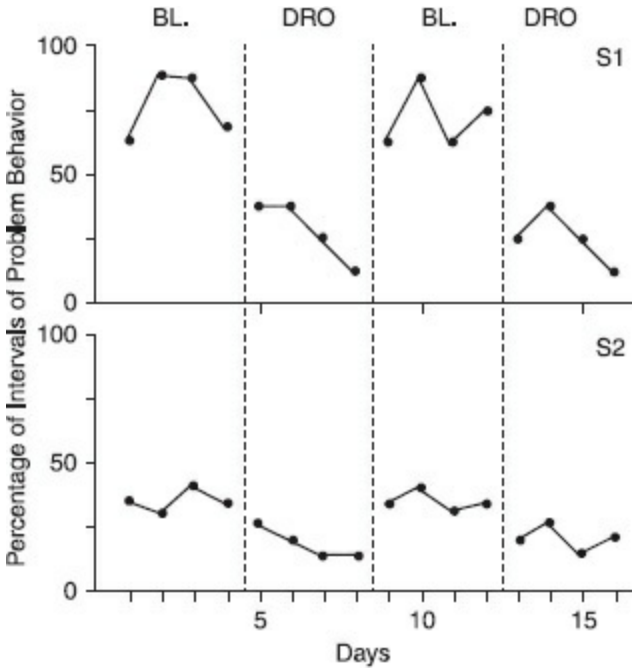


Figure 7.12b Graphs showing the effects of using inconsistent ordinate scales vs. consistent ordinate scales.

Source: Kennedy, C. H. (1989). Selecting consistent vertical axis scales. *Journal of Applied Behavior Analysis*, *22*, 338–339.

When logistically feasible, SCD researchers present all data. On occasion, however, when data have been collected over an extended period of time, it may be necessary to condense data in order to present it on a single graph. A procedure for condensing data, commonly referred to as "**blocking**," is infrequently used to reduce the number of data points plotted on a graph. This procedure entails calculating mean or median performance level of two or more adjacent days' data, thereby reducing the length of the abscissa and the number of data points presented on the graph. When blocking is used, proceed with caution. It is appropriate to block data only if blocking does not mask the variability of the data. The procedure is dangerous in that it is possible for a researcher to distort the actual data trends, and therefore it is rarely used. When data points are blocked, you should (a) note that the data have been blocked; (b) specify how many adjacent data points have been blocked within each condition (the number of data points blocked across conditions should be the same); (c) provide a rationale for blocking, assuring reader that blocking was not used to mask data variability, but rather to accentuate data trend and/or reduce figure size due to practical constraints (e.g., not blocking the data would have resulted in an illegible figure when duplicated); and (d) present a minimum of three blocked data points for each condition or phase, thereby allowing the reader to evaluate trend within each condition. As a rule, blocking is done post hoc; during the course of research all data are plotted. It is only after the study has been completed, and all data collected, that you can evaluate the appropriateness of the blocking procedure. The general rule regarding blocking is: Don't; if you must, proceed with caution and assure your reader that blocked data trends parallel and accurately

represent unblocked data.

Some researchers will add trend, median, and mean lines to their graphs to supplement point-by-point data plotted on a line graph; they should never be drawn as an alternative to plotting actual data points and data paths. These summative lines should be used sparingly, and as a general rule, we do not recommend their use. Their function is to highlight data trends and averages within and across conditions. When present on a graph, reviewers should not allow these to distract them from actual day-to-day data. These lines may distract readers from potential trend and variability present in a data path within and between conditions and make graphs needlessly complex.

When data are collected for multiple participants and you are plotting a statistical average (mean, median, or mode) of participant responses, you should generally plot, or specify in the text, the numerical range of responses. The range of responses for a group of participants has sometimes been shown on a figure by drawing a vertical line above and below the plotted data point to the upper and lower levels along the ordinate, thereby showing the two levels between which all students' responses fell. Range is an important statistic for readers when averages are plotted. It permits readers to evaluate consistency or stability of an individual or group's behavior within each condition.

Using Computer Software to Construct Graphs

Most researchers rely on computer software to graph their data, although practitioners may sometimes graph data by hand. Efforts to aide practitioners and applied researchers have been made by several authors specifically focusing on the Microsoft Office™ software typically loaded on personal computers (Barton, Reichow, & Wolery, 2007). Lo and Konrad (2007) and Carr and Burkholder (1998) outline steps for using Microsoft Excel™ to create a variety of SCD research design graphs while Hillman and Miller (2004) describe using Microsoft Excel™ for creating multiple baseline graphs. Alternatives to graphing within spreadsheets (e.g., Microsoft Excel™) include using Microsoft Word™ (Grehan & Moran, 2005) and Microsoft PowerPoint™ (Barton et al., 2007). With the availability of graphing software, it is important to note that adherence to all aforementioned guidelines for selecting and constructing graphic displays is crucial.

Tables

An alternative format for reporting data is the table. Data often reported in tables include participant demographics, condition variables, response definitions with examples and non- examples, and secondary data (e.g., reliability statistics, social validity data, generalization outcomes, number of trials or errors to criterion). Using a table to report supplemental or summative data can accomplish several things. Given the limited space of journal articles, presenting lengthy information in tabular form can condense it considerably. Table 7.1 shows a table with a considerable amount of information; displayed in a table, the data are organized and more comprehendible to readers. Tables also enable easy comparison of data. The trials to criterion and errors to criterion found in Table 7.2 are easily compared across sets of stimuli and participants. Occasionally, tables are used to demonstrate magnitude of data. Table 7.3 illustrates this by showing acquisition and maintenance of observational and incidental information (nutritional facts) for high school learners with moderate intellectual disability while being taught to bag groceries. Inserting the solid line draws the readers' attention to the immediate effect intervention had on students' behavior and that the effect was replicated within a multiple probe across dyads. Data gathered using a Likert scale are frequently summarized in tables. The information in Table 7.4 is a summary of selected social validity questions (cf. Hammond, Whatley, Ayres, & Gast, 2010). Without having to read each question and answer, the table allows readers to determine responses to each intervention component listed. Although tables efficiently highlight and summarize information, seldom are they used to present point-by-point data; rather, tables are primarily used for reporting supplemental or secondary data in SCD research studies. Tables provide an excellent format for summarizing some types of data; they are rarely used as a substitute for figures. When tables are used in research reviews, they should be used to summarize and synthesize variables to assist the reader in identifying relationships among studies; in general, text should supplement rather than duplicate information in tables.

Table 7.1 Organization of an Extensive Amount of Information

Instructional Data for Each Student and Training Set Through Criterion

Student/Set	No. training sessions	No. training trials	No. training errors	% training errors	Training time	Daily probe time	No. of Probe errors	% probe errors
Erol								
1	*15*	*45*	*0*	*0*	*135 min 18 s*	*45 min 05 s*	*17*	*37*
2	*11*	*33*	*0*	*0*	*99 min 16 s*	*33 min 08 s*	*9*	*27*
3	*5*	*15*	*0*	*0*	*45 min 44 s*	*15 min 12 s*	*5*	*33*
Total	*31*	*93*	*0*	*0*	*280 min 18 s*	*93 min 25 s*	*31*	*33*
Yunus								
1	*16*	*48*	*0*	*0*	*144 min 01 s*	*58 min 23 s*	*13*	*27*
2	*8*	*32*	*0*	*0*	*96 min 55 s*	*24 min 01 s*	*8*	*25*
3	*5*	*15*	*0*	*0*	*45 min 22 s*	*15 min 15 s*	*4*	*26*
Total	*29*	*95*	*0*	*0*	*286 min 18 s*	*97 min 39 s*	*25*	*26*
Yasemin								
1	*12*	*60*	*0*	*0*	*96 min 36 s*	*48 min 36 s*	*21*	*35*
2	*16*	*80*	*0*	*0*	*128 min 19 s*	*64 min 42 s*	*39*	*48*
3	*5*	*25*	*0*	*0*	*40 min 05 s*	*20 min 23 s*	*11*	*44*
Total	*33*	*165*	*0*	*0*	*265 min 00 s*	*133 min 41 s*	*71*	*43*
Grand Total	*93*	*353*	*0*	*0*	*831 min 36 s*	*324 min 45 s*	*127*	*35*

Source: Birkan, B. (2005). Using simultaneous prompting for teaching various discrete tasks to students with mental retardation. *Education and Training in Developmental Disabilities, 40*, 68–79.

Table 7.2 Example Table Showing Trials and Errors to Criterion Across Stimuli Sets for Three Students.

Participant & behavior set	Stimuli	Number of trials (days) to criterion	Errors to criterion
Colin			
Set 1	*blue, six*	*40 (5)*	*7.5*
Set 2	*pink, nine*	*24 (3)*	***0.0***
Set 3	*yellow, one*	*24 (3)*	***0.0***
Ser 4	*red, three*	*24 (3)*	***0.0***
Derek			
Set 1	*from, with*	*32 (4)*	***0.0***
Set 2	*down, once*	*32 (4)*	***0.0***
Set 3	*little, pretry*	*24 (3)*	***0.0***
Set 4	*left, walk*	*32 (4)*	***0.0***
Dustin			

Set 1	*3 x 4, .5 x 6*	*36(5)*	***0.0***
Set 2	*4 x 7, .6 x 10*	*64(5)*	***12.5***
Set 3	*7 x 3, 8 x 4*	*88(11)*	***13.6***
Set 4	*6 x 7, 3 x 8*	–	–

Source: 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*, 14–25.

Table 7.3 Example Table Showing Acquisition of Observational and Incidental Information Percentages of Correct Responding for Observational Learning of Incidental Information.*

		Probe Sessions and Dates							
Dyads	**Learners**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
		8/30	*9/13*	*9/26*	*10/11*	*10/23*	*11/1*	*11/8*	*11/15*
1	*Adam*	*0*	*20*	*20*	*40*	*20*	*20*	*40*	*40*
	Pete	*0*	*100*	*100*	*100*	*100*	*100*	*100*	*80*
2	*Robert*	*0*	*0*	*60*	*60*	*40*	*60*	*20*	*20*
	Barbara	*0*	*0*	*60*	*60*	*60*	*60*	*60*	*20*
3	*Emma*	*0*	*0*	*0*	*100*	*100*	*100*	*100*	*100*
	Danny	*0*	*0*	*0*	*60*	*60*	*60*	*40*	*40*
4	*Dot*	*0*	*0*	*0*	*0*	*100*	*80*	*100*	*80*
	Jim	*0*	*0*	*0*	*0*	*60*	*60*	*20*	*20*
5	*Mary*	*0*	*0*	*0*	*0*	*0*	*100*	*80*	*100*
	Cindi	*0*	*0*	*0*	*0*	*0*	*80*	*60*	*0*
6	*Michael*	*0*	*0*	*0*	*0*	*0*	*0*	*40*	*40*
	Cathy	*0*	*0*	*0*	*0*	*0*	*0*	*100*	*80*

Mean percentage correct responding at FRIO probe = 73.3% (range: 20%–100%)
Mean percentage correct responding at Probe 8 = 51.6% (range 0%–100%) [d]
Solid lines represent the occurrence of instruction and exposure to incidental information. The first probe after the vertical line for each dyad represents data collected after each dyad had reached the FRIO criterion point in the grocery-bagging program. Intermittent probe sessions were conducted when each dyad reached FRIO criterion. The Probe Session 8 column represents probes conducted for each learner 1 week after Dyad 6 reached FRIO criterion for the grocery-bagging skill.

Note: Line inserted to denote pre-intervention (left) and post-intervention (right) measurement occasions.

Source: Wall, M. E., & Gast, D. L. (1999). Acquisition of incidental information during instruction for a response-chain skill. *Research in Developmental Disabilities*, *20*, 31–50.

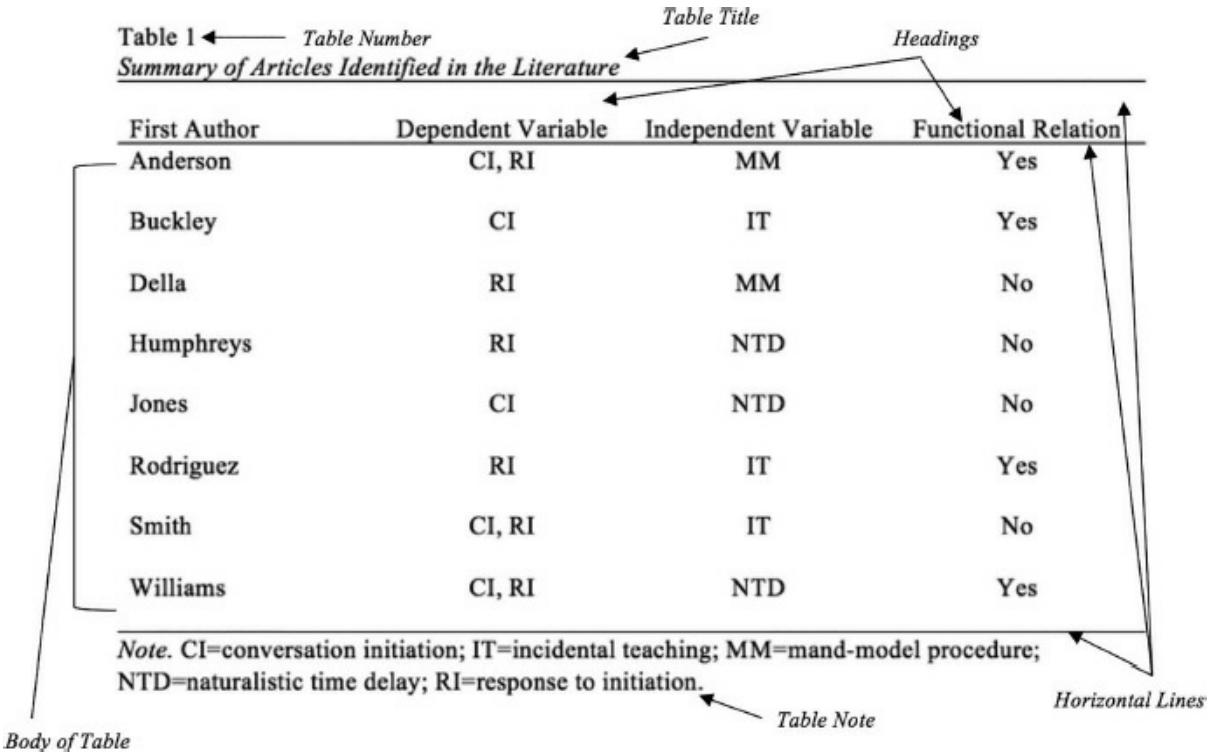
Table 7.4 Example Table Organizing Likert-Type Social Validity Data.

	Teacher responses on selected social validity issues[a]					
Intervention components	*Importance*	*n*	*Difficulty*	*n*	*Appropriateness*	*n*
Creating communicative opportunities	*Very (6–7)*	*9*	*Very (6–7)*	*0*	*Very (6–7)*	*9*
	Moderately (3–5)	*9*	*Moderately (3–5)*	*3*	*Moderately (3–5)*	*0*
	Not (1 2)	*0*	*Not (1 2)*	*6*	*Not (12)*	*0*
Modeling desired skill	*Very (6–7)*	*9*	*Very (6–7)*	*0*	*Very (6–7)*	*8*
	Moderately (3–5)	*0*	*Moderately (3–5)*	*1*	*Moderately (3–5)*	*1*
	Not (1 2)	*0*	*Not (1 2)*	*8*	*Not (1 2)*	*0*
Providing specific guidance	*Very (6 7)*	*7*	*Very (6 7)*	*0*	*Very (6 7)*	*9*
	Moderately (3–5)	*2*	*Moderately (3–5)*	*1*	*Moderately (3–5)*	*0*
	Not (1–2)	*0*	*Not (1–2)*	*8*	*Not (1–2)*	*0*

Note: n= 9
[a]*Teacher responses selected from anchored 7-poiut Likert scale*

Source: Johnston, S., Nelson, C., Evans, J., & Palazolo, K. (2003). The use of visual supports in teaching young children with autism spectrum disorder to initiate interactions. *Augmentative and Alternative Communication, 19*, 86–103.

Table 7.5 Basic Components of a Table



The *Publication Manual of the American Psychological Association, 6th edition*, outlines the parameters of table construction. Here, we discuss the elements most pertinent to applied SCD research; these components are outlined in Table 7.5: (a) Tables should be numbered in numerical order in the order they are mentioned in your text; (b) Table titles should be succinct; (c) Headings should be used to concisely organize the information you are sharing; (d) Subheadings may also be used under each heading, when necessary; (e) All headings should aid readers in finding pertinent information; (f)

Lines within tables should be limited to separating parts of the table to aid clarity for readers (e.g. around headings but not within the body); and (g) Vertical lines are not used. The size of a table will depend on the information being shared; careful consideration should be taken to fit the table within the text.

Summary

In this chapter we discussed the basic components and types of graphs and tables used to visually organize data collected in your research. Using the guidelines outlined in this chapter, you will be able to develop graphic displays appropriate for conducting visual analysis (see Chapter 8). While the information presented may appear cumbersome, it is imperative that you collect and organize your data accurately to ensure reliable data analysis. Selecting improper graphic displays or graphing data incorrectly may lead to unwarranted changes in instructional programs, incorrect conclusions of relations between dependent and independent variables, or unclear effects of interventions.

References

American Psychological Association. (2009). *Publication Manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Barton, E. E., & Reichow, B. (2012). Guidelines for graphing data with Microsoft Office 2007, Office 2010, and Office for Mac 2008 and 2011. *Journal of Early Intervention*, *34*, 129–150.

Barton, E. E., Reichow, B., & Wolery, M. (2007). Guidelines for graphing data with Microsoft® PowerPoint™. *Journal of Early Intervention*, *29*, 320–336.

Birkan, B. (2005). Using simultaneous prompting for teaching various discrete tasks to students with mental retardation. *Education and Training in Developmental Disabilities*, *40*, 68–79.

Carnine, D. W. (1976). Effects of two teacher presentation rates on off-task behavior, answering correctly, and participation. *Journal of Applied Behavior Analysis*, *9*, 199–206.

Carr, J. E., & Burkholder, E. D. (1998). Creating single-subject design graphs with Microsoft Excel™. *Journal of Applied Behavior Analysis*, *31*, 245–251.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Columbus, OH: Pearson.

Dart, E. H., & Radley, K. C. (2017). The impact of ordinate scaling on the visual analysis of single-case data. *Journal of School Psychology*, *63*, 105–118.

Grehan, P., & Moran, D. J. (2005). Constructing single-subject reversal design graphs using Microsoft Word™: A comprehensive tutorial. *The Behavior Analyst Today*, *6*, 235–256.

Hammond, D. L., Whatley, A. D., Ayres, K. M., & Gast, D. L. (2010). Effectiveness of video modeling to teach iPod use to students with moderate intellectual disabilities. *Education and Training in Autism and Developmental Disabilities, 45*, 525–538.

Heal, N. A., & Hanley, G. P. (2007). Evaluating preschool children's preference for motivational systems during instruction. *Journal of Applied Behavior Analysis*, *40*, 249–261.

Hillman, H. L., & Miller, L. K. (2004). Designing multiple baseline graphs using Microsoft Excel™. *The Behavior Analyst Today*, *5*, 372–424.

Hoch, H., McComas, J. J., Thompson, A. L., & Paone, D. (2002). Concurrent reinforcement schedules: Behavior change and maintenance without extinction. *Journal of Applied Behavior Analysis*, *35*, 155–169.

Johnston, S., Nelson, C., Evans, J., & Palazolo, K. (2003). The use of visual supports in teaching young children with autism spectrum disorder to initiate interactions. *Augmentative and Alternative Communication*, *19*, 86–103.

Kennedy, C. H. (1989). Selecting consistent vertical axis scales. *Journal of Applied*

Behavior Analysis, *22*, 338–339.

Kennedy, C. H., & Souza, G. (1995). Functional analysis and treatment of eye poking. *Journal of Applied Behavior Analysis*, *28*, 27–37.

Kubina, R. M., Kostewicz, D. E., Brennan, K. M., & King, S. A. (2017). A critical review of line graphs in behavior analytic journals. *Educational Psychology Review, 29,* 583–598.

Kubina, R. M., Morrison, R., & Lee, D. L. (2002). Benefits of adding precision teaching to behavioral interventions for students with autism. *Behavioral Interventions*, *17*, 233–246.

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., Lane, J. D., Shepley, C., & Kroll, S. M. (2016). Using teacher-implemented playground interventions to increase engagement, social behaviors, and physical activity for young children with autism. *Focus on Autism and Other Developmental Disabilities*, *31*, 163–173.

Ledford, J. R., Barton, E. E., Severini, K. E., Zimmerman, K. N., & Pokorski, E. (2017). Visual display of graphic data in single case design studies: Systematic review and expert preference analysis. Manuscript under review.

Lindsley, O. R. (1992). Precision teaching: Discoveries and effects. *Journal of Applied Behavior Analysis*, *25*, 51–57.

Lo, Y., & Konrad, M. (2007). A field-tested task analysis for creating single-subject graphs using Microsoft® Office Excel. *Journal of Behavioral Education*, *16*, 155–189.

Maglieri, K. A., DeLeon, I. G., Rodriguez-Catter, V., & Sevin, B. M. (2000). Treatment of covert food stealing in an individual with Prader-Willi syndrome. *Journal of Applied Behavior Analysis*, *33*, 615–618.

Mayfield, K. H., & Vollmer, T. R. (2007). Teaching math skills to at-risk students using home-based peer tutoring. *Journal of Applied Behavior Analysis*, *40*, 223–237.

Parsonson, B. S., & Baer, D. M. (1978). The analysis and presentation of graphic data. In T. Kratchwill (Ed.), *Single subject research: Strategies for evaluating change*. New York, NY: Academic Press.

Sanders, R. M. (1978). *How to plot data*. Lawrence, KS: H & H Enterprises.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, *14*, 47–69.

Shepley, S. B., Spriggs, A. D., Samudre, M., & Elliot, M. (2017). Increasing daily living independence using video activity schedules in middle school students with intellectual disability. *Journal of Special Education Technology*. doi: 10.1177/0162643417732294

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CN: Graphics Press.

Vanselow, N. R., & Bourret, J. C. (2012). Online interactive tutorials for creating graphs with Excel 2007 or 2010. *Behavior Analysis in Practice*, *5*, 40–46.

Wall, M. E., & Gast, D. L. (1999). Acquisition of incidental information during instruction for a response-chain skill. *Research in Developmental Disabilities*, *20*, 31–50.

Wilder, D. A., Atwell, J., & Wine, B. (2006). The effects of varying levels of treatment integrity on child compliance during treatment with a three-step prompting procedure. *Journal of Applied Behavior Analysis*, *39*, 369–373.

Williams, G., Perez-Gonzalez, L. A., & Queiroz, A. B. (2005). Using a combined blocking procedure to teach color discrimination to a child with autism. *Journal of Applied Behavior Analysis*, *38*, 555–558.

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*, 14–25.

8

Visual Analysis of Graphic Data

Erin E. Barton, Blair P. Lloyd, Amy D. Spriggs, and David L. Gast

Important Terms

formative visual analysis, behavior change, summative visual analysis, functional relations, adjacent conditions, within condition visual analyses, level, level change, level stability, trend, trend direction, accelerating, decelerating, zero celerating, steep, gradual, trend stability, variability, overlap, immediacy, stability, between conditions visual analyses, consistency, potential demonstrations of effect, demonstrations of effect, magnitude

Visual Analysis of Graphic Data

In single case design (SCD) research, replication is used to establish causal inference through repeated demonstrations of behavior change that coincide with changes in conditions. You should evaluate effectiveness of your independent variable (i.e., intervention) by repeatedly collecting, graphing, and analyzing data in the context of an appropriate SCD (Wolery & Harris, 1982). Visual analysis of graphed data, in contrast to statistical analysis of data, is the cornerstone of and most frequently used data analysis method in SCD research, particularly for determining whether a study demonstrates experimental control (Horner & Spaulding, 2010; Kratochwill et al., 2013). Researchers must be in constant contact with their data to ensure research participants are experiencing success and improvement; data should be collected repeatedly, graphed regularly, and analyzed frequently during an SCD study. The dynamic, formative nature of visual analysis facilitates an iterative process to identify effective interventions related to meaningful outcomes (Parsonson & Baer, 1978).

Visual analysis involves systematic procedures used to evaluate specific characteristics of data patterns and evaluate the presence of a functional relation. It facilitates formative evaluation of intervention effectiveness allowing for close examination of the data over time and across conditions. Visual analysis has several advantages. First, it can be used to evaluate data of individuals or small groups depending on the unit of analysis specified in the research question. Second, it is a dynamic process in that data are collected repeatedly, graphed as they are collected, and analyzed frequently. Formatively graphing and reviewing data facilitates informed, data-based decisions (e.g., condition changes, intervention adaptations) to ensure participants benefit from their involvement (Barton et al., 2016; Gast & Spriggs, 2014; Wolery, 2013). Third, visual analysis focuses on analysis of individual data patterns, thereby facilitating individualization, rather than group-based generalizations. Fourth, visual analysis of graphic data permits discovery of potentially interesting findings that may not be directly related to the original research question or program objective (Barton et al., 2016; Wolery, 2013). Unplanned or serendipitous findings (Sidman, 1960; Skinner, 1957) are possible because "primary" data are collected, graphed, and analyzed regularly; thus, formative analysis is critical to the dynamic nature of SCD research. Finally, graphic presentation of data permits independent analysis and interpretation of results (Parsonson & Baer, 1978). This transparency allows others to judge for themselves whether interventions have merit and whether the magnitude of results are socially valid. For these reasons, visual analysis of graphic data is the preferred strategy for SCD researchers. It is an approach that has proven to be both practical and reliable; therefore, we recommend its use.

Using Visual Analysis to Identify Behavior Change and Functional Relations

Visual analysis involves both formative and summative evaluation. **Formative visual analysis** is conducted within and across conditions to identify behavior change during the course of a study. **Behavior change** occurs when data patterns in one condition are different from data patterns in the subsequent, adjacent condition for the same variable(s). **Summative visual analysis** is conducted following study completion, across multiple opportunities to demonstrate behavior change to determine whether a **functional relation** exists between the independent variable and the dependent variable. Thus, visual analysis is used to make experimental decisions (formative, behavior change), identify the presence or absence of a functional relation (summative), and assess the magnitude of the effect (summative). As described in other chapters (9–13), recent standards for SCDs advocate for at least three demonstrations of temporally-related and consistent behavior change to establish experimental control and identify functional relations (e.g., Horner et al., 2005; Kratochwill et al., 2013), and recent research suggests that visual analysts adhere to this standard (Wolfe, Seaman, & Drasgow, 2016). In SCD research, data patterns are examined within and across **adjacent conditions**; when data in one condition differ from what is predicted based on the preceding condition, behavior change is demonstrated. Formative analysis is conducted in two steps: (1) within and across adjacent conditions analyses and (2) systematic examination of specific data characteristics; these are delineated in the next sections.

Formative Visual Analysis: Within Condition Analyses

In SCD research, data are graphed and analyzed throughout the study as data are collected (Parsonson & Baer, 1978). This process is dynamic, and experimental decisions are made based on data patterns. **Within condition visual analyses** are conducted to discern patterns within a single condition during a study. Within condition analyses of level, trend, and variability/stability are critical for determining when to change conditions, deciding whether adaptations need to be made, and providing information related to answering research questions. Data-informed decisions can be made, which might result in new, unexpected, or replicated findings that can be used to inform existing interventions.

Beginning with the initial condition—typically baseline—you should look for stability of data across a minimum of at least three to five sessions prior to changing conditions. You should make a priori decisions to set a criterion for changing conditions; however, the criterion should be based on within condition data patterns—not a set or maximum number of data points. For example, you might establish a condition change criterion of

stable responding at or near floor levels for a minimum of three consecutive data points before commencing intervention. Likewise, you might set a mastery criterion of three data points at a stable level of at least 90% unprompted correct trials for moving from an intervention to a maintenance condition. Condition change criteria should be made a priori based on hypothesized data patterns. These criteria will guide both formative and summative decisions about experimental control.

Figure 8.1 depicts two graphs from a study conducted by Hughes, Alberto, and Fredrick (2006); they examined use of an auditory prompting system for decreasing off-task behaviors in high school students with intellectual disabilities in community job sites. Within condition analyses for the initial baseline condition determined data were at a stable level—above 75%—with low variability and indicated a need for intervention (i.e., off-task behaviors were unacceptably high). Thus, it was appropriate to introduce the intervention. Although not specified by the authors, a minimum of 3 data points at a stable level might have been an appropriate condition change criterion for this study. However, given the level change between the first and second data points, additional data were needed to establish a stable level (e.g., it is possible that behavior change between data points 1 and 2 could have been the beginning of a consistent decreasing trend, indicating possible maturation). Thus, within-condition formative analysis was potentially used to delay the condition change.

Level

The term **level** refers to the amount of behavior that occurs, as indicated by the ordinate scale value (Kennedy, 2005). Level is often the characteristic of highest interest for behavior change, and is generally described as low, moderate, or high. You can also characterize level by describing the range of dependent variable values in a condition (e.g., 10% or fewer of intervals; 90–100% accuracy). Less often, it has been described as a median value. Despite the typical use of means/averages (e.g., in between group research and outside of research contexts), we do not recommend their use for summarizing level because the relatively small number of data points make the mean susceptible to outliers (i.e., results in the mean being a poor representation of level).

The bottom panel of Figure 8.1 displays percentage of intervals with noncompliance for Veronica. The percentage of intervals with noncompliance are at a high level, at or above 75% of intervals, for the initial *no prompts* condition and a low level, 15% of intervals or fewer, for the initial *reminder* condition. The subsequent conditions show similarly stable levels.

Occasionally researchers make condition changes when data levels are not

stable, although within-condition level change might compromise confidence in decisions about behavior change and functional relations. For example, data in Figure 8.2 are from a study conducted by Dunlap and colleagues (1994); they examined use of choice making on task engagement and disruptive behaviors in elementary school students with emotional and behavioral disorders. Figure 8.2 shows data from Wendell, one of three study participants. Wendell had low level stability across both dependent variables in baseline: disruptive behavior (range = 10–45% of intervals) and task engagement (range = 40–95% of intervals). During the initial intervention condition, he had stable levels at or near 100% of intervals for task engagement and at or near 0% of intervals for disruptive behavior.

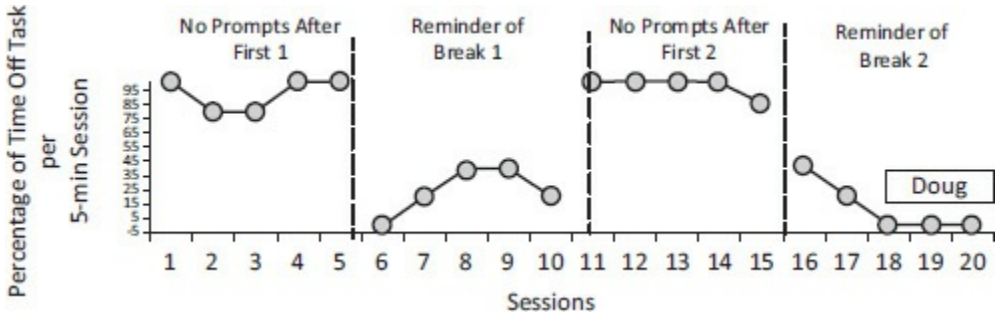


Figure 8.1a Representation of high levels in baseline conditions, used in within-condition analysis to make formative decisions.

Source: Hughes, M. A., Alberto, P. A., & Fredrick, L. L. (2006). Self-operated auditory prompting systems as a function-based intervention in public community settings. *Journal of Positive Behavior Interventions*, *8*, 230–243.

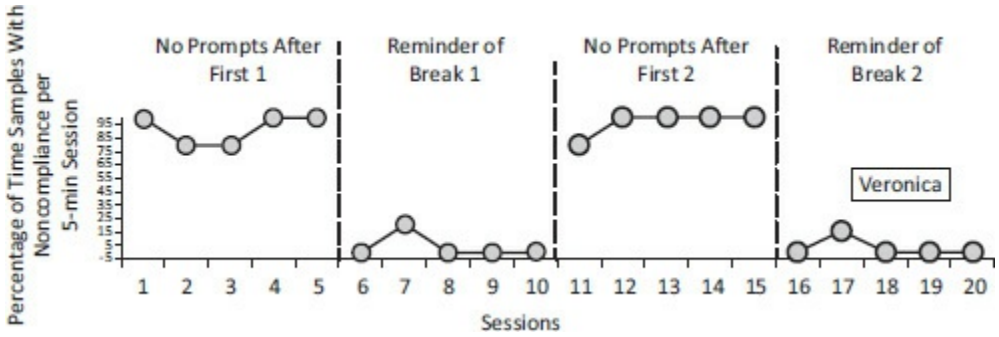


Figure 8.1b Representation of high levels in baseline conditions, used in within-condition analysis to make formative decisions.

Source: Hughes, M. A., Alberto, P. A., & Fredrick, L. L. (2006). Self-operated auditory prompting systems as a function-based intervention in public community settings. *Journal of Positive Behavior Interventions*, *8*, 230–243.

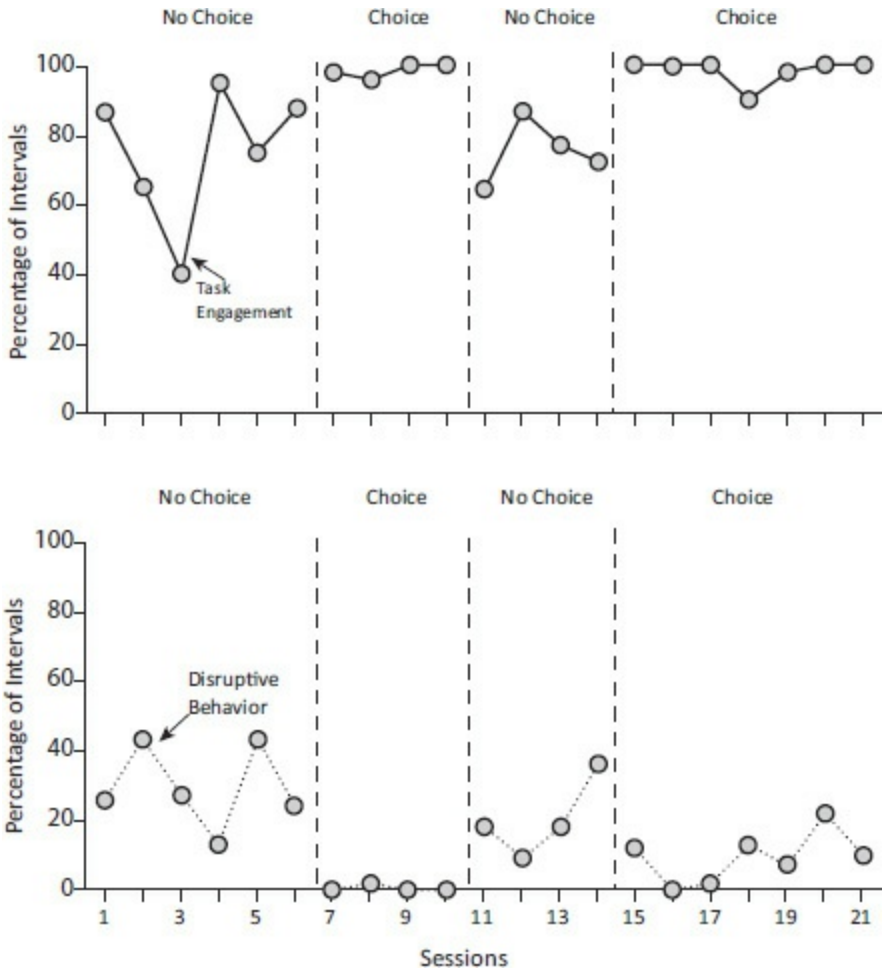
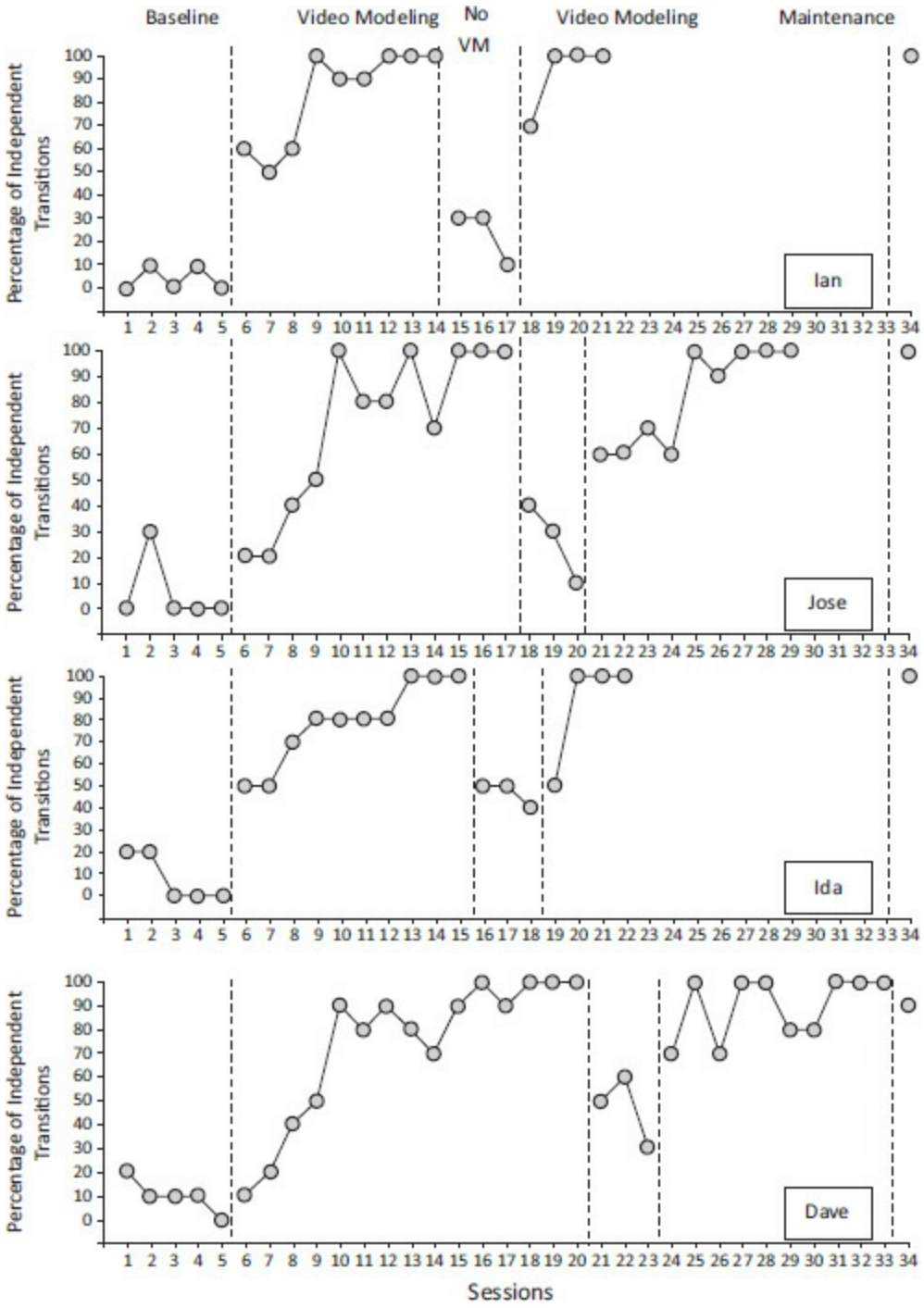


Figure 8.2 Representation of variability in levels of both dependent variables in baseline conditions; if data variability are expected in baseline conditions, stability is not always required for formative condition change decisions, although summative decisions regarding functional relations may be impacted.

Source: Dunlap, G., DePerczel, M., Clarke, S., Wilson, D., Wright, S., White, R., & Gomez, A. (1994). Choice making to promote adaptive behavior for students with emotional and behavioral challenges. *Journal of Applied Behavior Analysis, 27*, 505–518.



Figurue 8.3 Representation of low levels in baseline conditions, used in within-condition analysis to make formative decisions.

Source: Cihak, D., Fahrenkrog, C., Ayres, K. M., & Smith, C. (2010). The use of video modeling via a video iPod and a system of least prompts to improve transitional behaviors for students with autism spectrum disorders in the general education classroom. *Journal of Positive Behavior Interventions*, *12*, 103–115.

Figure 8.3 is from a study conducted by Cihak, Fahrencog, Ayres, and Smith (2010); they examined use of video modeling delivered via iPad for improving transition behaviors in students with autism spectrum disorders (ASD). Data show relatively low levels of correct performance during baseline conditions, with ranges varying

from 0–10% independent transitions to 0–30% independent transitions.

Trend

Trend is the slope and direction of a data series or the direction data are moving over time (increasing, decreasing, or remaining the same; Kennedy, 2005). When visually analyzing data, three characteristics can be described: trend direction, trend magnitude, and trend stability. **Trend direction** is referred to as **accelerating** (increasing in ordinate value over time), **decelerating** (decreasing in ordinate value over time), or **zero celerating** (data series is parallel to the abscissa). Trend can further be characterized by magnitude, and is often described as **steep** or **gradual** and paired with direction (e.g., steep accelerating trend or gradual decelerating trend). You should also describe whether the direction of a trend is improving (therapeutic) or deteriorating (contra-therapeutic) based on the behavior of interest (e.g., a steep accelerating trend during intervention is desirable for acquisition of target behaviors, but the same trend is undesirable [contra-therapeutic] if the goal is to decrease problem behaviors). To increase confidence in functional relations, trend direction and stability should align with hypothesized data patterns.

The contra-therapeutic trend represents a common data pattern in SCD data that might occur within a condition and particularly prior to the introduction of the independent variable. Contra-therapeutic trends refer to trends that are in the opposite direction of the hypothesized direction of improvement and can establish need for the intervention. Though contra-therapeutic trends occurring in baseline might seem to provide evidence that immediate intervention is needed, it is optimal to collect data until stability is established, due to the possibility of regression to the mean (i.e., that data are likely to improve even without intervention based on random fluctuations; Kazdin, 2011).

Variability

Variability is fluctuation from one data point to the next and is the opposite of stability; in data with no trend (i.e., zero celerating), variability can be summarized as the range of data values within a condition or as the percentage of data points falling within a given stability envelope (Franklin, Gorman, Beasley, & Allison, 1996; see Tools section below). In data with trends, it can be calculated via a stability envelope around a split middle trend line (Lane & Gast, 2013, see Tools section below). However, in general, data are described as stable or variable without numerical quantification (Kennedy, 2005). Variability might be a function of extraneous events (e.g., health issues, sporadic sleep patterns, caregiver changes), which can be temporary or permanent. Data are generally reported as either highly variable, somewhat variable, or stable; there are no guidelines

for quantifying the magnitude of variability. It is generally recommended that conditions be extended when data patterns are highly or somewhat variable. However, highly variable data might establish need for an intervention that produces stable levels of responding. That is, variability might be the predicted pattern of the dependent variable under baseline conditions, in which case condition changes might proceed if the expected pattern of behavior change is a decrease in variability. In general, even when expected, variability indicates need for additional data in a condition (e.g., more than the minimal three data points; Kennedy, 2005; Parsonson & Baer, 1978); additional data establish that variability is likely to continue in the absence of intervention.

As shown in the initial *video modeling* condition in Figure 8.3, Jose's, Ida's, and Dave's behaviors show an accelerating trend with high trend stability, which is the predicted pattern of change.

Figure 8.4 is a graph from a study conducted by Jones, Lerman, and Lechago (2014); they used video modeling to teach social responses to children with ASD. During the intervention condition, percentage of correct responses had a steep, accelerating trend with high trend stability across participants.

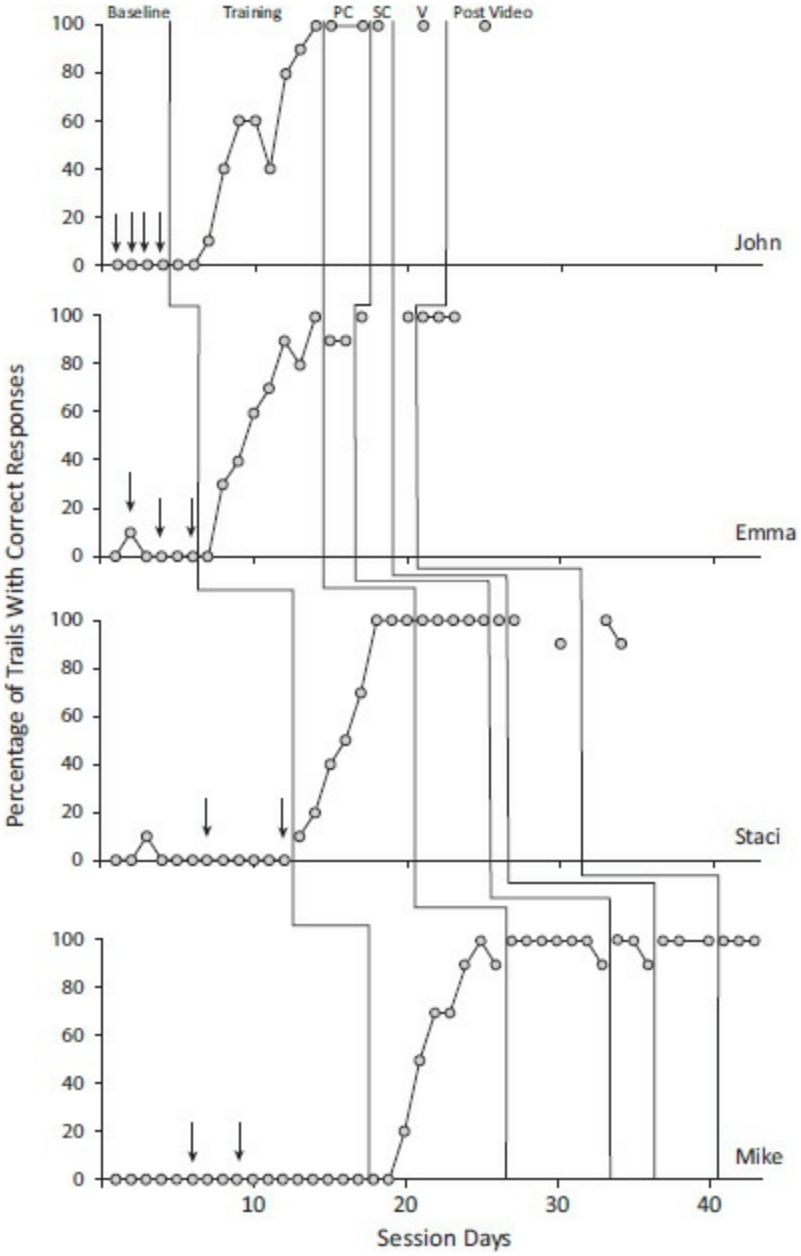


Figure 8.4 Representation of a steep accelerating trend during training conditions.

Source: Jones, J., Lerman, D. C., & Lechago, S. (2014). Assessing stimulus control and promoting generalization via video modeling when teaching social responses to children with autism. *Journal of Applied Behavior Analysis*, *47*, 37–50.

Figure 8.5 shows graphs from a study conducted by Wills and Mason (2014); they examined effects of a technology-aided self-monitoring intervention on the on-task behaviors of two high school students with disabilities. For the participant shown in Figure 8.5a, a decelerating, contra-therapeutic trend in on-task behavior occurred during the initial baseline condition; however, the final 5 data points demonstrated more stability.

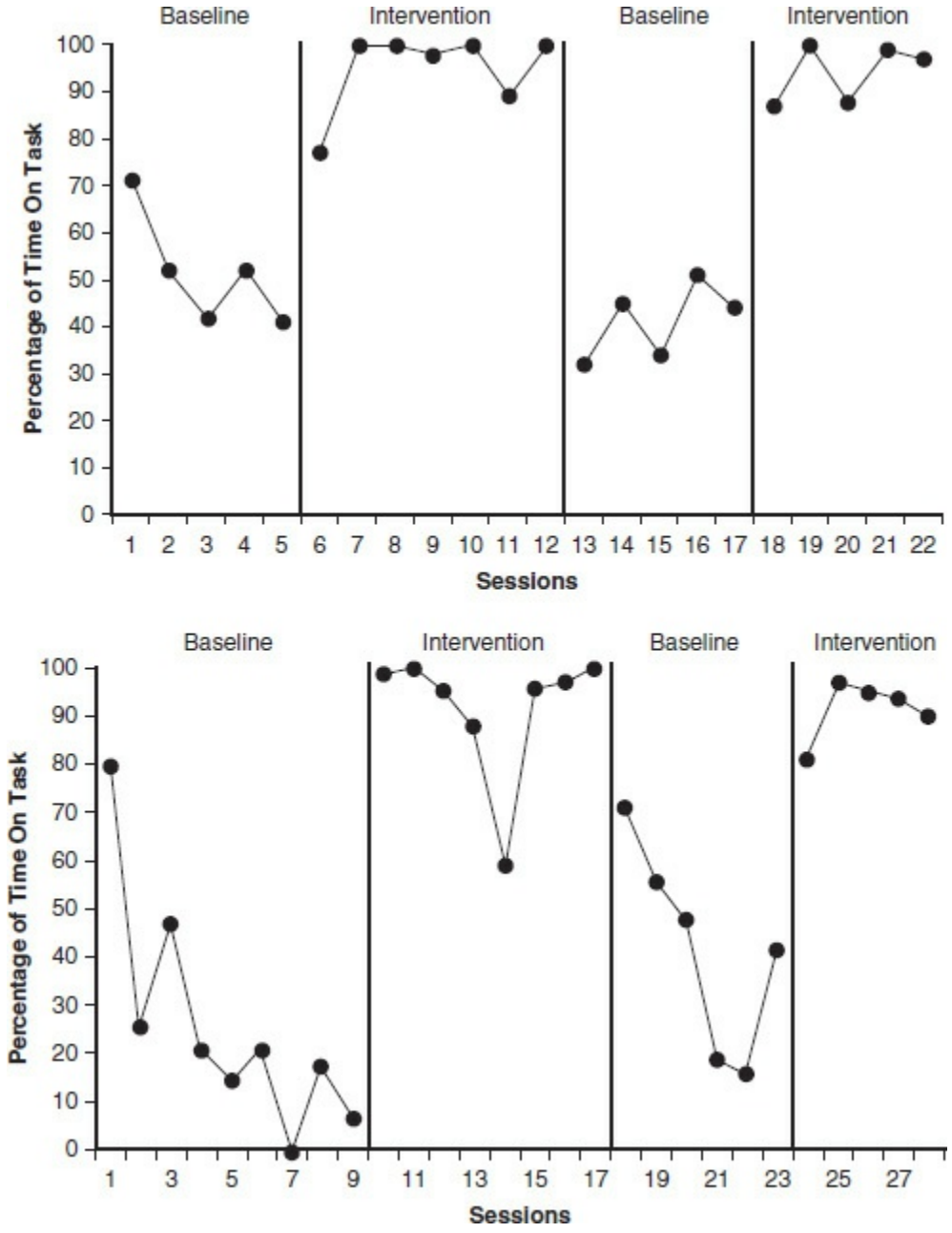


Figure 8.5a-b Representation of a contra-therapeutic trend during baseline conditions.

Source: Wills, H. P., & Mason, B. A. (2014). Implementation of a self-monitoring application to improve on-task behavior: A high-school pilot study. *Journal of Behavioral Education, 23,* 421–434.

Figure 8.6 is a graph from an article describing several studies (Barton et al., 2016); the one represented here was an assessment of environmental arrangement (EA) with and without a system of least prompts (SLP) on rate of social interactions of young children. Initially, baseline data were highly variable (although relatively stable for the final three data points), and introduction of the EA condition resulted in similarly variable data. Thus, a modification was implemented during the third

condition; this condition resulted in considerably less variable data.

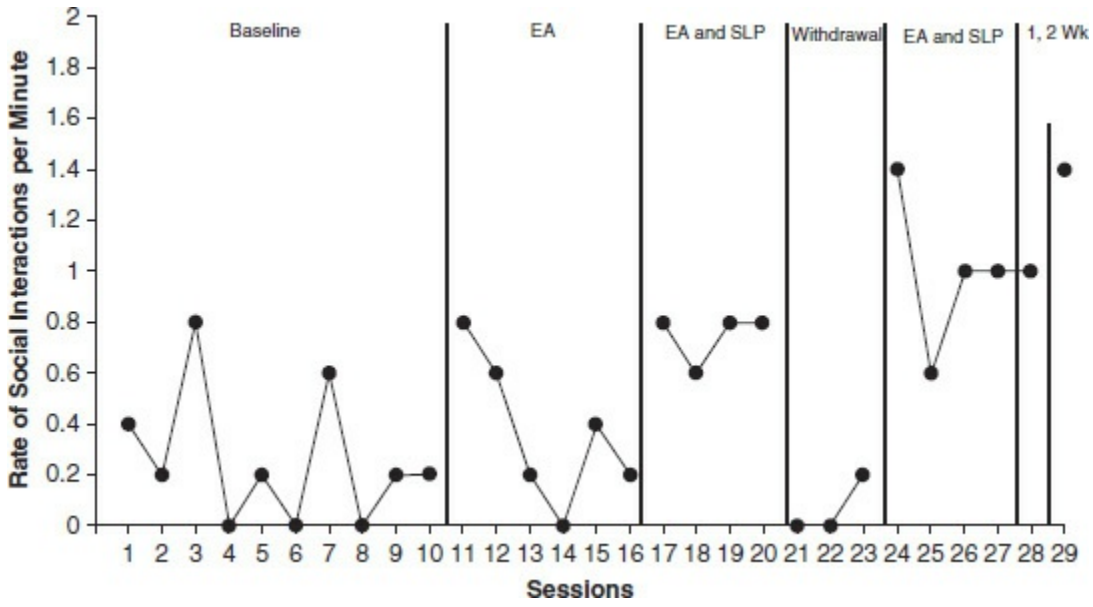


Figure 8.6 Representation of high levels of variability in baseline and initial intervention conditions, used in within-condition analysis to make formative decisions about condition modifications.

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

Stability

Stability is predictability and consistency of data values within a condition (Hersen & Barlow, 1976) or lack of fluctuations in adjacent data points (i.e., lack of variability). Perceptions of stability can be influenced by scales and ranges of y-axes (Parsonson & Baer, 1978; see Chapter 7). Data stability assumes that in the absence of environmental changes, the current data pattern would not change. Data can demonstrate level stability or trend stability (or both). Calculating a stability envelope is one way to quantify stability; calculations for the stability envelope are described below.

Figure 8.7 is a graph from a study by Barton, Fuller, and Schnitz (2016); they examined use of performance-based feedback sent via email to increase pre-service teachers' use of recommended practices. Jasmine had low stable levels of all three target behaviors—emotion labeling, choices, and promoting social interactions—during baseline conditions, which indicates it was appropriate to commence intervention. During email feedback she had relatively stable levels of choices (range = 10–21) and prompting social interactions (range = 15–18).

Figure 8.8 is a graph from a study by Adamo and colleagues (2015); they used a

multi-component intervention package (including modeling, prompting, and praise) to increase physical activity in young children with Down syndrome. Ramona had stable levels of unprompted moderate to vigorous physical activity (MVPA) during the initial baseline condition (range = 0–11% of intervals), which indicates it was appropriate to commence intervention. During the initial intervention condition, Ramona's MVPA was stable at a moderate level (range = 20–32% of intervals). The second baseline condition also had a low, stable level (range = 3–12% of intervals). During the second intervention condition, Ramona had a stable level of MVPA at a moderate level (range = 26–31% of intervals).

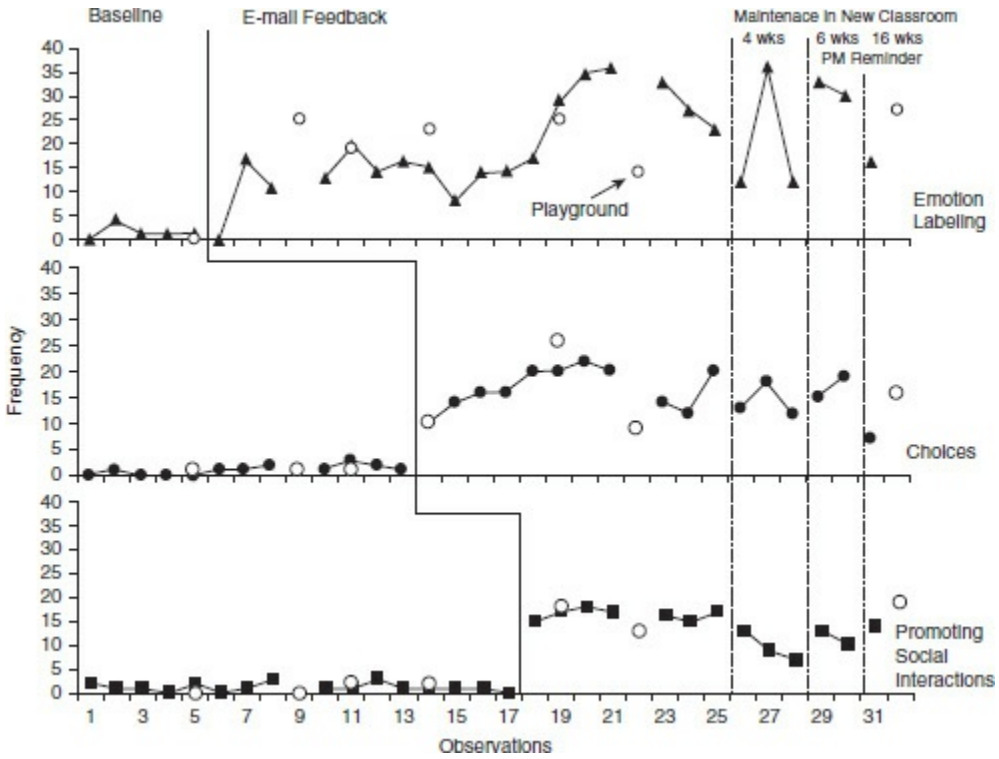


Figure 8.7 Representation of stable baseline data, used in within-condition analysis to make formative decisions.

Source: Barton, E. E., Fuller, E. A., & Schnitz, A. (2016). The use of email to coach preservice early childhood teachers. *Topics in Early Childhood Special Education*, *36*, 78–90.

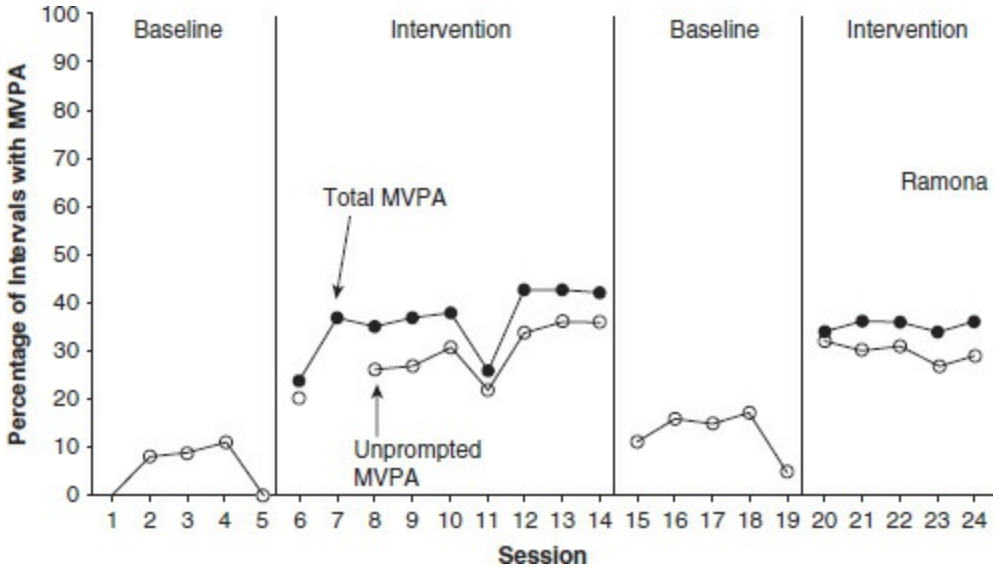


Figure 8.8 Representation of stable baseline data, used in within-condition analysis to make formative decisions.

Source: Adamo, E. K., Wu, J., Wolery, M., Hemmeter, M. L., Ledford, J. R., & Barton, E. E. (2015). Using video modeling, prompting, and behavior-specific praise to increase moderate-to-vigorous physical activity for young children with Down syndrome. *Journal of Early Intervention*, *37*, 270–285.

Formative Visual Analysis: Adjacent Condition Analyses

Visual analysis can be used throughout a study to make informed decisions or changes about design and study variables while maintaining experimental control *and* producing improved outcomes. The objective of **between conditions visual analysis** (adjacent conditions analysis) is to identify if behavior change has occurred. In SCD research a particular condition (B) is introduced and re-introduced to one (e.g., A-B-A-B design) or more than one (e.g., multiple baseline design) data series to evaluate whether there is a functional relation between independent and dependent variables. **Functional relations** are unequivocal demonstrations that an independent variable (intervention) produced reliable and consistent change in a dependent variable (target behavior). The purpose of SCD research is to determine if behavior change occurs when the intervention is introduced, and whether the behavior change can be reliably replicated. When conducting a between conditions analysis it is important to remember that only data in adjacent conditions can be directly compared. Condition change decisions are made formatively by examining level, trend, and variability and comparing hypothesized changes to actual data patterns. Typically, once data stabilize in level or trend in the predicted direction, magnitude, or pattern, you can introduce the next planned condition or end the study. Analysis of data across adjacent conditions entails determining: (a) changes in data patterns (level, trend, and variability), (b) immediacy of change, (c) amount of overlapping data across adjacent conditions, and (d) consistency of data

patterns across similar conditions.

Changes in Data Patterns

When comparing data across two adjacent conditions, data patterns immediately prior to and following the condition change should be examined. Generally, researchers are most interested in changes in level and/or trend direction. For example, researchers are likely to be interested in the level of data when studying problem behavior under different conditions and are likely to be interested in trend when studying reading rates.

> In Figure 8.4, there was a stable, low level with a zero celerating trend during initial baseline conditions. Conversely, during intervention conditions, the graph shows a steep, accelerating trend with high trend stability. Thus, changes in data patterns occurred, and were consistent across tiers.

Immediacy of Change

Immediacy of change across adjacent conditions is the degree to which behavior change occurs as soon as the intervention is introduced (Horner et al., 2005). When a large change in level occurs immediately after introduction of a new condition, it is referred to as an *abrupt* change in level, which is indicative of an immediately "powerful" or immediately effective intervention (Parsonson & Baer, 1978). Generally, immediate and abrupt change in the dependent variable that coincides with a condition change provides a clear indication of behavior change. The more rapid (or immediate) the effect, the more convincing the inference that change in outcome measures was due to manipulation of the independent variable. However, delayed changes might occur and do not necessarily preclude identification of functional relations; in these cases, confidence in functional relations is increased when (a) delay is predicted a priori (e.g., as might be the case with some academic skills), (b) latency to change (number of data points prior to change) is consistent across conditions or tiers, and (c) magnitude of change in level or trend is consistent across conditions or tiers (Lieberman, Yoder, Reichow, & Wolery, 2010; Parsonson & Baer, 1978).

Overlap

Overlap refers to values of data in one condition that are in the same range of values of data in the subsequent, adjacent condition (Kennedy, 2005). Overlap can be reported as the proportion of data from one condition that is of the same level as data from an adjacent condition (e.g., percentage of overlapping data). Confidence in behavior change and the presence of a functional relation is inversely related to the proportion of

overlapping data across adjacent conditions (Parsonson & Baer, 1978). Larger separation and smaller proportion of overlap are generally associated with more compelling demonstrations of effect.

Figure 8.9 represents a study conducted by Fettig, Schultz, and Sreckovic (2015); they examined use of coaching on parents' implementation of function-based interventions to reduce their children's challenging behaviors. The percentage of parent-implemented strategies used had a low, stable level during initial baseline conditions across all children, which indicated parents were not using function-based interventions. However, when training commenced, levels immediately increased across all children, and immediately increased again once coaching commenced. The percentage of parent-implemented strategies used eventually stabilized at ceiling levels (100%).

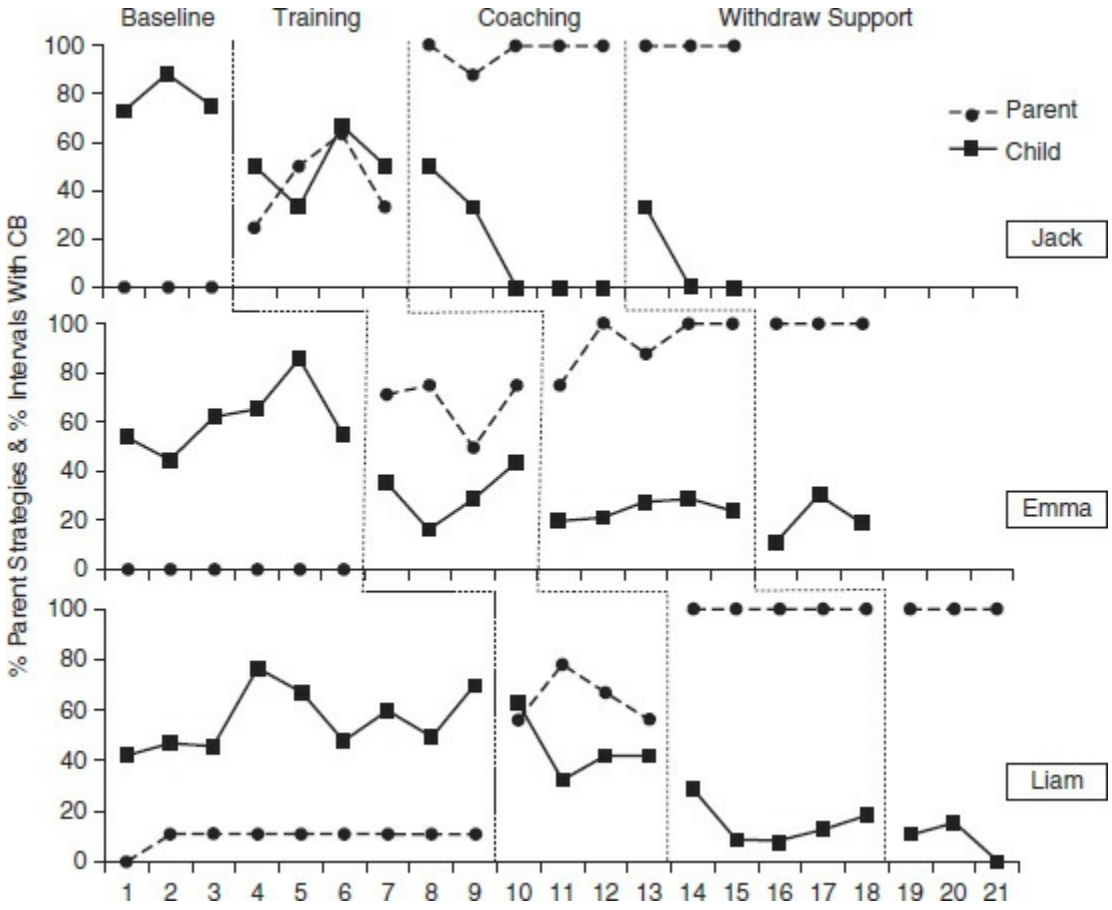


Figure 8.9 Representation of immediate and near-immediate change between baseline and training conditions.

Source: Fettig, A., Schultz, T. R., & Sreckovic, M. A. (2015). Effects of coaching on the implementation of functional assessment-based parent intervention in reducing challenging behaviors. *Journal of Positive Behavior Interventions*, *17*, 170–180.

In the top panel of Figure 8.1, there is no overlap in noncompliance across adjacent conditions (the minimum values of the first and third conditions is higher than maximum values of the second and fourth conditions). Thus, there was an immediate change, with no overlap, increasing confidence in the presence of a functional relation.

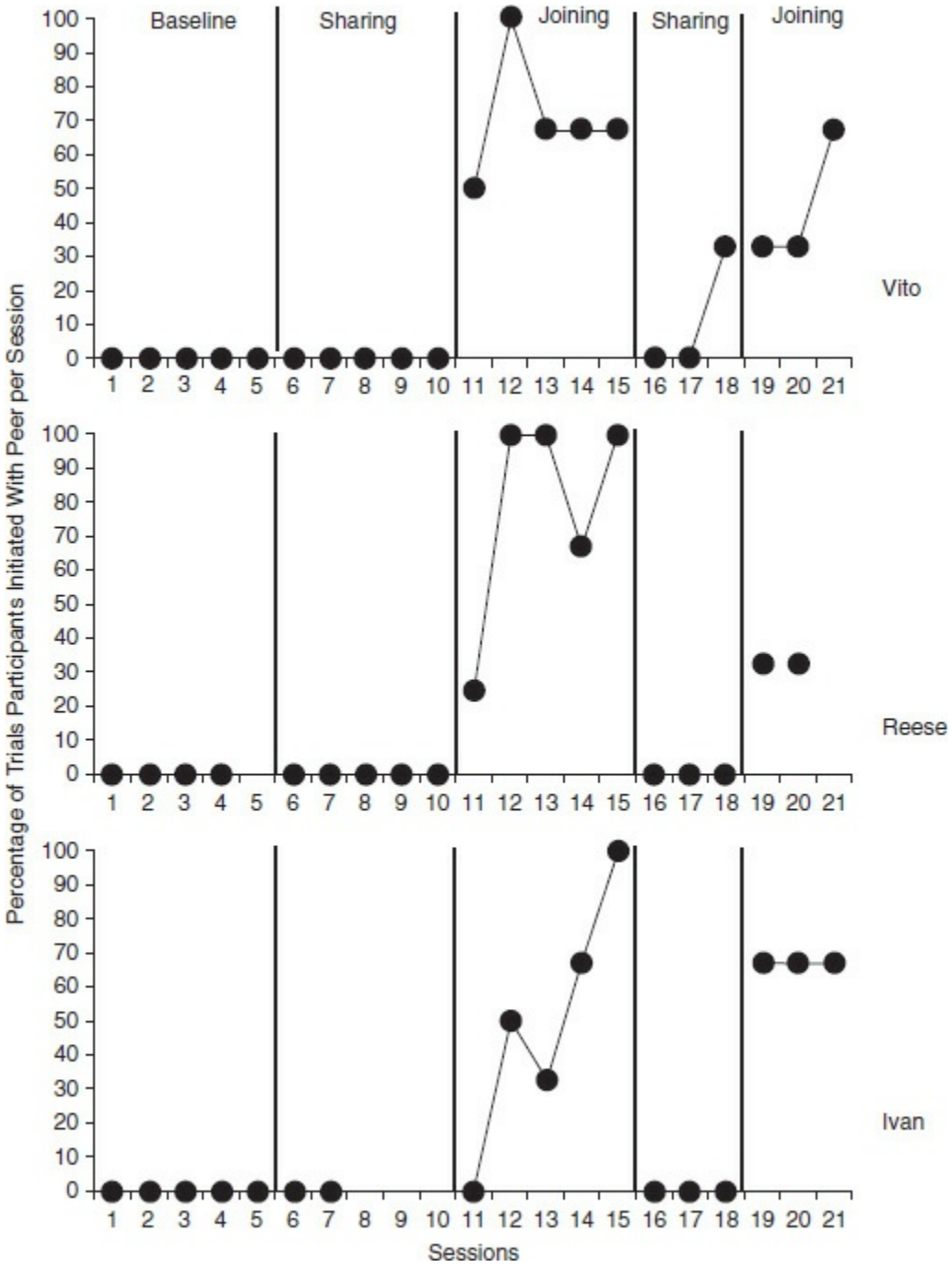


Figure 8.10 Representation of no overlap between conditions (Sharing and Joining) for Reese and low to moderate overlap between conditions for Vito and Ivan.

Source: Plavnick, J. B., MacFarland, M. C., & Ferreri, S. J. (2015). Variability in the effectiveness of a video modeling intervention package for children with autism. *Journal of Positive Behavior Interventions*, *17*, 105–

115.

Figure 8.10 represents a graph from a study by Plavnick, MacFarland, and Ferreri (2015); they used video modeling to teach three children with ASD to initiate to their peers. Each participant made no peer initiations during baseline and *sharing* conditions (i.e., low, stable levels). When the initial *joining* condition commenced, all children showed an immediate change in data patterns. Data for Vito and Reese showed an immediate change in level and no overlap with prior or subsequent adjacent conditions. Data for Ivan showed an immediate increase in trend with one overlapping data point (i.e., 20% of joining data points overlapped with prior or subsequent adjacent conditions). During the second intervention condition (i.e., joining), data for Reese and Ivan showed no overlapping data with the previous condition. However, two of three data points (67%) overlapped with the previous sharing condition for Vito.

Consistency

Consistency refers to the extent to which data patterns in one condition are similar to data patterns in other conditions (Parsonson & Baer, 1978). Confident determination that a functional relation exists requires consistency in data patterns between iterations of the same condition (e.g., Baseline 1 and Baseline 2) and inconsistency in data patterns between different, adjacent conditions (e.g., Baseline 1 and Intervention 1). Consistency also applies to behavior change across conditions. For example, the immediacy and magnitude of behavior change should be consistent each time similar condition changes occur.

Summative Visual Analysis

Identifying Functional Relations

Summative visual analyses are used to draw conclusions about the presence of functional relations and the magnitude of change. A functional relation can be identified when (a) there is a sufficient number of **potential demonstrations of effect** (i.e., at least three opportunities to demonstrate behavior change contingent on condition change), and (b) visual analysis suggests that consistent changes in data occur for all potential demonstrations (i.e., there are at least three actual **demonstrations of effect**), given that you have chosen a methodologically sound design (see Chapters 9–13) and threats to internal validity have been appropriately controlled for (see Chapter 1). Generality of findings are further enhanced when similar conditions generate similar effects across different researchers, programs, participants, behaviors, and conditions (replication; see Chapter 4).

Data patterns are compared across similar conditions to determine whether comparable conditions of an experiment have a similar effect on the dependent variable. Consistent data patterns across similar conditions are critical for establishing replicable, predictable patterns of behavior under specific conditions. Consistency with a previously predicted pattern of behavior across similar conditions increases the likelihood of identifying a functional relation; the greater the consistency, the more likely the data represent a functional relation.

The presence of a functional relation can be confirmed when (a) there is a successful attempt to replicate effects of a condition and (b) similar conditions generate similar levels and trends within (intra-participant replication) and across (inter-participant replication) participants in a study. Establishing a clear pattern of responding during similar conditions and showing consistent patterns of behavior change when conditions change increases confidence that the independent variable had an effect on the dependent variable(s). A minimum of three demonstrations of behavior change is required to establish experimental control.

Figure 8.2 shows three intra-participant replications of behavior change when intervention is introduced, withdrawn, and then re-introduced. Three inter-participant replications are required to establish experimental control in a multiple baseline across participants design. This is illustrated in Figure 8.9, which demonstrates replicated effects for three participants (i.e., Emma, Jack, & Liam).

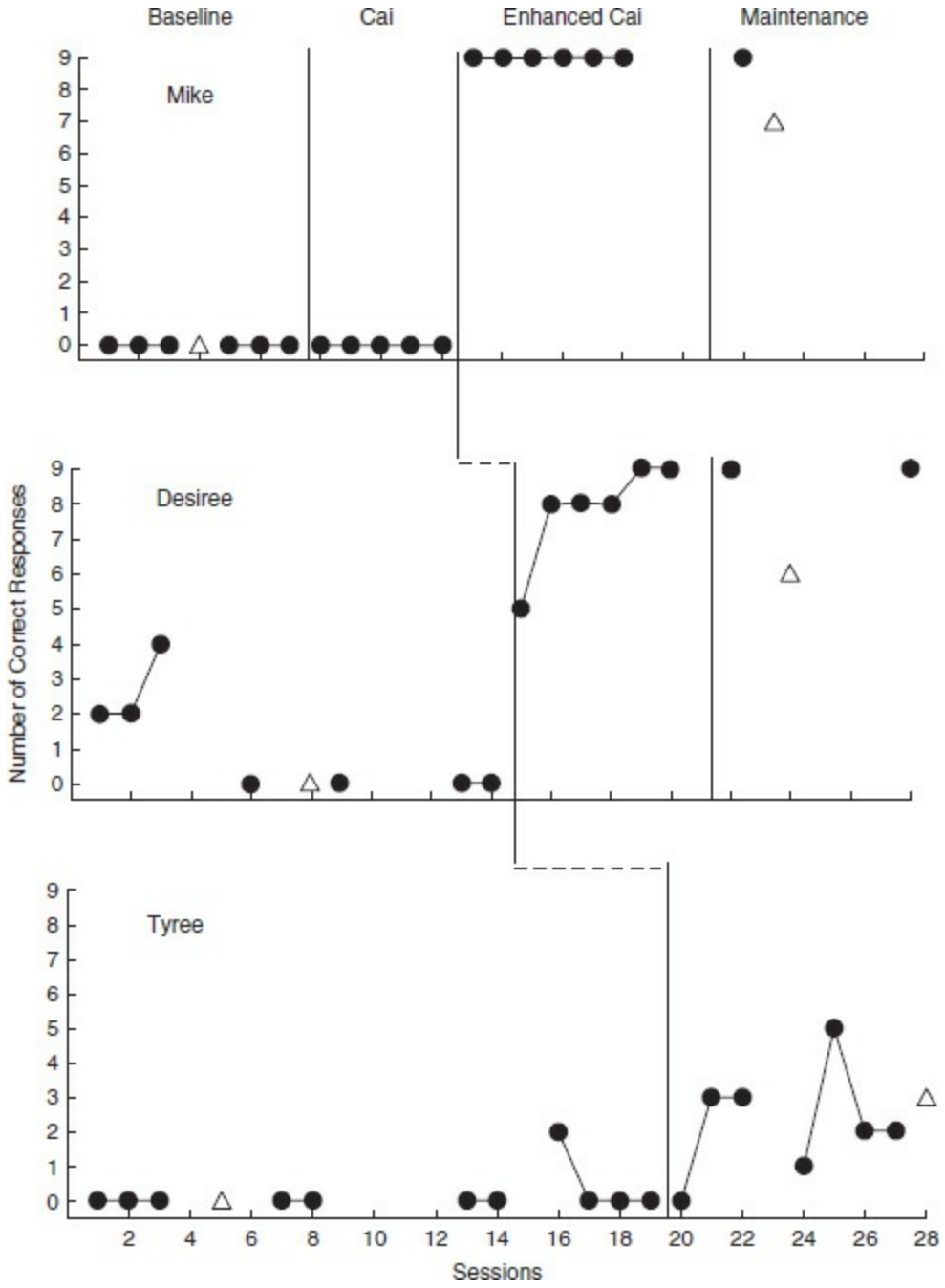


Figure 8.11 Representation of three potential demonstrations of effect, with two clear demonstrations of effect and one equivocal outcome.

Source: McKissick, B. R., Spooner, F., Wood, C. L., & Diegelmann, K. M. (2013). Effects of computer- assisted explicit instruction on map-reading skills for students with autism. *Research in Autism Spectrum Disorders*, *7*, 1653–1662.

In Figure 8.8, the percentage of intervals with MVPA showed stable levels at or below 11 percentage of intervals during baseline conditions. When the intervention was introduced, data immediately increased to a stable level at approximately 30 across both intervention conditions; data were predictable within and across similar conditions. This consistency within similar conditions is indicative of a functional

relation for Ramona. Data also should be consistent across similar conditions (e.g., across participants, behaviors, contexts) in time-lagged designs.

McKissick, Spooner, Wood, and Diegelman (2013) examined an "enhanced computer- assisted explicit instructional package" on correct responding of three elementary school students with ASD. As shown in Figure 8.11, there was a change in level and trend of large magnitude—for Mike and Desiree (i.e., two of three participants). However, the amount of overlap and variability across baseline and intervention conditions for Tyree precludes determination of functional relation.

Assessing Magnitude of Change

If a functional relation is present, the **magnitude**, or amount of behavior change may be of interest. After a functional relation is established, magnitude of the effect is assessed by comparing the amount and consistency of change across conditions and cases within a study that is directly attributed to the intervention. Smaller proportions of overlap are more likely to demonstrate functional relations and larger magnitudes of change. Likewise, immediate effects are more likely to demonstrate functional relations and larger magnitudes of change, although functional relations can be established when a gradual or small change is hypothesized (predicted) and consistent across tiers. Generally magnitude or the amount of consistent change is rated as small, medium, or large. Magnitude ratings should consider level, trend, and variability of behavior prior to introducing the independent variable and subsequent changes during intervention conditions. Because consistency of change is the primary factor when drawing conclusions regarding functional relations, studies including functional relations might include small, medium, or large magnitudes of effect. Thus, the magnitude of behavior change may be of interest for social validity evaluation (see Chapter 6), although it is generally not associated with internal validity.

The data in Figure 8.8 show clear, consistent behavior change at three different points in time indicating a functional relation (assuming threats to internal validity were minimized and data were collected using procedures that meet minimum design standards). However, the magnitude of change across conditions was small; authors may have hypothesized this magnitude of change given previous research on physical activity of young children and information about motor and physical abilities of young children with Down syndrome.

Conversely, data in Figure 8.3 show clear, consistent behavior change at three different points in time indicating a functional relation, but magnitude of change across conditions was large. Participants had low to no independent transitions during baseline and 100% independent transitions by the end of the intervention. Again, authors likely hypothesized a large level change given knowledge about

participants, dependent variables, and previous research.

As shown in Figure 8.12, Hemmeter, Snyder, Kinder, and Artman (2011) used a multiple baseline across participants design to examine use of performance feedback delivered via email on teachers' use of descriptive praise. There were four opportunities for behavior change with four potential inter-participant replications. There was an immediate, small change in level and no overlap in frequency of descriptive praise for Teachers A, B, and C when the intervention was introduced. Further, each of these teachers had stable data patterns within both baseline and intervention conditions. Teacher D had an immediate change with introduction of the intervention, but her use of descriptive praise decreased to baseline levels after three sessions. Given amount of overlap across conditions and high variability in the intervention condition, a clear determination of behavior change could not be made. Teacher D's levels increased when the intervention was adapted to include a criterion. Thus, despite having three inter-participant replications of behavior change at three different times, lack of behavior change for the fourth participant reduces confidence in presence of a functional relation.

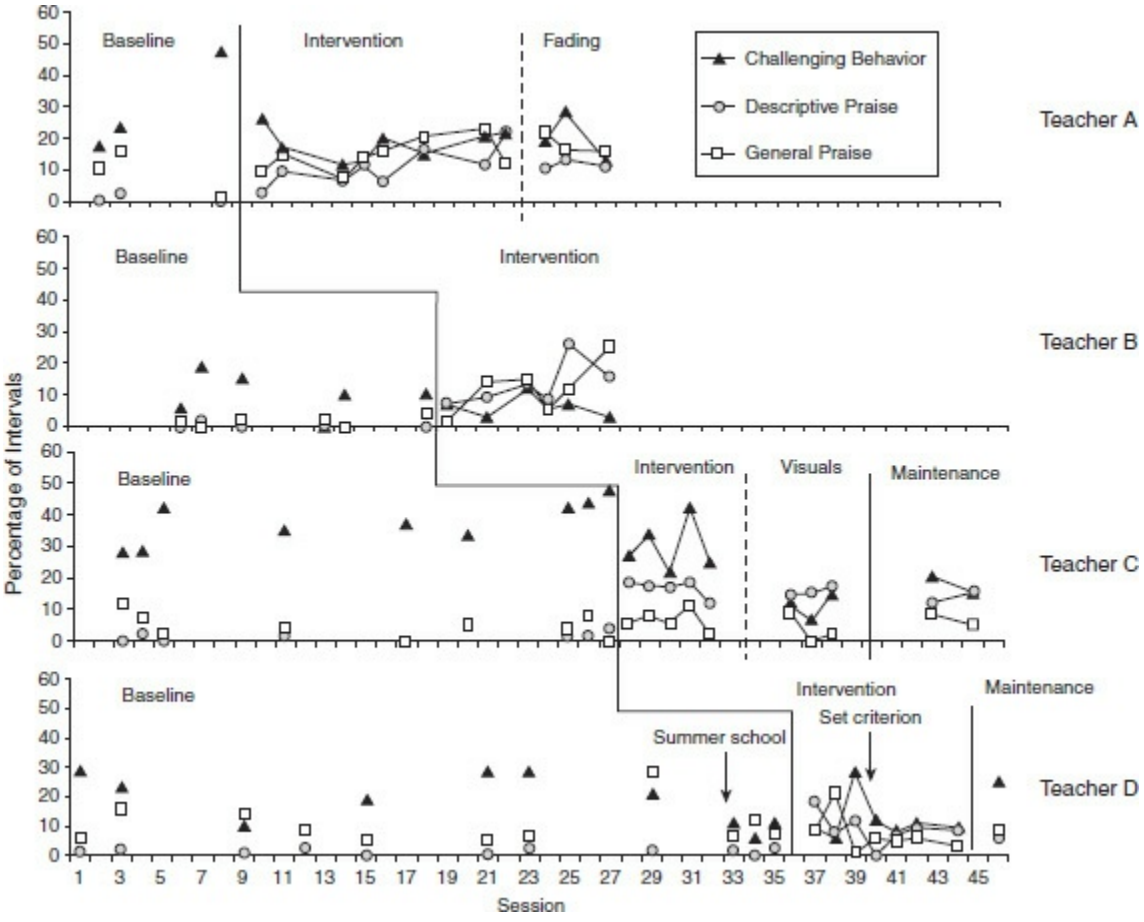


Figure 8.12 Representation of four potential demonstrations of effect, with three clear demonstrations of

effect and one effect after a criterion modification was implemented.

Source: Hemmeter, M. L., Snyder, P., Kinder, K., & Artman, K. (2011). Impact of performance feedback delivered via electronic mail on preschool teachers' use of descriptive praise. *Early Childhood Research Quarterly*, *26*, 96–109.

Summative evaluations also should compare opportunities for behavior change (potential demonstrations of effect) to occurrences of behavior change (actual demonstrations of effect). You might plan for a study to have more than three opportunities for behavior change to account for possible attrition or based on the phenomena being studied. For example, a multiple baseline across participants design with four participants has four opportunities for inter-participant replication. If there is clear behavior change and four inter-participant replications, experimental control is established and a functional relation can be identified. However, if there is clear behavior change and three inter-participant replications but no change for the fourth participant, confidence in presence of a functional relation is weakened. In this case, you can consider contextual or participant characteristics that might explain why behavior change did not occur for the fourth participant.

In sum, summative evaluations should consider each opportunity for *and* occurrence of adjacent condition behavior change. When visual analysis raises questions about experimental control, you must identify why the experimental demonstration was weakened or jeopardized. Through such post hoc analysis you will be able to redesign a study controlling for previously uncontrolled variables. In addition, these analyses provide an excellent source for identifying future research questions. This again illustrates the flexibility and usefulness of SCD research for identifying and improving interventions to ensure therapeutic effects.

Systematic Process for Conducting Visual Analysis

You should visually inspect graphs for the following: (1) Adequate number of data points within conditions to establish data patterns; (2) clear patterns within conditions in level, trend, or stability; (3) behavior change between adjacent conditions in level, trend, and/or variability; (4) degree of overlap and immediacy of change in data patterns across adjacent conditions; (5) consistency of changes across conditions and cases; (6) predicted patterns of change; and (7) magnitude of change across conditions and cases. A systematic process for conducting visual analysis is provided below and depicted in Figure 8.13.

1. Review the graph for equal and appropriate scaling of axes and to identify data series, conditions, representation of time, and unit of analysis (e.g., participant, behavior, context).

2. Examine research questions for the predicted pattern(s) of change in dependent variables.
3. Review number of data points per condition. Evaluate data stability within condition and determine if there is an adequate amount of data to establish a predictable pattern in each condition.
    a. If yes, proceed to Step 4.
    b. If no due to high variability in a baseline condition that is hypothesized to stabilize with introduction of the independent variable, proceed to Step 4.
    c. If no because variability precludes identifying a predictable pattern of behavior within more than one condition, or precludes evaluation of behavior change across adjacent conditions, discontinue visual analysis; experimental control cannot be established.
4. Analyze level, trend, and variability/stability of data in each condition. Determine if there are clear data patterns within all conditions.
    a. If yes, identify level change and stability, trend direction and stability, and amount of variability for each condition and move to Step 5.
    b. If no, discontinue visual analysis—experimental control cannot be established.
5. Analyze level, trend, and variability/stability of data across adjacent conditions. Determine if behavior change occurred across adjacent conditions. Using information from Step 4a, compare adjacent conditions for changes in magnitude of levels and stability, changes in trend direction and stability, and changes in variability or range of dependent measure values across adjacent conditions.
    a. If yes, behavior change occurred across adjacent conditions, proceed to Step 6.
    b. If no, discontinue visual analysis—experimental control cannot be established.
6. Analyze consistency of behavior change across conditions. For time-lagged and sequential introduction and withdrawal designs (see Chapters 9–10), experimental control is established and functional relations are identified when data patterns change with introduction of the independent variable at three different and temporally related time points. For designs that use rapid iterative alternation (see Chapter 11), it is established when data patterns in one condition are differentiated from other condition(s). Consider any anomalies or outliers in data. Determine if function relations exist.
    a. If yes, decide if data changes are consistent with predicted patterns.
    b. If yes, identify magnitude of change across conditions.
    c. If no, discontinue visual analysis—experimental control cannot be established.
7. Make a summative conclusion regarding experimental control and functional relation to answer your research questions.

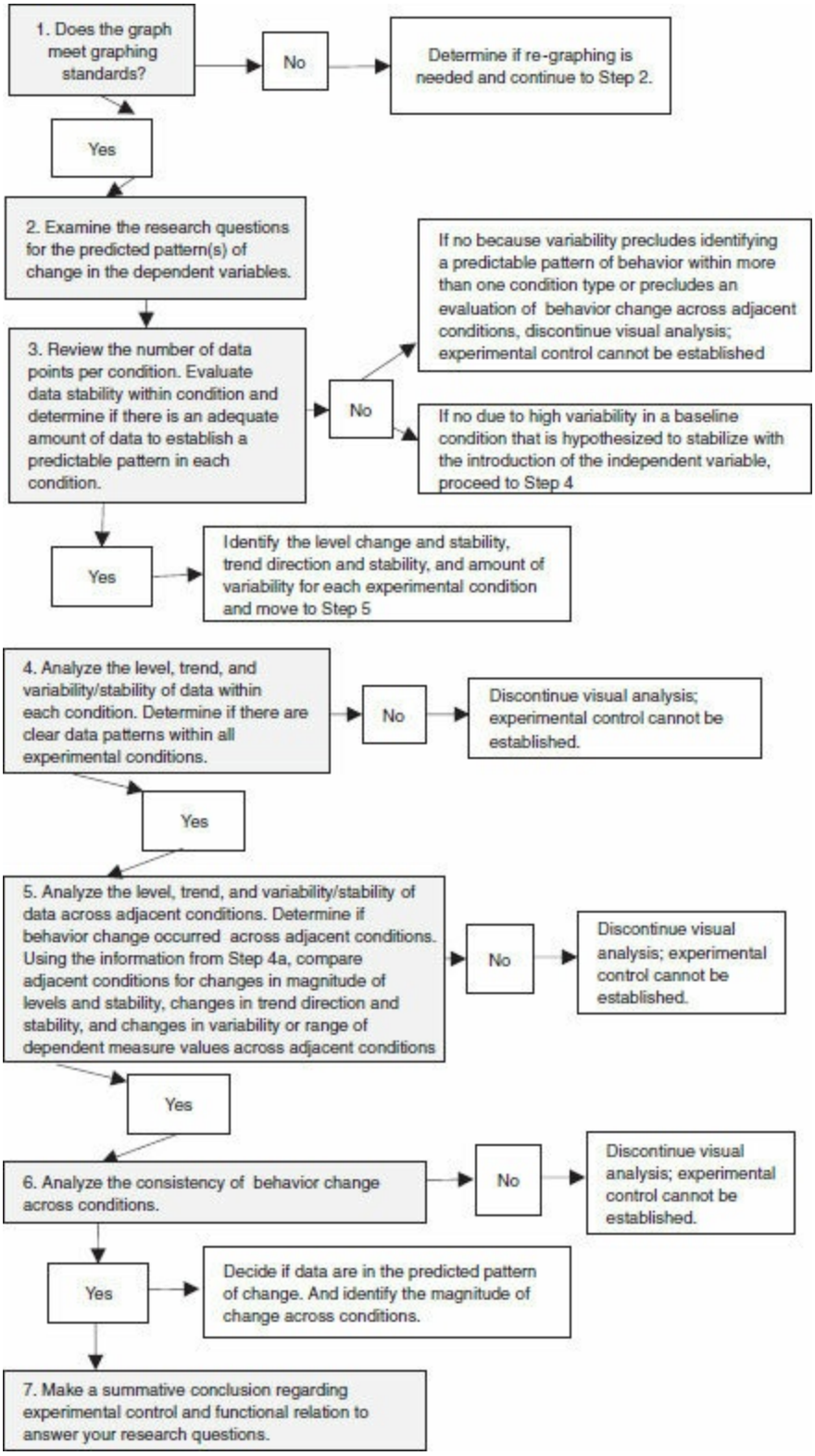


Figure 8.13 Depiction of the visual analysis process.

Planning and Reporting Visual Analyses

Like any quantitative analysis of experimental data, visual analysis requires a plan. While the specifics of the plan should be determined based on features of the research question and experimental design, the analysis plan should address several critical components. These components include (a) deciding how often data will be graphed, (b) considering how data will be graphically displayed, (c) determining which data characteristics will be the focus of within- and between-condition analyses, and (d) identifying design-related criteria that will impact visual analysis. In many cases, it is also important to identify a priori modifications in the event of unexpected data patterns (e.g., no behavior change following intervention). Below, we elaborate on each of these components of a visual analysis plan as they apply across design types. Design-specific guidelines are described in Chapters 9–12.

Determining a Schedule for Graphing Data

While the practice of graphing data applies across research methodologies, the timing and frequency with which data are graphed is unique in SCDs. In group experimental, quasi- experimental, and correlational designs, graphs may be prepared for descriptive purposes after all study data have been collected. Visual analysis of SCD data, in contrast, requires regular and frequent graphing of data *throughout* the study. For this reason, deciding how often data will be graphed is an important part of the planning process. When determining a schedule for graphing data, you should ensure data are graphed regularly enough to (a) inform decision- making with respect to implementing the design as planned and (b) identify relevant threats to internal validity that can be detected visually (e.g., history, maturation, testing). As a general rule, the more frequently we graph our data, the better positioned we are to formatively analyze the data. Whether it is necessary to graph data following each session or data collection opportunity, however, depends on the design-related criteria. If, in the context of a multiple baseline across participants design, the decision for when to change conditions from baseline to intervention is based on meeting a specific criterion for level stability, it may be necessary to graph data following each session. Conversely, alternating treatments designs commonly incorporate an element of randomization in sequencing conditions. This means that the sequence of sessions per condition series is determined randomly. Once determined, each series of sessions will be completed in the randomly selected order, regardless of data patterns observed. Thus, in this case, it would be sufficient to update graphs following one or more *series* of sessions. Even in cases for which experimental change decisions cannot be made, more frequent graphing allows you to detect threats to internal validity and take steps to control for them.

Considering Graphic Display

Other aspects of graphic display must also be considered to facilitate formative and summative visual analysis. As described in Chapter 7, the full range of the ordinate (*y*-axis) should be represented to evaluate level and variability within conditions, and changes in level and/or variability between conditions. It is also important to ensure that the abscissa (*x*-axis) accurately preserves time and sequence in which data are collected. This is particularly important for experimental designs in which conditions are introduced in a time-lagged fashion across tiers (i.e., multiple baseline and multiple probe designs) and those in which conditions are rapidly alternated from one session to the next (i.e., alternating treatments and adapted alternating treatments). Designs with time-lagged introduction of the intervention require concurrent measurement of target behaviors across participants, behaviors, or contexts; thus, it is critical for data to accurately depict relative sequence in which sessions are conducted across tiers (see Chapter 10). Comparison designs in which independent variables are rapidly alternated are prone to a special threat to internal validity known as sequence effects. Preserving sequence of rapidly alternating conditions helps to detect and address such threats. Regardless of design type, the *x*-axis should accurately reflect unexpected interruptions or extended breaks in data collection. That is, if sessions are completed on a daily basis for Sessions 1–10, but a participant was absent from school for a week between Sessions 10 and 11, the spacing between Sessions 10 and 11 should reflect this break on the graph.

Visual analysis can become more difficult as the number of data paths increases—especially when these data paths are overlapping. Thus, when multiple dependent measures are included in a study, the decision to plot them on the same graph should be made with caution. Seemingly minor formatting decisions can make a difference in these cases, such as selecting condition series labels that are visually distinct, and ensuring lines and condition symbols are fine and small enough to distinguish overlapping data paths. Additional dependent variables also may be graphed on secondary *y*-axes to minimize overlap among data paths. Even when multiple dependent variables are included in a study, one dependent variable must be selected to drive design-related decisions, and priority should be given to graphically display the primary dependent variable clearly and accurately. Limited publication space sometimes necessitates plotting more than one dependent variable on a graph, but formative analyses can be conducted with data on separate graphs to minimize complexity.

Finally, we recommend against overreliance on mean, median, or trend lines when visually analyzing SCD data. While such tools may be used as judgment aids, they only provide estimates of a single characteristic of the data, and have potential to distract from observed data patterns. When such tools are used as part of formative data analysis, we recommend using median-based over mean-based estimates, as the latter are unduly influenced by extreme or outlying data points. In addition, we recommend removing median or trend lines prior to conducting summative analysis and when disseminating data in presentations or publications. In our opinion, the addition of mean,

median, or trend lines to graphs are more likely to bias readers toward Type I errors (i.e., identifying presence of a functional relation when there is none).

Identifying Relevant Data Characteristics

Identifying which data characteristics will be the focus of within- and between-condition analyses is another important component of planning for visual analysis. While all data characteristics can be considered, identifying what types of behavior changes are expected can strengthen internal validity of a study. Such specification has been described in terms of making 'elaborate' predictions: "The more elaborate the prediction, the fewer alternative explanations are plausible when the data support the predictions" (Shadish, Hedges, Horner, & Odom, 2015, p. 19). With careful consideration of the research question and associated variables, we can hypothesize within-condition patterns and between-condition changes in behavior we expect to observe if the intervention works as expected. For example, when evaluating effects of an academic intervention on reading fluency, we might predict a gradual increase in trend following initiation of the intervention. Alternatively, when evaluating effects of a differential reinforcement procedure to decrease rates of disruptive behavior, we might predict an immediate and abrupt decrease in level of disruptive behavior following intervention. In other cases, the goal of an intervention may be to increase consistency or stability of a behavior. For example, variability in class attendance may be predicted to decrease from baseline to intervention. Finally, there may be interventions for which we expect change in behavior to happen following some delay. When we make such specific predictions, and our data support them, our confidence that observed changes are due to the intervention increases. It is worth noting, however, that unexpected data patterns do not necessarily prevent drawing conclusions. Rather, data patterns that are inconsistent with our predictions provide opportunities to reconsider how the intervention impacts behavior, and under what conditions.

Identifying Design-Related Criteria

Relevant data characteristics that will support your predictions should also inform another component of planning for visual analysis: identifying design-related criteria. Two such criteria are determining (a) the minimum number of sessions per condition and (b) explicit criteria for changing conditions. These decisions should be informed by existing SCD standards (e.g., CEC, 2014; Kratochwill et al., 2013) as well as your expected within- and between-condition data patterns. For example, the What Works Clearinghouse (WWC) standards require a minimum of five data points per condition to 'meet standards without reservations' (Kratochwill et al., 2013). However, there may be cases in which fewer than five data points are sufficient to establish a stable pattern. For example, suppose you are evaluating effects of a systematic prompting procedure to

teach 10 sight words to students with developmental disabilities. A baseline condition consisting of only three sessions may be sufficient if the percentage of correct responding is consistently at zero. If, on the other hand, the dependent variable is a percentage of time spent academically engaged—a readily reversible behavior that may vary from one session to another—the minimum number of sessions per condition may need to exceed five. As a general guideline, the more variable the data, the longer the condition should be. While more data will always be preferable for the purpose of evaluating experimental control, practical and ethical considerations must also be considered when determining condition length. When interventions target high-risk behaviors, such as physical aggression or self-injury, abbreviated baseline conditions may be warranted to minimize delays to initiating intervention—especially when baseline data demonstrate a clear need for intervention (e.g., high rates or contra-therapeutic trends). Selecting criteria for changing conditions should also depend on the research question and independent and dependent variables. In general, studies focusing on acquisition of non-reversible behaviors lend themselves to absolute criteria for changing conditions (e.g., three consecutive sessions exceeding 90% correct responses), whereas studies targeting reversible behaviors tend to require relative criteria for changing conditions (e.g., four of five consecutive sessions below median baseline responding).

Reporting Visual Analyses

Each critical component of the visual analysis plan should be made transparent. Summative analysis should be described using visual analysis terminology and in a way that matches the logic of the experimental design used. You should avoid reporting means per condition, as condition means are not the basis on which conclusions of functional relations are drawn. Rather, summative analysis should focus on the extent to which (a) within-condition data patterns were stable, (b) hypothesized between-condition shifts in data patterns were detected, and (c) these shifts consistently co-occurred with each change in condition.

Visual Analysis Applications

In the following section, we present summative visual analyses based on published examples of SCD studies. Each example reflects a different method of ordering conditions (i.e., design type): sequential introduction and withdrawal of the independent variable (A-B-A-B or withdrawal design); time-lagged introduction of the independent variable across tiers (multiple baseline design); and rapid alternation of the independent variable across sessions (alternating treatments design). These examples were selected based on their clear graphic depiction of the experimental design and evidence supporting conclusions of functional relations.

Summative Visual Analysis Application: A-B-A-B Withdrawal Design

Wills, H. P., & Mason, B. A. (2014). Implementation of a self-monitoring application to improve on-task behavior: A high-school pilot study. *Journal of Behavioral Education*, *23*, 421–434.

The graph in Figure 8.5 depicts results of a study evaluating effects of a self-monitoring intervention on percentage of time on-task for two high school students receiving special education services. An A-B-A-B withdrawal design was used in which the self-monitoring intervention was sequentially introduced and withdrawn to provide three opportunities for demonstrating an effect. Researchers selected five as the minimum number of sessions per condition, with additional sessions conducted in the presence of trends or variability in on-task behavior. Results in Figure 8.5b show that with the exception of the first baseline data point, initial baseline percentages of on-task behavior showed high **level stability**, ranging from 41–51%. When the self-monitoring intervention was introduced, an **immediate** increase in **level** of on-task behavior was observed, with no **overlapping** data points with the initial baseline condition and high **level stability** with five of seven sessions at or approaching 100% on-task behaviors. When the intervention was withdrawn, an **immediate** decrease in **level** of on-task behavior was observed, with no **overlapping**data points with the previous self-monitoring condition and with ranges approximating levels observed in the initial baseline condition (32–51%; i.e., **consistency** across similar conditions). When the self-monitoring intervention was re-introduced, **levels** of on-task behavior **immediately** increased, with no **overlapping** data points with the previous baseline condition. **Immediate** and abrupt changes in level that **consistently** co-occurred with changes in condition, and lack of **overlap** between adjacent conditions, support a conclusion of a functional relation between the intervention and increases in on-task behavior.

Summative Visual Analysis Application: Multiple Baseline Across Participants

Lambert, J. M., Bloom, S. E., & Irvin, J. (2012). Trial-based functional analysis and functional communication training in an early childhood setting. *Journal of Applied Behavior Analysis*, *45*, 579–584.

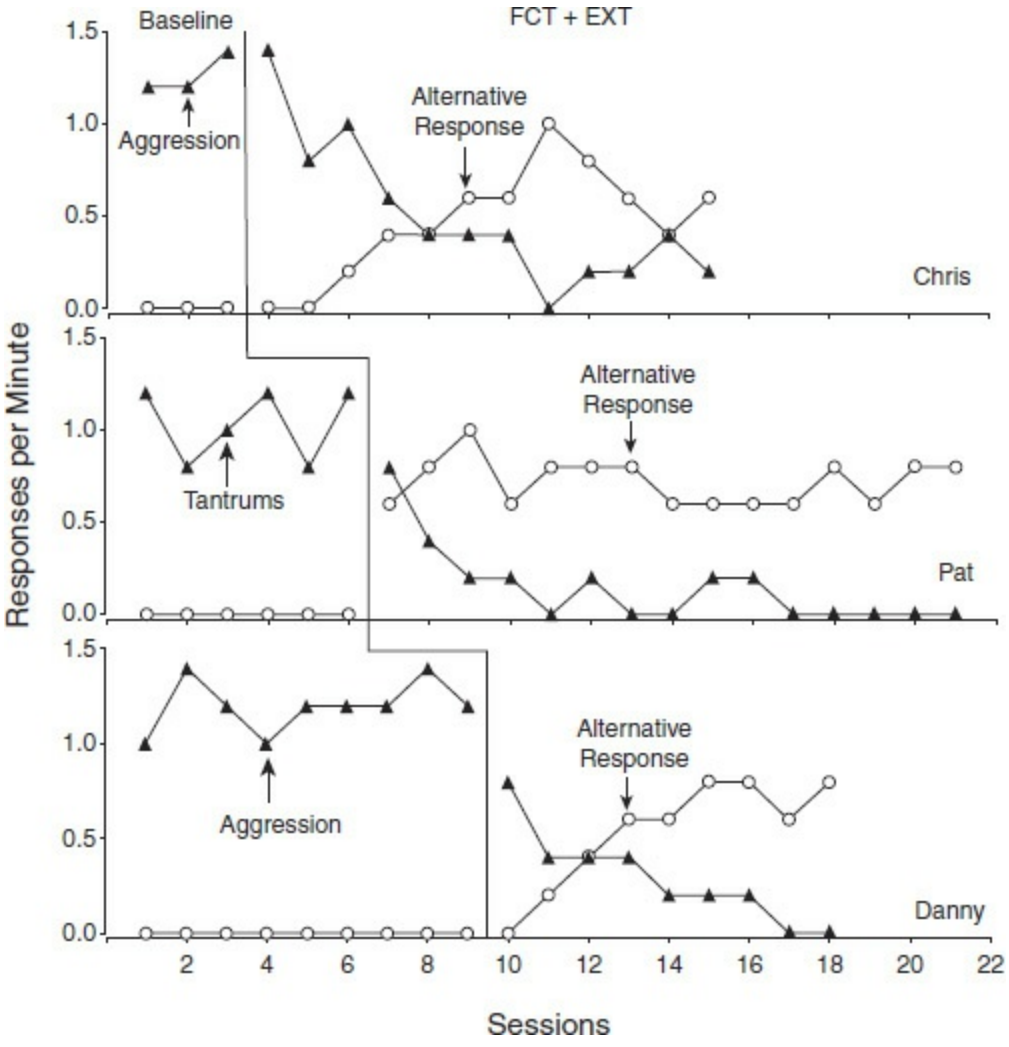


Figure 8.14 Visual analysis application for multiple baseline across participants design.

Source: Lambert, J. M., Bloom, S. E., & Irvin, J. (2012). Trial-based functional analysis and functional communication training in an early childhood setting. *Journal of Applied Behavior Analysis*, *45*, 579–584.

The graph in Figure 8.14 is from a study evaluating effects of functional communication training (FCT) with extinction on rates of problem behavior and alternative communication for three young children with developmental delays. A multiple baseline design was used in which an FCT + Extinction intervention was introduced in a time-lagged fashion across three participants. Each data path

represents a distinct target behavior, with problem behavior depicted by closed triangles and alternative communication responses depicted by open circles. For all three participants, baseline **levels** of problem behavior were relatively high (ranging from approximately 0.75 to 1.5 per minute). When the intervention was introduced to each participant, **immediate** changes in **trend** were observed, with **levels** of problem behavior decreasing to less than 0.5 per minute for one participant, and to zero rates for two participants. These changes in **trend** and **level** occurred only when the intervention was introduced to each participant, and at no other time. **Levels** of alternative communication were **stable** and at zero during baseline sessions for all participants. Following introduction of the FCT + Extinction intervention, **accelerating trends** occurred within 1–3 intervention sessions across all three participants. For one participant (Pat) there was an **immediate** increase in **level** in addition to an initial increasing **trend**. Three demonstrations of (a) decreases in problem behavior and (b) increases in alternative communication at three points in time (i.e., when the intervention was introduced to each participant) support a conclusion of a functional relation between FCT + Extinction and both target behaviors across three participants.

Summative Visual Analysis Application: Alternating Treatments Design

Rispoli, M., O'Reilly, M., Lang, R., Machalicek, W., Davis, T., Lancioni, G., & Sigafoos, J. (2011). Effects of motivating operations on problem and academic behavior in classrooms. *Journal of Applied Behavior Analysis*, *44*, 187–192.

The graph in Figure 8.15 shows results from a study evaluating effects of pre-session access to preferred items on levels of academic engagement during subsequent instructional sessions. An alternating treatments design was used in which sessions with and without pre-session access to tangibles were rapidly alternated across school days. In this graph, each data path represents a different condition; both reflect the same dependent variable (percentage of intervals with academic engagement). **Levels** of academic engagement were higher in the *pre-session access* condition relative to the *no pre-session access* condition. Importantly, this **differentiation** in responding was **consistent** across sessions, producing five demonstrations of effect (i.e., one demonstration for each condition pair). This **consistency** in **level change** between conditions, as well as relative **level stability** within conditions, supports a conclusion of a functional relation between pre-session access to preferred items and percentage of intervals with academic engagement for this participant.

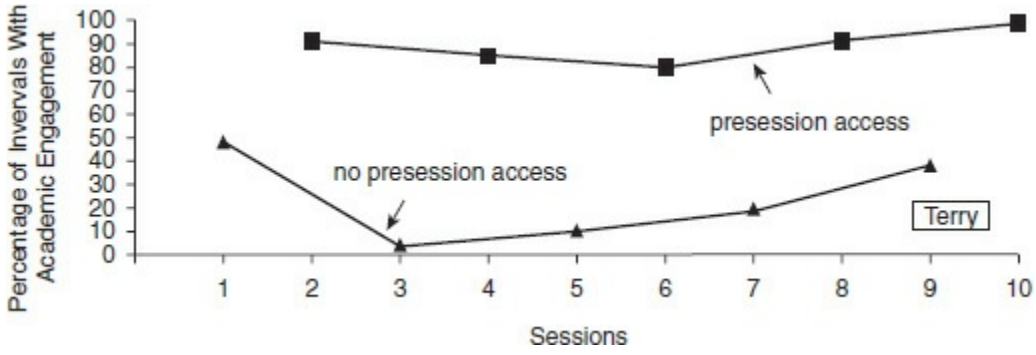


Figure 8.15 Visual analysis application for alternating treatments design.

Source: Rispoli, M., O'Reilly, M., Lang, R., Machalicek, W., Davis, T., Lancioni, G., & Sigafoos, J. (2011). Effects of motivating operations on problem and academic behavior in classrooms. *Journal of Applied Behavior Analysis*, *44*, 187–192.

Visual Analysis Tools

In this section, we describe three tools that may be used as judgment aids for formative and summative visual analysis. We review the split middle method to estimate trend, stability envelopes to estimate variability, and percentage of non-overlapping data to estimate overlap. Rather than presenting an exhaustive summary of existing visual analysis tools, we focus on these three based on their relative widespread use and because each tool addresses a distinct characteristic of data. We describe steps to apply each tool, identify conditions in which each tool may be most informative, and caution against overreliance on any single tool, as people, not judgment aids, make decisions about functional relations.

Split Middle Method to Estimate Trend

The split middle method (White & Haring, 1980) is a tool that can be used to estimate trend within conditions and compare trends between conditions. Steps to use the split middle method are as follows (depicted in Figures 8–16, 8–17, and 8–18):

1. Within each condition, draw a vertical line that divides the number of data points in half. If the total number of data points is an even number, the vertical line will cross the data path between two data points (see baseline condition of Figure 8.16); if the total number of data points is an odd number, the vertical line will cross through a data point (see DRO condition of Figure 8.16).
2. Within each half of the condition, draw another vertical line that divides that number of data points in half. Then, draw a horizontal line at the median value for each half such that it intersects the vertical line (see Figure 8.17).
3. Within each condition, draw a line through the points at which the vertical and horizontal lines from Step 2 intersect (see Figure 8.18).
4. Adjust the line drawn in Step 3 such that there are an equal number of data points above and below the line. This adjustment may not be necessary if the number of data points above and below the line is already equal. This line is the split-middle trend line.

Split middle trend lines are most useful when within condition trends or between condition changes in trend are of primary interest and data show moderate or high variability within conditions. In addition, while this method requires a minimum of four data points, the accuracy of the split middle method to estimate trend increases with the number of data points in each condition. When using trend lines as judgment aids, you should be cautious in the degree to which these impact decisions about functional relations. As mentioned earlier, to minimize potential for Type I error bias, we

recommend removing trend lines for summative analysis and prior to disseminating graphed data for independent analysis.

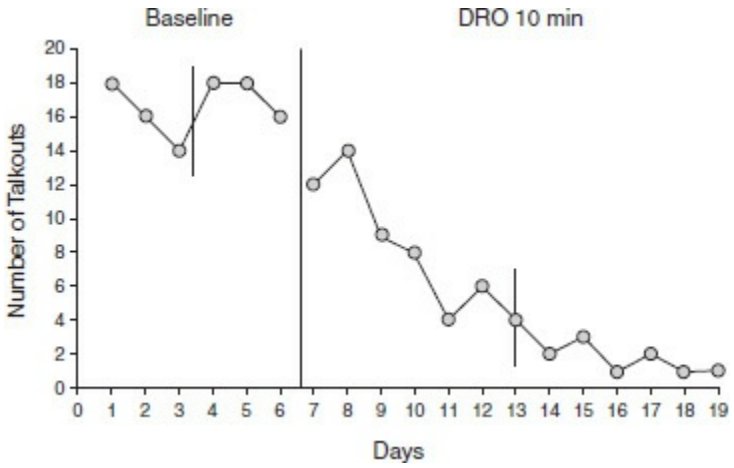


Figure 8.16 Depiction of calculation of split middle. Step 1: Within each condition, draw a vertical line to divide the number of data points in half.

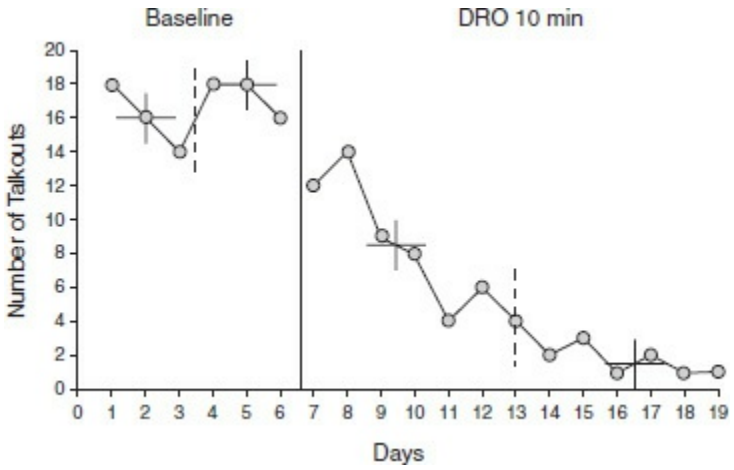


Figure 8.17 Depiction of calculation of split middle. Step 2: Within each half of each condition, draw another vertical line to divide the number of data points in half. Then draw a horizontal line at the median value such that it intersects the vertical line.

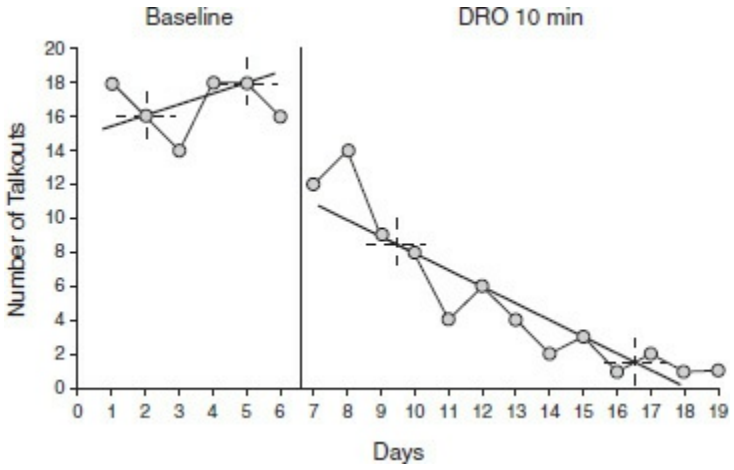


Figure 8.18 Depiction of calculation of split middle. Step 3: Within each condition, draw a line through the points of intersection from Step 2. When necessary, adjust each line such that there is an equal number of data points above and below it.

Stability Envelopes to Estimate Level or Trend Stability

Stability envelopes can be used to estimate stability in level or trend within conditions. The primary advantage of using stability envelopes is to ensure consistency in experimental decisions related to data stability. They consist of two parallel lines that are drawn on either side of a median or trend line. Though stability envelopes also may be drawn around mean lines, we recommend basing them on median values, which are less influenced by extreme data values. Steps to draw a stability envelope around a median line are as follows (depicted in Figure 8.19):

1. Calculate the median level of all data point values in a condition. The median level of a data series is the middle data point value if all values are ordered from low to high. If the number of data points is even, the median is the average of the two middle values.
2. Draw a median line parallel to the abscissa that intersects the median value (see solid median line in Figure 8.19).
3. Select a percentage used to determine level stability (e.g., 30%) and multiply this value by the median value. The product represents the size of the stability envelope.
4. Draw two parallel lines above and below the median line to form a stability envelope; the distance between these lines must match the product calculated in Step 3. Adjust the envelope up or down to capture as many data points as possible (see dashed lines in Figure 8.19).
5. Calculate the percentage of data points falling within the stability envelope and compare it to the stability criterion to make experimental decisions.

The percentage selected for stability envelopes may depend on factors such as number of opportunities to respond, or whether the behavior of interest is trial-based or free operant. Generally, larger stability envelopes may be used for free operant behaviors than trial-based responding, and for trial-based responding when the number of opportunities is few. Stability envelopes should be calculated only once for a behavior and placed over the median or trend line of the original condition and all other conditions introduced to that behavior. If the behavior does not occur during an initial baseline condition, stability envelopes may be calculated based on the median value of the first intervention condition. Stability envelopes also may be used to evaluate trend stability using the same steps above, but the lines forming the envelope are drawn such that they are parallel to a trend line rather than a median line. While lines must remain parallel to the trend line, the envelope may be adjusted up or down to capture as many

data points as possible. Additional research is needed to determine appropriate use of stability envelopes for characterizing single case data.

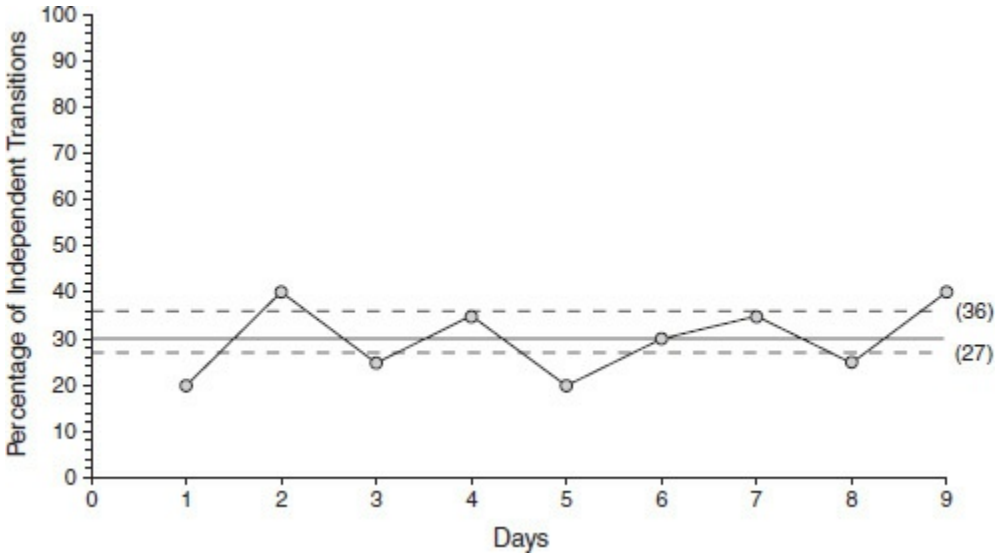


Figure 8.19 Illustration of steps to calculate percentage of non-overlapping data when levels of target behavior are expected to increase following the condition change (top graph) and when levels of target behavior are expected to decrease following the condition change (bottom graph).

Percentage of Non-Overlapping Data to Estimate Between-Condition Level Change

The percentage of non-overlapping data (PND; Scruggs & Mastropieri, 1998) may be used to estimate level change between two adjacent conditions. Steps to calculate PND are outlined below (depicted in Figure 8.20):

1. Determine the range of data point values in the first condition (Condition A).
2. Count the number of data points in the second condition (Condition B) that fall outside of this range, in the predicted direction.
3. Divide the number of Condition B data points that fall outside the range of Condition A by the total number of data points in Condition B.
4. Multiply the quotient by 100 to yield a percentage.

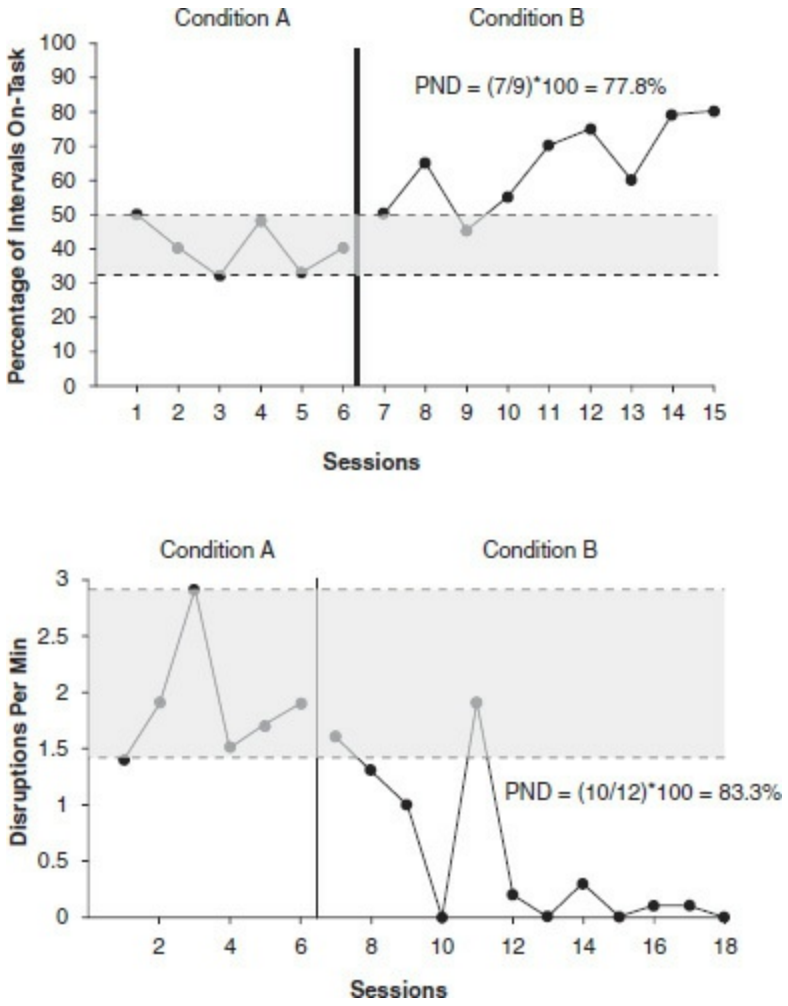


Figure 8.20 Illustration of steps to calculate percentage of non-overlapping data when levels of target behavior are expected to increase following the condition change (top graph) and when levels of target behavior are expected to decrease following the condition change (bottom graph).

The higher the PND, the more consistent and abrupt the level change between adjacent conditions (note this is unrelated to the *size* of the level change). A PND of 100% indicates no overlap in the ranges of values between two adjacent conditions. While calculating PND can be useful when differences in level between conditions are of primary interest, it should not be used in isolation to determine between-condition behavior change. There are several scenarios in which relying on PND alone can lead to incorrect conclusions, including when accelerating or decelerating trends are present in one or more conditions. For example, PND can be compromised when one or more baseline data points reach a therapeutic floor or ceiling (Figure 8.21a) or when baseline data points are highly variable (Figure 8.21b). Additionally, PND can indicate no behavior change (0%) when there is a clear change in trend direction between conditions (Figure 8.21c). Or, PND can indicate behavior change (100%) when there is a consistent accelerating or decelerating trend across conditions (Figure 8.21d). Because PND is also affected by the number of data points in the intervention condition (Figure 8.21e), it will be a more interpretable estimate as the number of data points increases. A final

cautionary note related to PND is that it does not reflect magnitude of behavior change between conditions (Figure 8.21f). PND would be the same regardless of whether the number of words spelled correctly increased from 0 to 1 from baseline to intervention or from 0 to 100 from baseline to intervention (i.e., 100%). It only reflects the overlap between conditions, and should only be used for this specific purpose.

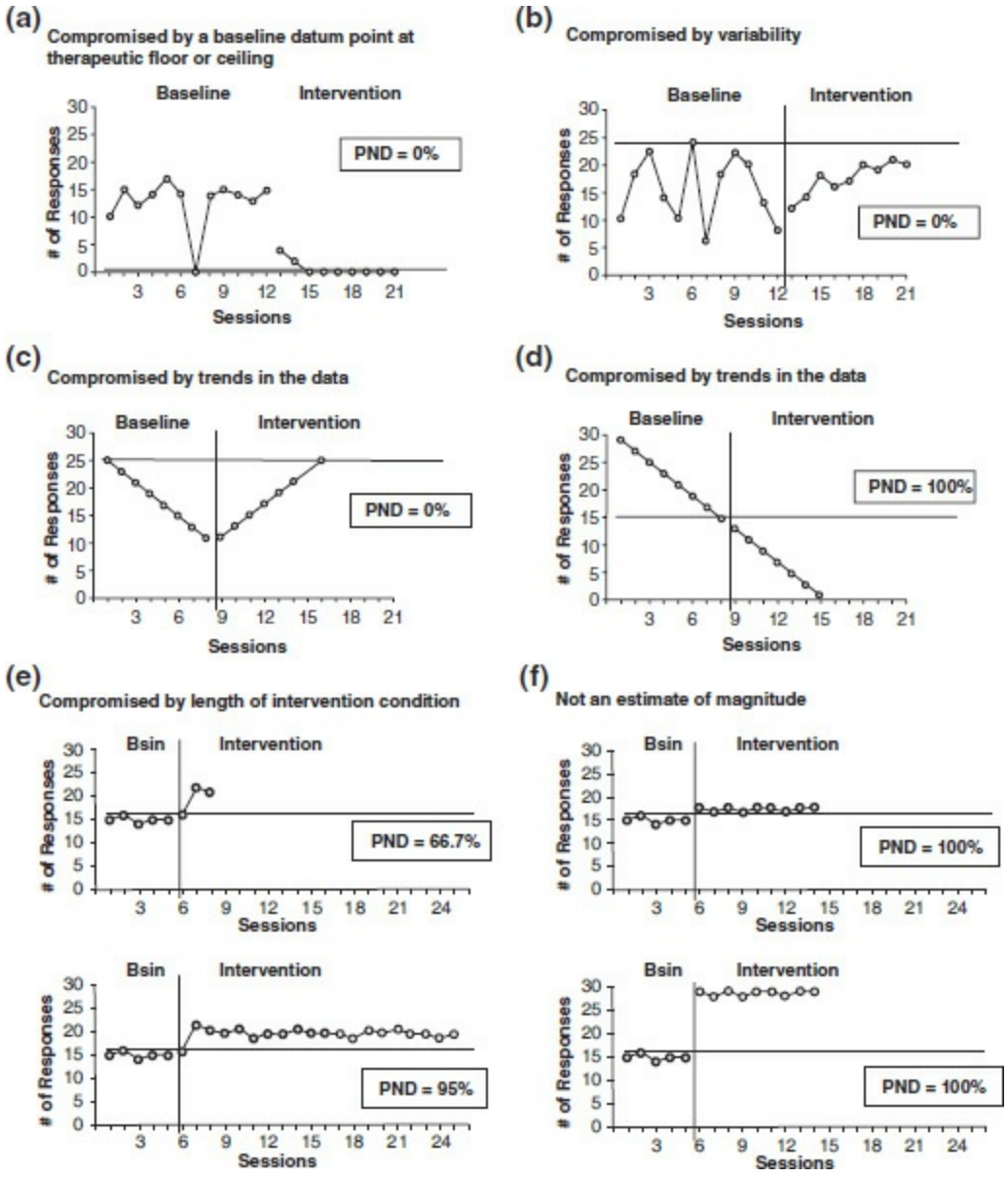


Figure 8.21 Problems associated with percentage of non-overlapping data (PND).

Source: M. Wolery, personal communication, January 15, 2008

Visual Analysis Protocols

Causal inferences regarding the influence of the independent variable on the dependent variable should only be made when data analyses were completed systematically and objectively and when the method of analysis produces reliable and consistent results when conducted by different individuals (Kratochwill & Levin, 2014; Shadish, Cook, & Campbell, 2002). There has been increased criticism of visual analysis including concerns that procedures are not standardized and can lead to subjective judgments about behavior change and magnitude of effects, which could lead to disagreements about functional relations (Kazdin, 2011; Lieberman et al., 2010). Recent data suggest SCD researchers do not consistently report visual analysis procedures or use standard visual analysis terms to describe results (Barton, Fettig, & Meadan, 2017). Inconsistent results across different visual analysts can impact credibility of individual studies and credibility and usefulness of SCD research. Validated tools and standardized protocols for conducting visual analyses might minimize disagreements regarding functional relation.

Researchers have argued for creation and use of formal guidelines to operationalize visual analysis processes (Furlong & Wampold, 1982; Kazdin, 1982) and much effort has been placed in creating trainings and protocols to enhance reliability of visual analysis (e.g., Swoboda, Kratochwill, Horner, Levin, & Albin, 2012; Wolfe & Slocum, 2015). For example, the WWC developed evidence criteria to use with their design standards; they describe six features of SCD data and outline four steps for data analysis. Maggin, Briesch, and Chafouleas (2013) developed a protocol for visual analysis by adapting the WWC evidence criteria (Kratochwill et al., 2013). Their protocol guides visual analysts through within condition analysis, between conditions analysis, identification of functional relations, and strength of experimental control (Maggin et al., 2013). This protocol has been reliably used in systematic reviews of SCD research (Qi, Barton, Collier, Lin, & Montoya, 2017), but has yet to be validated. Wolfe, Barton, and Meadan (2017) developed a systematic protocol that walks visual analysts through a series of questions regarding within condition data patterns and data patterns across contrasting adjacent conditions. Aggregated responses to questions results in a total score for each study. Fisher, Kelley, and Lomas (2003) introduced a method for improving visual analysis of graphed data called the conservative dual criterion (CDC) method. The CDC method, as proposed by Fisher and colleagues (2003), provides guidelines for evaluating changes across conditions while considering various features of graphed data and was developed to improve descriptive quantitative methods such as the split-middle technique (White & Haring, 1980). The CDC evaluates and blends multiple sources of data—consistent changes in level and trend across conditions—by setting criterion lines based on trends and mean lines and hypothesized direction of change. The number of data points above and below the dual criterion lines is used to make conclusions about

overall systematic changes, which can be used to inform decisions about functional relation. Whether using these protocols or a researcher-developed one, you should have an established plan for visual analysis, and you should use that plan systematically.

Summary

As Kennedy (2005) described, SCD research can be compared to a chess match, in which your next move is determined in part by what the data say. Visual analysis of graphic data is the process by which these formative analyses are conducted. Visual analysis of graphic data is also critical to summatively determining whether or not behavior changes that occurred during the study are attributable to condition changes (i.e., whether a functional relation was demonstrated) and if so, how large those changes are.

References

Adamo, E. K., Wu, J., Wolery, M., Hemmeter, M. L., Ledford, J. R., & Barton, E. E. (2015). Using video modeling, prompting, and behavior-specific praise to increase moderate-to-vigorous physical activity for young children with Down syndrome. *Journal of Early Intervention*, *37*, 270–285.

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.

Barton, E. E., Fettig, A., & Meadan, H. (2017). Comparison of visual analysis and non-overlap methods in the evaluation of parent implemented functional assessment based interventions. *Manuscript Under Review*.

Barton, E. E., Fuller, E. A., & Schnitz, A. (2016). The use of email to coach preservice early childhood teachers. *Topics in Early Childhood Special Education*, *36*, 78–90.

Barton, E. E., Ledford, J. R., Lane, J. D., Decker, J., Germansky, S. E., Hemmeter, M. L., & Kaiser, A. (2016). The iterative use of single case research designs to advance the science of EI/ECSE. *Topics in Early Childhood Special Education*, *36*, 4–14. Cihak, D., Fahrenkrog, C., Ayres, K. M., & Smith, C. (2010). The use of video modeling via a video iPod and a system of least prompts to improve transitional behaviors for students with autism spectrum disorders in the general education classroom. *Journal of Positive Behavior Interventions*, *12*, 103–115.

Council for Exceptional Children (2014). *Standards for evidence based-practices in special education*. Arlington, VA: Author.

Dunlap, G., DePerczel, M., Clarke, S., Wilson, D., Wright, S., White, R., & Gomez, A. (1994). Choice making to promote adaptive behavior for students with emotional and behavioral challenges. *Journal of Applied Behavior Analysis*, *27*, 505–518.

Fettig, A., Schultz, T. R., & Sreckovic, M. A. (2015). Effects of coaching on the implementation of functional assessment—based parent intervention in reducing challenging behaviors. *Journal of Positive Behavior Interventions*, *17*, 170–180.

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

Franklin, R. D., Gorman, B. S., Beasley, T. M., & Allison, D. B. (1996). Graphical display and visual analysis. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single case research*. Mahwah, NJ: Lawrence-Erlbaum.

Furlong, M. J., & Wampold, B. E. (1982). Intervention effects and relative variation as dimensions in experts' use of visual inference. *Journal of Applied Behavior Analysis*, *15*, 415–421.

Gast, D. L., & Spriggs, A. D. (2014). Visual analysis of graphic data. In D. L. Gast & J. R. Ledford (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (2nd ed., pp. 176–210). New York, NY: Routledge.

Hemmeter, M. L., Snyder, P., Kinder, K., & Artman, K. (2011). Impact of performance feedback delivered via electronic mail on preschool teachers' use of descriptive praise. *Early Childhood Research Quarterly*, *26*, 96–109.

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

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

Horner, R., & Spaulding, S. (2010). Single-case research designs. In N. J. Salkind (Ed.), *Encyclopedia of research design* (pp. 1386–1394). Thousand Oaks, CA: Sage.

Hughes, M. A., Alberto, P. A., & Fredrick, L. L. (2006). Self-operated auditory prompting systems as a function- based intervention in public community settings. *Journal of Positive Behavior Interventions*, *8*, 230–243.

Jones, J., Lerman, D. C., & Lechago, S. (2014). Assessing stimulus control and promoting generalization via video modeling when teaching social responses to children with autism. *Journal of Applied Behavior Analysis*, *47*, 37–50.

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

Kazdin, A. E. (2011). *Single case research designs: Methods for clinical and applied settings*. (2nd ed.). New York, NY: Oxford.

Kennedy, C. H. (2005). *Single case designs for educational research*. Boston, MA: Allyn & Bacon.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, *34*, 26–38.

Kratochwill, T. R., & Levin, J. R. (2014). *Single-case intervention research: Methodological and statistical advances*. Washington, DC: American Psychological Association. Lambert, J. M., Bloom, S. E., & Irvin, J. (2012). Trial- based functional analysis and functional communication training in an early childhood setting. *Journal of Applied Behavior Analysis*, *45*, 579–584.

Lambert, J. M., Bloom, S. E., & Irvin, J. (2012). Trial-based functional analysis and functional communication training in an early childhood setting. *Journal of Applied Behavior Analysis*, *45*, 579–584.

Lane, J. D., & Gast, D. L. (2013). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation*, *24*, 445–463.

Lieberman, R. G., Yoder, P. J., Reichow, B., & Wolery, M. (2010). Visual analysis of multiple baseline across participants graphs when change is delayed. *School Psychology Quarterly*, *25*, 28–44.

Maggin, D. M., Briesch, A. M., & Chafouleas, S. M. (2013). An application of the what works clearinghouse standards for evaluating single-subject research: Synthesis of the self-management literature base. *Remedial and Special Education*, *34*, 44–58.

McKissick, B. R., Spooner, F., Wood, C. L., & Diegelmann, K. M. (2013). Effects of

computer-assisted explicit instruction on map-reading skills for students with autism. *Research in Autism Spectrum Disorders*, *7*, 1653–1662.

Parsonson, B. S., & Baer, D. M. (1978). The analysis and presentation of graphic data. In T. R. Kratochwill (Ed.), *Single subject research: Strategies for evaluating change*. New York, NY: Academic Press.

Parsonson, B. S., & Baer, D. M. (1992). The visual analysis of data, and current research into the stimuli controlling it. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 15–40). Hillsdale, NJ: Lawrence Erlbaum.

Plavnick, J. B., MacFarland, M. C., & Ferreri, S. J. (2015). Variability in the effectiveness of a video modeling intervention package for children with autism. *Journal of Positive Behavior Interventions*, *17*, 105–115.

Qi, C. H., Barton, E. E., Collier, M., Lin, Y. L., & Montoya, C. (2017). A systematic review of effects of social stories interventions for individuals with autism spectrum disorder. *Focus on Autism and Other Developmental Disabilities*. doi: 10.1177/1088357615613516

Rispoli, M., O'Reilly, M., Lang, R., Machalicek, W., Davis, T., Lancioni, G., & Sigafoos, J. (2011). Effects of motivating operations on problem and academic behavior in classrooms. *Journal of Applied Behavior Analysis*, *44*, 187–192.

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification*, *22*, 221–242.

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

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). The *role of between-case effect size in conducting, interpreting, and summarizing single-case research* (NCER 2015–2002) Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

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

Skinner, B. F. (1957). *Verbal behavior*. New York, NY: Appleton-Century-Crofts.

Swoboda, C., Kratochwill, T., Horner, R., Levin, J., & Albin, R. (2012). *Visual analysis training protocol: Applications with the alternating treatment, multiple baseline, and ABAB designs*. Eugene, OR: University of Oregon. Retrieved from www.singlecase.org

What Works Clearinghouse. (2014). *Procedures and standards handbook* (Version 3.0). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v3_0_standards_handbook.pdf

White, O. R., & Haring, N. G. (1980). *Exceptional teaching* (2nd ed.). Columbus, OH: Charles E. Merrill.

Wills, H. P., & Mason, B. A. (2014). Implementation of a self-monitoring application to improve on-task behavior: A high-school pilot study. *Journal of Behavioral*

Education, *23*, 421–434.

Wolery, M. (2013). A commentary: Single-case design technical document of the what works clearinghouse. *Remedial and Special Education*, *34*, 39–43.

Wolery, M., & Harris, S. R. (1982). Interpreting results of single-subject research designs. *Physical Therapy*, *62*, 445–452.

Wolfe, K., Barton, E. E., & Meadan, H. (2017). *Systematic visual analysis protocol*. Unpublished protocol, University of South Carolina, Charleston, SC.

Wolfe, K., Seaman, M. A., & Drasgow, E. (2016). Interrater agreement on the visual analysis of individual tiers and functional relations in multiple baseline designs. *Behavior Modification*, *40*, 852–873.

Wolfe, K., & Slocum, T. A. (2015). A comparison of two approaches to training visual analysis of AB graphs. *Journal of Applied Behavior Analysis*, *48*, 472–477.

9
Withdrawal and Reversal Designs

David L. Gast, Jennifer R. Ledford, and Katherine E. Severini

Important Terms

single case experimental designs, baseline logic, withdrawal design, experimental, corollary behaviors, multitreatment design, reversal design

Baseline logic serves as the foundation for all single case design (SCD) research. That is, all SCDs are mere extensions or elaborations of the basic A-B paradigm, wherein behavior is measured repeatedly across two adjacent conditions: baseline (A) and intervention (B). In this chapter we describe those SCDs commonly referred to as "withdrawal" or "reversal" designs—one of the earliest and simplest SCDs, they involve repeating this basic A-B comparison by introducing and withdrawing the intervention with one or more participants. Historically these designs have been referred to as simple and repeated time series designs (Birnbrauer, Peterson, & Solnick, 1974; Campbell & Stanley, 1966; Glass, Willson, & Gottman, 1975). Although stand-alone "A" and "B" designs are theoretically possible, they are not useful for evaluating intervention effectiveness and do not use baseline logic (i.e., comparing data from a single case under two different conditions). Thus, in this chapter, we will introduce A-B and A-B-A designs but will focus primarily on the A-B-A-B design given its experimental utility. Some authors refer to designs with at least three demonstrations of effect (i.e., A-B-A-B rather than A-B) as ***single case experimental designs (SCEDs)***. For simplicity, and because the extent to which causal relations are possible relies on a number of factors (including number of demonstrations), we will refer to all designs as SCDs. In this

chapter, we describe withdrawal designs and their use by applied researchers investigating the effectiveness of a wide range of interventions in educational and clinical settings. We discuss how baseline logic applies to this class of designs, and how threats to internal validity are evaluated. We then present guidelines for their use, discussing advantages and limitations of these designs.

Non-experimental Variations

A-B Designs

The A-B design, sometimes referred to as the "simple time series design" (Birnbrauer et al., 1974), represents the most basic non-experimental SCD. This design requires that the dependent variable be measured repeatedly under controlled baseline (A) and intervention (B) conditions. In theory, stand-alone "A" designs and "B" designs exist, in which the experimenter collects data either during typically occurring conditions ("A" design) or after the introduction of an intervention ("B" design), but the A-B design is the least complex design in which behavior *change* can be evaluated. In the A-B design, researchers collect repeated observations in the baseline condition until data are stable, then introduce the intervention. During intervention, the target behavior is again repeatedly measured, using the same measurement procedures used in the baseline condition. Any changes in the target behavior are *presumed* to be a function of the independent variable (i.e., only correlational conclusions are possible). However, since there is no direct intra-participant replication (i.e., the effect is not replicated with the same participant) there is no assurance that the independent variable is responsible for observed behavior changes. This design, despite low internal validity and weak confidence in conclusions, may be beneficial in practice when more complex designs are not feasible. As mentioned above, this basic comparison also serves as the basis for all other SCDs.

A-B-A Designs

Current guidelines for methodology indicate there must be a minimum of three potential demonstrations of effect (i.e., one demonstration and two replications), thus making the A-B-A-B design the current standard for demonstrating a functional relation. Although a researcher would *not* select the A-B-A design a priori to evaluate an intervention, these designs may appear in the literature primarily due to participant attrition during the course of a study.

Like the A-B design, the target behavior is repeatedly measured under baseline (A_1) and intervention (B) conditions. After the dependent variable has stabilized during intervention, you reintroduce the baseline condition (A_2) to the target behavior. Compared to the A-B design, the A-B-A design includes an additional demonstration of effect, strengthening the argument that the independent variable was responsible for observed changes in the dependent variable if behavior changes in the expected direction for each condition change (A_1-B, B-A_2). Conclusions can be strengthened further by extending the design to an A-B-A-B design and by replicating the experimental effect with other individuals (inter-participant replication), thereby strengthening internal and

external validity.

Despite a historical acceptance of A-B-A designs (i.e., fewer than three potential demonstrations of effect), the A-B-A design is susceptible to numerous threats to internal and external validity. First, it is possible that the introduction and withdrawal of the independent variable coincided with naturally occurring cyclical variations of the target behavior. This threat can be minimized by varying the number of observation periods in each condition and by reintroducing the intervention (B_2) (i.e., expanding the design to an A-B-A-B design). Second, there is the likelihood that dependent variable levels in A_1 will not be fully retrieved in A_2, though they should be approximated. Such sequential confounding is not uncommon in this class of designs; the additional replication in designs with more demonstrations of effect renders it less problematic than it is in A-B-A designs.

The A-B-A design is more useful than the basic A-B design from an experimental perspective. However, you would not select this design at the outset to evaluate intervention effectiveness due to the practical and ethical considerations of terminating a study with a participant in a baseline condition. From a research perspective, if ethically defensible and practical, it would be more appropriate to expand to an A-B-A-B design, thereby replicating the effect of the independent variable on the target behavior.

A-B-A-B Withdrawal Designs

The A-B-A-B design, also referred to as the "reversal design" (Baer, Wolf, & Risley, 1968), "withdrawal design" (Leitenberg, 1973), "operant design" (Glass et al., 1975), and "equivalent time series design" (Birnbrauer et al., 1974; Campbell & Stanley, 1966), has been one of the most frequently used SCDs in behavioral research. Regardless of its label, the A-B-A-B design permits a clear and convincing demonstration of experimental control because it requires the repeated introduction and withdrawal (or reversal) of an intervention. The A-B-A-B design extends the A-B-A design by adding an additional replication of effect: after implementing the first baseline condition (A_1), first intervention condition (B_1), and second baseline condition (A_2), the intervention condition is reintroduced (B_2). The most important feature of the A-B-A-B design is that it evaluates a direct replication of effect (i.e., the last two conditions [A_2-B_2] replicate the first two conditions [A_1-B_1]), with the same participant and the same behavior (direct intra- participant replication). **Withdrawal designs** refer to designs that follow the A-B-A-B condition ordering paradigm, wherein A refers to baseline conditions and the second A condition occurs when an intervention is *withdrawn*.

A-B-A-B designs increase our confidence that changes in the dependent variable are due to the intervention and *only* the intervention because there are three potential demonstrations of effect (i.e., A_1-B_1, B_1-A_2, A_2-B_2); this is the minimum number of potential demonstrations required for the design to be considered *experimental* in nature. By **experimental**, we mean that causal attributions can be made and functional relations can be demonstrated (i.e., we can say the independent variable caused the change in the dependent variable). Confidence is further strengthened when the magnitude of change in the dependent variable is immediate and abrupt (e.g., correct responding improves from 50% in the last session of A_1 to 90% in the first session of B_1), and when levels observed in the first baseline condition (A_1) are fully retrieved in the second baseline condition (A_2). Though immediate and abrupt changes in both trend and level are desirable, a believable demonstration of causality is still possible when a gradual reversal in trend is observed and when the first baseline condition level is approached, but not fully recovered. In spite of reservations by some educators and clinicians to use the A-B-A-B design, it continues to be the simplest evaluation paradigm for evaluating causality with behaviors that are reversible.

Internal Validity

Experimental control is demonstrated in the A-B-A-B design when the level and trend of a target behavior improves (relative to baseline) under intervention conditions (B_1 and B_2) and deteriorates under subsequent baseline conditions (A_2). Each replication of effect

strengthens the internal validity of results. Though one of the most frequently used designs in SCD research, some threats to internal validity are common to A-B-A-B designs; controlling for likely threats is necessary to achieve sufficient internal validity.

Maturation threats may be likely if either the baseline or intervention conditions occur for an extended period of time. This threat can be controlled for by: (a) using condition lengths that are of sufficient length to establish data patterns but not longer than necessary; (b) intervening on behaviors that are unlikely to slowly improve over time in the absence of intervention; and (c) removing the intervention in the second baseline condition. When withdrawal of the intervention results in immediate and large change in level, in a contra-therapeutic direction, it is unlikely that behavior change is due to maturation effects, even if a therapeutic trend is present in baseline.

Due to the nature of the sequential withdrawal and implementation of intervention in A-B-A-B designs, procedural infidelity and carryover effects may be likely. Procedural infidelity may be likely immediately after condition changes, but can be minimized by training implementers to a pre-determined criterion and providing implementation supports (e.g., checklist reminders) throughout the study. Similarly, carryover effects are likely when participants cannot easily discern differences between conditions. This threat can be controlled for by continuing data collection in one condition until data are stable (i.e., until the contingencies in that condition are distinguishable). In addition, you can plan to use correlated stimuli to help participants understand which condition is in effect. Some interventions have natural correlated stimuli (e.g., token boards are present during a token board intervention condition, and absent during baseline conditions); other condition stimuli can be taught (e.g., Reinforcer A is available when the light is on; Reinforcer B is available when the light is off). This can minimize the extent to which behavior change in one condition carries over to the next.

A-B-A-B designs are sensitive to attrition threats in the second baseline condition when behaviors are expected to deteriorate again, but can be minimized by including an explicit description of the withdrawal procedures during the consent process. Testing threats are likely if baseline conditions are aversive, thus researchers should devise non-aversive baseline conditions and re-start a new, modified baseline condition if data do not stabilize within a reasonable timeframe.

Similar to other designs, sampling bias is likely when multiple individuals meet inclusion criteria but only some are included as participants in the study. To control for this threat, researchers should *randomly select* final participants from the eligible individuals (e.g., if 6 children in a clinic qualify for participation but only 3 participants are desired, randomly choose 3 of the 6 children). Finally, A-B-A-B designs are only appropriate for reversible behaviors; if non- reversible behaviors are used, the behavior will not deteriorate upon removing the intervention and will prevent demonstration of experimental control. If non-reversible behaviors are of interest, a different design type should be used.

The A-B-A-B design is not particularly sensitive to history, instrumentation, or data instability threats; typical procedures for detecting and controlling for these threats

should be used. Common threats to internal validity and methods to detect, control for, and report threats are displayed in Table 9.1. Design-specific guidelines for visual analysis are available in Appendix 9.1.

Table 9.1 Common Threats to Internal Validity, and Methods to Detect and Control for Threats.

	A-B-A-B and Other Sequential Introduction and Withdrawal Designs			
	Likelihood	**Detect**	**Control**	**Report**
History	No particular likelihood associated with these designs	Visual analysis: An abrupt change in data that is not concurrent with a condition change	Continue condition until data are stable	Anecdotally describe known conditions that may have attributed to non-experimental behavior change (e.g., illness)
Maturation	May be likely if extremely long BL or intervention conditions are conducted; not typical of these designs	Visual analysis: Trend does not become steeper on introduction of intervention or behavior does not reverse during the second BL condition	Use short condition lengths (e.g., 5 sessions over 5 days); use for behaviors unlikely to gradually improve without intervention	Describe possibility of maturation threat if there is a therapeutic trend in BL
Instrumentation	Similar likelihood to other designs	Visual analysis: Differences between observers, particularly if one is blind	Use blind observers; carefully formulate and pilot definitions and recording systems; train observers to a criterion; have discrepancy discussions	Describe all reliability procedures and results; explicitly say whether observers were blind; describe reasons for low agreement
Procedural Infidelity	Likely immediately after condition changes, and in return to baseline conditions	Formative analysis of direct observational recording of fidelity data	Train implementers to criterion; re-train if necessary; provide supports to implementers such as reminder checklists; ensure implementers understand value of withdrawal	Describe all fidelity procedures and results, including training, supports, and re-training
Testing	May be likely if BL conditions are aversive	Visual analysis: Deteriorating or therapeutic trends in BL conditions	Design non-aversive BL conditions; continue condition until data are stable *or* re-start a modified BL condition that is less aversive and continue data collection until data are stable	Describe likely testing threats and solutions
Attrition Bias	Likely during the second BL condition due to nature of behavior change	Author report	Clearly describe withdrawal and reinstatement of intervention procedures during consent process; minimize time in BL conditions when behaviors are seriously detrimental to participant or attrition is likely (e.g., 3 sessions)	Describe attrition in written report and report all data from all participants, even if design was not completed

	Likelihood	Detect	Control	Report
Sampling Bias	Likely when multiple available participants meet inclusion criteria but only some are included	Reliant on author description	Randomly choose from all available participants who are eligible for participation or include all eligible participants	Describe number of participants who met criteria given your constraints (e.g., in a specific setting); if all were not chosen, describe methods for choosing participants who were included
Adaptation	Likely when observations are apparent	Participant behavior changes over time during BL conditions	Continue BL until data are stable	Describe anecdotal evidence that BL change was due to adaptation; discuss degree to which later BL data are potentially more representative of "typical" behavior
Hawthorne Effect	Likely when participants are sensitive to perceived desirable behaviors	Participant behavior is inconsistent with expectations when study begins	Use covert measurement; do not implement intervention condition if BL data do not indicate need; continue data collection to determine whether effect is temporary and change conditions only when behavior are stable	Describe anecdotal evidence that BL change was due to Hawthorne effect; discuss degree to which later BL data are potentially more representative of "typical" behavior
Multitreatment Interference	Likely in the form of carryover effects when participants cannot identify differences between conditions	Visual analysis: Delayed change in behavior when new condition is implemented	Continue all conditions until data are stable	Discuss potential presence of carryover effects when presenting results of visual analysis
Instability	No particular likelihood associated with these designs	Visual analysis: Changes in the y value of data that reduce ability to predict value of next data point given no condition change	(a) Change conditions only after data are stable *or* (b) when variability has been established over 5 or more data points *and* large between-condition changes in level are expected	Describe degree to which data instability within conditions impacted conclusions due to uncertainty regarding between-condition changes
Irreversibility of Behaviors	Likely if a non-reversible behavior is used	After behavior improves during the first intervention condition, it does not deteriorate when the intervention is removed	Use a different design type when evaluating non-reversible behaviors	If behavior does not reverse, limited conclusions can be drawn regarding the initial A-B change due to lack of experimental control

Note: BL = baseline

Concurrent Measurement of Additional Dependent Variables

Because the A-B-A-B design is, in some ways, the simplest SCD (e.g., it does not require concurrent monitoring of multiple participants, behaviors, or contexts), it may be feasible for researchers to measure multiple dependent variables in the course of a study. This could include monitoring behaviors that are functionally or topographically similar to the target behavior, in which case you are assessing response generalization (e.g., when teaching children to respond to peers, you could also measure initiations). You could also monitor behaviors that are not functionally or topographically similar to the target behavior. For example, it may be that an intervention designed to decrease the frequency of aggressive behaviors may result in a concurrent increase in engagement or appropriate use of verbal communication to make requests. This would be an important positive side effect of the intervention that may be especially important to stakeholders. Side effects can be either negative (e.g., an intervention that decreases one problem behavior results in the replacement of that behavior with another problem behavior) or positive (e.g., an intervention designed to improve engagement also results in increased peer interactions). Whichever the case, the concurrent monitoring of non-target behaviors, also called **corollary behaviors**, has practical implications for practitioners—interventions with positive effects on more than one behavior are desirable from an efficiency standpoint, while those with negative effects on some behaviors are undesirable. When two dependent variables are measured, one should be explicitly named the primary dependent variable, for which experimental decisions would be made. For example, a researcher assessing an intervention designed to improve engagement in free play activities might also measure proximity to peers. However, she would designate a priori that the engagement measure would be used to make formative decisions about condition changes. While concurrent monitoring of multiple behaviors is possible with all SCDs, it is perhaps most feasible with the A-B-A-B withdrawal design. Thus, we recommend it when resources permit.

Procedural Guidelines

When using an A-B-A-B design, adhere to the following guidelines:

1. Identify and define a reversible target behavior.
2. Select a sensitive, reliable, valid, and feasible data collection system and pilot the system and your behavior definitions.
3. Determine a priori frequency of reliability and fidelity data collection (e.g., 33% of sessions), and conduct data collection for the duration of the study.
4. Collect continuous baseline data (A) on target behaviors for a minimum of 3 consecutive days or until data are stable.
5. Introduce Intervention (B) only after data stability has been established in the initial baseline (A) condition.

6. Collect continuous data on target behaviors during Intervention (B) for a minimum of 3 consecutive days or until data are stable.
7. After a stable data pattern occurs under the intervention (B) condition, withdraw the intervention and re-introduce baseline (A) condition.
8. Repeat steps 4–6.
9. Replicate with similar participants.

Advantages

The A-B-A-B design provides a convincing demonstration of causality in applied research. It controls for many of the deficiencies associated with the A-B-A design by (a) ending in an intervention condition which is practically and ethically beneficial, and (b) providing two opportunities to replicate the positive effects of intervention (A_1 to B_1; A_2 to B_2). The A-B-A-B design can be extended to a multitreatment design (e.g., A-B-A-B-C-B-C), thereby permitting you the flexibility of comparing another intervention with the initial intervention. This is a particularly useful option when the first intervention (B) results in positive changes in the target behavior that are therapeutic but do not meet the therapeutic or educational outcome objective (e.g., Falcomata, Roane, Hovanetz, Kettering, & Keeney, 2004). In such cases, a new intervention may be introduced alone (C) or in combination with the first intervention (BC).

Limitations

The primary limitations of the A-B-A-B design relate to practical and ethical concerns rather than experimental considerations. For many practitioners who are responsible for programming durable behavior changes, even a brief withdrawal of an effective intervention may be deemed unethical. This is particularly true when target behaviors are dangerous to the client or student (e.g., eye gouging) or others (e.g., fighting). Such concerns are valid and cannot be discounted. However, you may view condition A_2 (withdrawing the intervention) as an empirical check or "probe" to see what effect an abrupt withdrawal will have on the target behavior. If the target behavior returns to unacceptable levels, that indicates that you will have to plan an additional condition after B_2 is reintroduced, one that systematically brings the individual's behavior under self control (self-management strategy) or under the control of natural contingencies. In the latter case you may have to systematically thin the reinforcement schedule (e.g., CRF to FR_2 to VR_3) or teach others to implement the intervention in the natural environment. Rusch and Kazdin (1981) outlined three strategies (i.e., sequential-withdrawal, partial-withdrawal, and partial-sequential-withdrawal) that may facilitate behavior maintenance if the total withdrawal of the intervention in the second baseline condition (A_2) results in a contra-therapeutic trend. On the other hand, the A-B-A-B design is best suited for behaviors in which we expect to continue to be under the control of current

environmental variables rather than a history of learning. For example, a series of studies have evaluated the impact of visual supports on children with disabilities (cf. Zimmerman, Ledford, & Barton, 2017) and a number of those studies used A-B-A-B design variations. This is reasonable because the participants (and many of us!) require continued use of visual supports (i.e., schedules, lists) to maintain optimal appropriate behavior. For similar reasons, the A-B-A-B design also may be used to evaluate the effectiveness of assistive technology and the use of adaptive equipment, including communication devices (Mechling & Gast, 1997), visual activity schedules (Bryan & Gast, 2000; Spriggs, Gast, & Ayres, 2007), and alternative seating equipment (Schilling & Schwartz, 2004). These interventions may be required long-term, in order to prevent behaviors from reverting to baseline levels, so consideration for supporting indigenous implementers to continue "the B condition" (the intervention) after study completion is crucial from a practical and ethical standpoint, but not an experimental one.

Due to ethical concerns, some applied researchers find it difficult to discontinue an effective intervention during the second baseline condition. If implementers and other stakeholders do not support withdrawing an effective intervention for even a brief period, the behavior probably will not reverse during the second baseline condition, thus jeopardizing a demonstration of experimental control. For this reason it is critical that procedural reliability data be collected during A_2, as during all conditions, to ensure planned condition procedures are followed. Implementers and other stakeholders should be informed of the purpose of withdrawing the intervention (i.e., a test of behavior maintenance under non-intervention conditions that will increase confidence that the behavior change is due to the intervention and not some unidentified variable).

A third limitation of the A-B-A-B design is that it is *not* appropriate for evaluating interventions with behaviors that are not likely to be reversed (e.g., writing one's name, completing an assembly task, solving addition problems, learning a mnemonic to self monitor behavior). The A-B-A-B design can be used in these and similar situations if the reason for failure on such tasks is one of motivation rather than skill acquisition. Otherwise, a multiple baseline or multiple probe design is more appropriate for evaluating experimental control. Some researchers have measured non-reversible behaviors using A-B-A-B designs by assigning slightly different behaviors for each session (e.g., a different set of sight words of approximately the same difficulty). This is not an appropriate use of the A-B-A-B paradigm because changes between sessions can be due to a number of factors rather than only condition changes (e.g., word difficulty, background knowledge, idiosyncratic interests that impact learning rate).

Conclusions

The A-B-A-B design represents the clearest and most convincing research paradigm for evaluating and demonstrating a functional relation between independent and dependent variables when a target behavior is reversible. Historically the A-B-A-B withdrawal design has been one of the most frequently used SCDs. It improves upon the A-B and A-

B-A designs by providing three potential demonstrations of effect with the same participant, thereby strengthening the internal validity of the findings. Although experimental control is demonstrated for single participants when A-B-A-B designs are used, we recommend that multiple participants be recruited to improve external validity. Thus, we recommend that you include at least three participants in your study, regardless of the experimental design, in accordance with current recommendations (e.g., Barlow & Hersen, 1984; Cooper, Heron, & Heward, 2007; Horner et al., 2005; Shadish, Hedges, Horner, & Odom, 2015; Tawney & Gast, 1984), and methodically identify differences between conditions. In light of the flexibility of the A-B-A-B design and clear evaluation of experimental control, it deserves serious consideration by practitioners and applied researchers. Table 9.2 summarizes several studies that used an A-B-A-B design to evaluate experimental control.

Table 9.2 Studies Using A-B-A-B Designs.

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
Bruzek, J. L., & Thompson, R. H. (2007). Antecedent effects of observing peer play. *Journal of Applied Behavior Analysis, 40*, 327–331.	Number: 4 Sex: 2 F, 2 M Age: 26–42 months Disability/diagnosis: TD	Setting: Room equipped with an adjoining observation area Arrangement: Dyads	High preference, medium preference, and low preference reinforcement	Percentage of time (in-zone) Responses per minute
Bryan, L., & Gast, D. L (2000). Teaching on-task and on-schedule behaviors to high-functioning children with autism via picture activity schedules. *Journal of Autism and Developmental Disorders, 30*, 553–567.	Number: 4 Sex: 3 M; 1 F Age: 7–8 Disability/diagnosis: Autism	Setting: Resource classroom Arrangement: Individual	Picture activity schedule book + graduated guidance to teach use	Percentage of on-task and on-schedule behaviors

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
Field, C., Nash, H. M., Handwerk, M. L., & Friman, P. C. (2004). A modification of the token economy for nonresponsive youth in family-style residential care. *Behavior Modification, 28*, 438–457.	Number: 3 Sex: 2 M, 1 F Age: 11–12 Disability/Diagnosis: ADHD, CD, dysthymia, PTSD, ODD, RAD	Setting: Youth's home and school Arrangement: Group	Multiple exchange token economy system	Mean frequency of intense behavior episodes Mean percent of privileges earned
Gibson, J., Pennington, R. C., Stenhoff, D., & Hopper, J. (2009). Using desktop videoconferencing to deliver interventions to a preschool student with autism. *Topics in Early Childhood Special Education, 29*, 214–225.	Number: 1 Sex: M Age: 4 Disability/diagnosis: Autism	Setting: Preschool classroom Arrangement: Large group	FCT (teach hand-raising response to access preferred items; restrict access contingent on elopement)	Percentage of intervals out of area
Gresham, F. M., Van, M. B., Cook, C. R. (2006). Social skills training for teaching replacement behaviors: Remediating acquisition deficits in at-risk students. *Behavioral Disorders, 31*, 363–377.	Number: 4 Sex: 2 M, 2 F Age: 6–8 Disability/diagnosis: At risk for EBD	Setting: Pull-out setting as described in SSIG curriculum Arrangement: Group	SST and differential reinforcement of other behavior (DRO)	Duration (percentage of time) for total disruptive behavior, time alone, and negative social interaction
Hagopian, L. P., Kuhn, S. A., Long, E. S., & Rush, K. S. (2005). Schedule thinning following communication training: Using competing stimuli to enhance tolerance to decrements in reinforcer density. *Journal of Applied Behavior Analysis, 38*, 177–193.	Number: 3 Sex: M Age: 7–13 Disability/diagnosis: PDD, ADHD, mild ID, autism, moderate ID	Setting: Treatment rooms Arrangement: Individual	FCT with extinction versus FCT with extinction and access to competing stimuli	Responses per minute (problem behavior) Responses per minute (communication)
Kern, L., Starosta, K., Adelaman, B. E. (2006). Reducing pica by teaching children to exchange inedible items for edibles. *Behavior Modification, 30*, 135–158	Number: 2 Sex: M Age: 8–18 Disability/diagnosis: severe ID, autism	Setting: Hospital, school Arrangement: Individual	Exchanging inedible for preferred edible item	Frequency of pica attempts and exchanges per hour
Kuhn, S., Lerman, D., Vorndran, C., & Addison, L. (2006). Analysis of factors that affect responding in a two-response chain in children with developmental disabilities. *Journal of Applied Behavior Analysis, 39*, 263–280.	Number: 5 Sex: 4 M, 1 F Age: 3–11 Disability/diagnosis: autism, DD, Ds, obsessive compulsive disorder, seizure disorder, ID, disruptive behavior disorder, ADHD	Setting: Library, cafeteria, and classrooms at school, multiple rooms at hospital Arrangement: Individual	Extinction, satiation, and unchaining in behavior chaining	Frequency of responding

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
Locchetta, B. M., Barton, E. E., & Kaiser, A. (2017). Using family-style dining to increase social interactions in young children. *Topics in Early Childhood Special Education*, *37*, 54–64.	Number: 9 (3 target, 6 peers) Sex: 4 M, 5 F Age: 39–60 months Disability/diagnosis: DD + seizure disorder, DD, TD	Setting: Inclusive early childhood program Arrangement: Small groups	Family-style dining and asking open-ended questions to the group	Rate of initiations and other communicative behaviors
Posavac, H. D., Sheridan, S. M., & Posavac, S. S. (1999). A cueing procedure to control impulsivity in children with attention deficit hyperactivity disorder. *Behavior Modification*, *23*, 234–253.	Number: 4 Sex: M Age: 9 Disability/ Diagnosis: ADHD, depression, bipolar mood disorder, LD	Setting: 8-week outpatient treatment program Arrangement: Group	Visual reminder, goal evaluation, constructive feedback, and reinforcement implemented in context of social skills program	Frequency of hand raising and talk-outs
Theodore, L. A., Bray, M. A., & Kehle, T. J., & Jenson, W. R. (2001). Randomization of group contengencies and reinforcers to reduce classroom disruptive behavior. *Journal of School Psychology*, *39*, 267–277.	Number: 5 Sex: M Age: unreported Disability/diagnosis: Emotional disorder	Setting: Self-contained classroom Arrangement: Group	Randomized group contingencies with reinforcement; posting of classroom expectations	Percentage of intervals with disruptions

Note: ID=intellectual disability, DD=developmental delay, CD=conduct disorder, TD=typically developing, Ds=Down syndrome, SD=seizure disorder, ADHD=attention deficit hyperactivity disorder, LD=learning disability, EBD=emotional/behavioral disorder, PTSD=post traumatic stress disorder, ODD=oppositional defiant disorder, RAD=reactive attachment disorder, M=male, F=female

Variations of the A-B-A-B Design

The A-B-A-B design uses a versatile paradigm for evaluating intervention effectiveness. Unlike group research designs, which are static, these designs are dynamic. For example, if you design a study using the A-B-A-B design and discover that the effect Intervention (B) has on the target behavior is negligible (i.e., A=B), it is *not* necessary for you to return to baseline (A) because conditions A and B are functionally equivalent; rather you have the flexibility to introduce a new intervention (C) or to combine a new condition with B (BC). If intervention C has a measurable positive effect on the target behavior you can proceed by returning to condition B. In this example you initially chose the A-B-A-B design to evaluate your intervention (B); however, because it had no effect on the dependent variable, you changed the design to an A-B-C-B-C design (or A-B-BC-B-BC design). Because of such flexibility, there are numerous studies in the applied research literature that differ from the basic A-B-A-B design and yet demonstrate experimental control. These variations of A-B-A-B designs are termed **multitreatment designs** and can also be planned for use a priori to compare interventions; we discuss these designs at length in Chapter 11. In this chapter we overview two other common variations and extensions of the A-B-A-B design: the B-A-B design and the A-B-A'-B reversal design.

Applied Example 9–1: A-B-A-B Design

Ahearn, W. H., Clark, K. M., & MacDonald, R. P. F. (2007). Assessing and treating vocal stereotypy in children with autism. *Journal of Applied Behavior Analysis*, *40*, 263–275.

Ahearn, Clark, and MacDonald (2007) studied the effects of a response interruption and redirection procedure (RIRD) on vocal stereotypy for children with autism. Two males (Mitch and Peter) and two females (Nicki and Alice) ages 3 to 11 years participated in the study. Each child was diagnosed with autism spectrum disorder and was referred by either educational or clinical staff due to vocal stereotypy. Professional service providers determined that each participant engaged in vocal stereotypy at rates that interfered with educational activities and previous assessments showed the behaviors were not maintained via social contingencies. Baseline and intervention sessions were 5 minutes in duration and were conducted in a room with appropriate materials and equipment.

Two dependent variables were measured: percentage of session with vocal stereotypy and the number of appropriate vocalizations. Data for both dependent variables were collected using duration per occurrence recording. Vocal stereotypy was converted to a percentage by dividing the total number of seconds of stereotypy by the total number of seconds in the session and multiplying by 100. Data on appropriate vocalizations were graphed as a number of vocalizations; authors reported duration was similar across occurrences. IOA data were collected for a minimum of 32% of sessions and was calculated using exact agreement. Mean reliability of vocal stereotypy was 99% for Mitch, 90% for Peter, 96% for Alice, and 93% for Nicki. IOA for cumulative number of appropriate vocalizations was 100% for all participants across all conditions.

The RIRD procedure included the delivery of teacher praise and, if possible, honoring the participant's appropriate requests. If the participant engaged in vocal stereotypy, the teacher gained the participant's attention and prompted the participant to engage in appropriate language. Mitch, Paul, and Nicki were prompted to answer social questions (e.g., "What color is your shirt?") and Alice was prompted to engage in vocal imitation. The vocal redirection tasks included skills participants had in their behavior repertoire. RIRD was discontinued when participants correctly responded to three consecutive opportunities.

The relation between percentage of intervals in which each participant engaged in vocal stereotypy and cumulative number of appropriate vocalizations was evaluated within the context of an A-B-A-B withdrawal design. Figure 9.1 presents data for Mitch's and Peter's percentage of vocal stereotypy per session and number of appropriate vocalizations across experimental conditions (baseline and RIRD).

During the baseline condition (A_1), percentage of session with vocal stereotypy was moderate to high for Mitch (range = 30% to 70%) and variable for Peter (range = 10% to 55%). Appropriate vocalizations were low (range = 0 to 5) for both participants. Upon introduction of RIRD (B_1), there was an immediate decrease in percentage of vocal stereotypy (range = 5% to 18%) for Mitch and low stable responding for Peter (range = 2% to 18%). An abrupt increase in the number of appropriate vocalizations (range = 6 to 13) occurred for Mitch with an absolute level change from 0 to 5. Peter also demonstrated an increase in the number of appropriate vocalizations (range = 0 to 8). With the return to RIRD (B_2), percentage of session of vocal stereotypy for Peter returned to low levels (range = 0% to 14%) and number of appropriate vocalizations returned to high, but variable levels (range = 1 to 9).

As displayed in Figure 9.1, these data provide a convincing evaluation and demonstration of a functional relation between percentage of time participants engaged in vocal stereotypy and number of appropriate vocalizations for Mitch. Results were replicated for Alice for both dependent variables and for percentage of session with vocal stereotypy for Nicki.

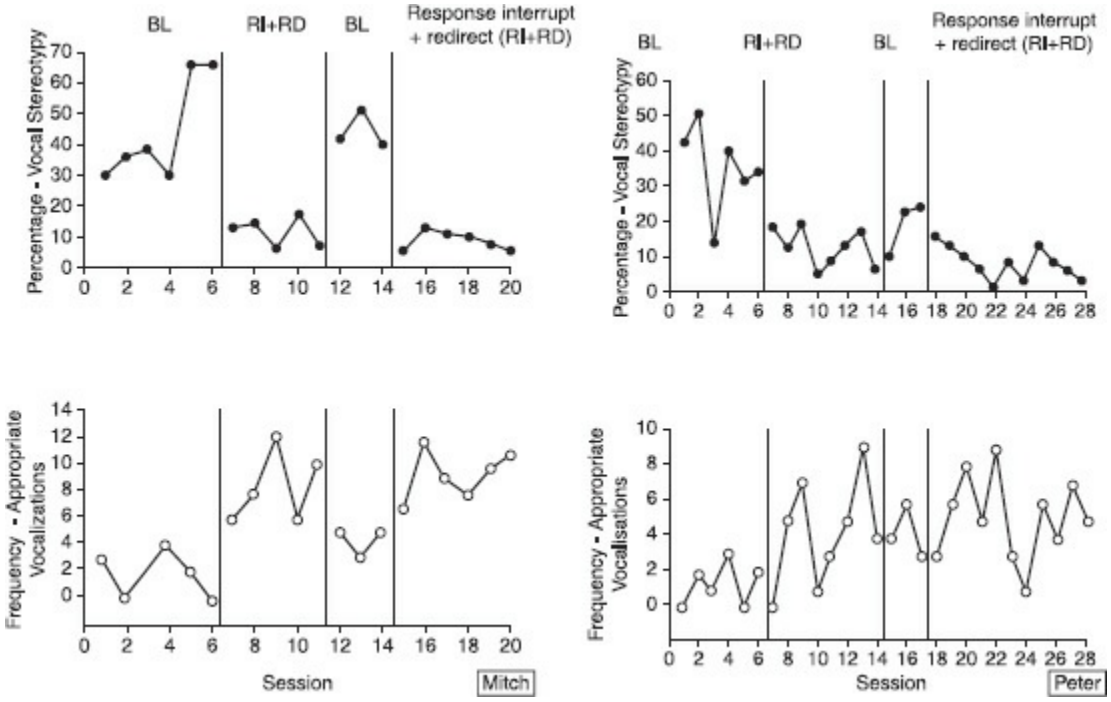


Figure 9.1 A-B-A-B designs for two participants, with two dependent variables measured for each.

Source: Ahearn, W. H., Clark, K. M., & MacDonald, R. P. F. (2007). Assessing and treating vocal stereotypy in children with autism. *Journal of Applied Behavior Analysis, 40*, 263–275.

B-A-B Designs

The B-A-B design is a research design you may use when a student or client exhibits

self-injurious, physically aggressive, or otherwise highly undesirable behaviors. For ethical reasons, due to potential danger to the student or others, you may not have the opportunity to collect baseline data. It is important to remember that although it may be understandable that baseline data are not collected, it prevents the evaluation of experimental control. The absence of pre-intervention behavior measures precludes assessment of baseline data patterns of the behavior prior to the introduction of the intervention (B_1). Thus, there are no empirical means for (a) comparing the effects of intervention to the pre-intervention data and (b) assessing whether the level and trend in baseline (A_1) replicate the level and trend prior to the introduction of B_1. Because of these experimental limitations few B-A-B designs are found in the recent applied research literature. To ensure a sufficient number of replications, you can modify a B-A-B design by adding two additional conditions (B-A-B-A-B design); this may be a similarly undesirable variation due to additional baseline sessions but might be appropriate if quickly garnering stakeholder support is critical (e.g., a child's teachers or parents may be more willing to tolerate baseline conditions after you have demonstrated an intervention exists that will result in behavior change).

Applied Example 9–2: A-B-A-B with Concurrent Monitoring

Carnine, D. W. (1976). Effects of two teacher-presentation rates on off-task behavior, answering correctly, and participation. *Journal of Applied Behavior Analysis*, *9*, 199–206.

This study evaluated the effects of fast and slow presentation rates on the off-task, correct answering, and participation behaviors of two "low achieving" first-grade students. The two participants, one male and one female, were two of the four children in the lowest- performing first-grade reading group in their school. The children were instructed in reading using the Level I DISTAR program for 30 minutes each day. Instruction occurred in the rear of the classroom while other students worked independently or were instructed in small-groups in other areas of the room. The teacher conducted the instruction during the first 33 sessions, and the student teacher conducted the last 5 sessions.

During each instructional session, two data collectors collected data on each participant's off-task, correct answering, and participation behaviors using trial-based event recording, and calculated a percentage of behavior occurrence by dividing occurrences by total trials and multiplying by 100. IOA on each dependent measure was calculated using the point-by-point method, which yielded a mean percentage agreement for all measures above 90%.

The independent variable was task presentation rate. Two experimental conditions, slow-rate presentation (A) and fast-rate presentation (B), were alternated to assess their effects on the three dependent variables. During the slow-rate presentation condition the teacher silently counted to five after each child's response before presenting the next task. In contrast, during the fast-rate presentation condition the teacher presented the next task immediately after each response. The teacher presented the lesson exactly as it was written in the DISTAR program. The teacher delivered general verbal praise at a constant rate across conditions by utilizing cues from a preprogrammed tone from an audio-cassette recorder, equipped with an earplug. This constant schedule across conditions prevented confounding of verbal praise and presentation rate over the course of the study. Observers recorded task presentation rate: after each block of 10 trials observers recorded the duration of time it took to complete the block of 10 trials and reset their stopwatches. The presentation rate was calculated by dividing total instructional time for a session by number of tasks presented during that session. IOA data were collected during 87% of the sessions.

The effect of slow-rate presentation (A) and fast-rate presentation (B) on participants' off-task, correct answering, and participation behaviors was evaluated

within the context of an A-B-A-B-A'-B' design. The teacher instructed the reading group during the first four phases of the study (A_1-B_1-A_2-B_2), while the student teacher conducted the group during the last two phases (A'_3-B'_3) of the investigation. Including the student teacher permitted a brief assessment of stimulus generalization across teachers.

Figure 9.2 illustrates the mean percentage of the three dependent variables for Subject 1. It should be noted that off-task data drove the design (i.e., decisions to move conditions were based on these data) and that condition labels were omitted above "% off task behavior" but are identified above the other two behaviors monitored. For Subject 1, during A_1 (slow-rate condition) both level and trend were stable using an "80% of the data points falling within a 20% range" as the definition for stability. Upon introduction of B_1, there was an immediate, though modest, change in level. Subsequent days in the fast-rate condition resulted in a stable, zero celeration trend near the floor. This change in level was replicated across subsequent condition comparisons and was replicated with "Subject 2."

30"/>This study indicated that faster presentation rate may result in decreased off-task behavior and increased correct responses and participation. The direct intra-participant replication of effect across the two conditions with two different teachers increases the internal validity and reliability of findings. The generality of these findings was demonstrated by replicating different responding patterns with two participants (i.e., direct inter-participant replication).

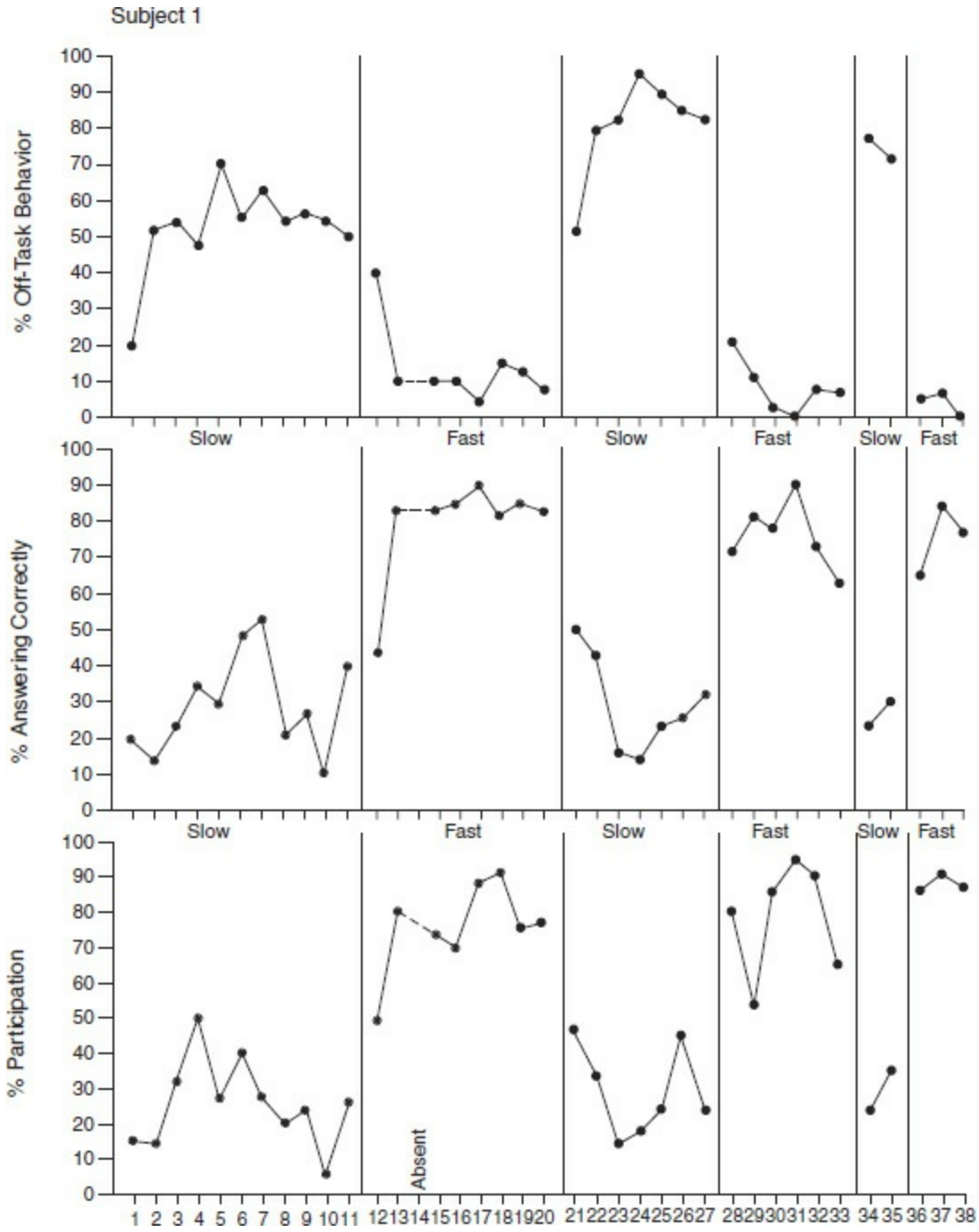


Figure 9.2 Graphs from A-B-A-B-A-B design (with final A-B comparison presented by a different implementer).

Source: Carnine, D. W. (1976). Effects of two teacher-presentation rates on off-task behavior, answering correctly, and participation. *Journal of Applied Behavior Analysis*, *9*, 199–206.

Murphey, Ruprecht, Baggio, and Nunes (1979) used a B-A-B design to evaluate differential reinforcement of other behavior plus mild punishment (contingent water squirts) intervention on the frequency of self-choking behavior of a young adult with profound intellectual disabilities. Figure 9.3 displays the number of self-chokes emitted by the participant during each condition of the investigation. The data indicate the mean frequency for self-chokes during initial treatment (B_1), withdrawal of treatment (A_1), and

reinstatement of treatment (B_2) conditions were 22, 265, and 24, respectively. The withdrawal of the treatment package resulted in an immediate and abrupt change in frequency of self-choking in a contra-therapeutic direction. This level was reversed immediately upon reintroduction of treatment procedures (B_2). Although an initial baseline condition prior to the introduction of the treatment package would have permitted a comparison with the pre-intervention levels of the behavior and strengthened the demonstration of experimental control, the immediate and abrupt level changes between conditions give credence to the effectiveness of the intervention while maintaining increased acceptability. The investigators' decision to omit an initial baseline condition illustrates the dilemma that sometimes confronts applied researchers who deal with potentially dangerous behaviors in educational and clinical settings.

Although it may be impractical or unethical to collect baseline data for an extended time period with dangerous behaviors, it may be possible to collect baseline data over a shortened period to establish a baseline rate. Kennedy and Souza (1995), for example, collected only one 6-minute session of baseline data for a 19-year-old male with a profound intellectual disability who exhibited a high rate of eye-gouging, before introducing their eye-goggle intervention condition. Although abbreviated, the collection of pre-intervention data strengthened their demonstration of experimental control, clearly showing the effectiveness of the eye-goggle condition in immediately and abruptly decreasing the number of seconds the participant engaged in eye-poking.

If you determine it is unethical and/or impractical to collect pre-intervention data, we recommend that you proceed as follows when implementing a B-A-B design: (a) justify on ethical and/or practical grounds why pre-intervention data cannot be collected; (b) introduce an intervention (B_1), based on a functional behavior assessment, and look for an immediate and abrupt level change in behavior in a therapeutic direction; (c) conduct a *brief* withdrawal of the intervention (A_1) after the behavior reaches the established therapeutic criterion level in (B_1); (d) reintroduce the intervention (B_2) after a *brief* reversal in level and/or trend are observed. Assuming you are monitoring the frequency of an inappropriate behavior, a demonstration of effect will be established when the initial introduction of the independent variable results in a low and ideally therapeutic level of the target behavior (B_1), followed by an immediate, though brief, increase in the frequency of the behavior (A_1). Once a change in level or trend is observed (A_1 compared to B_1), reintroduce the intervention (B_2), and ideally, an immediate and abrupt change in level and trend that replicates B_1 will be observed.

It is preferable to collect even brief baseline data prior to introducing intervention. Without an initial baseline measure it is impossible to evaluate the effect of the intervention on the natural frequency of the behavior. In contrast to the A-B-A design, however, the B-A-B design has the advantage of ending with intervention and allowing two demonstrations of intervention effectiveness. If practical and ethical considerations permit, a more believable demonstration of causality is possible with the more complete A-B-A-B design or a B-A-B-A-B design.

A-B-A'-B Reversal Designs

We have chosen to use the notation A-B-A'-B for a class of SCDs that are procedurally "*true reversals*," in that the independent variable is withdrawn from one behavior and applied to a second, possibly incompatible behavior that is being concurrently measured. Thus, **reversal designs** involve reversing intervention contingencies during A_2, rather than simply withdrawing the intervention. For example, Goetz and Baer (1973) conducted a no-reinforcement (A) baseline condition and measured the number of different block forms built by children, and then reinforced "new" form building (B). During the second A condition (A'), they instead reinforced "old" forms (previously built within the session).

Withdrawal vs. Reversal Design Distinction

Leitenberg (1973) restricts the use of the term *reversal design* to those SCDs where the independent variable is truly reversed in the third condition (A_2), *not* simply withdrawn. Operationalized, the reversal design usually entails concurrently monitoring two behaviors during the first baseline condition (e.g., hands on desk and hands in lap). Historically, the two monitored behaviors have been incompatible, but this is not required. After a stable baseline level and trend are established with both behaviors, the independent variable is applied to one of the behaviors (e.g., hands on desk) during B_1. If the intervention strategy has a positive effect on this behavior, then it is applied to the concurrently monitored behavior (hands in lap) in the third condition (commonly referred to as A'). It is at this juncture that the reversal design is distinguished from the withdrawal design. Not only is the intervention withdrawn from the target behavior in the reversal design, it is applied to a concurrently monitored behavior during the third (A') condition. If there is a decrease in the one behavior (hands on desk) and a concomitant increase in the incompatible behavior (hands in lap), then a functional relation between the independent and two dependent variables is demonstrated. When the independent variable is reintroduced to the first behavior (hands on desk), experimental control is further strengthened by reversing data trends of the two behaviors in B_2.

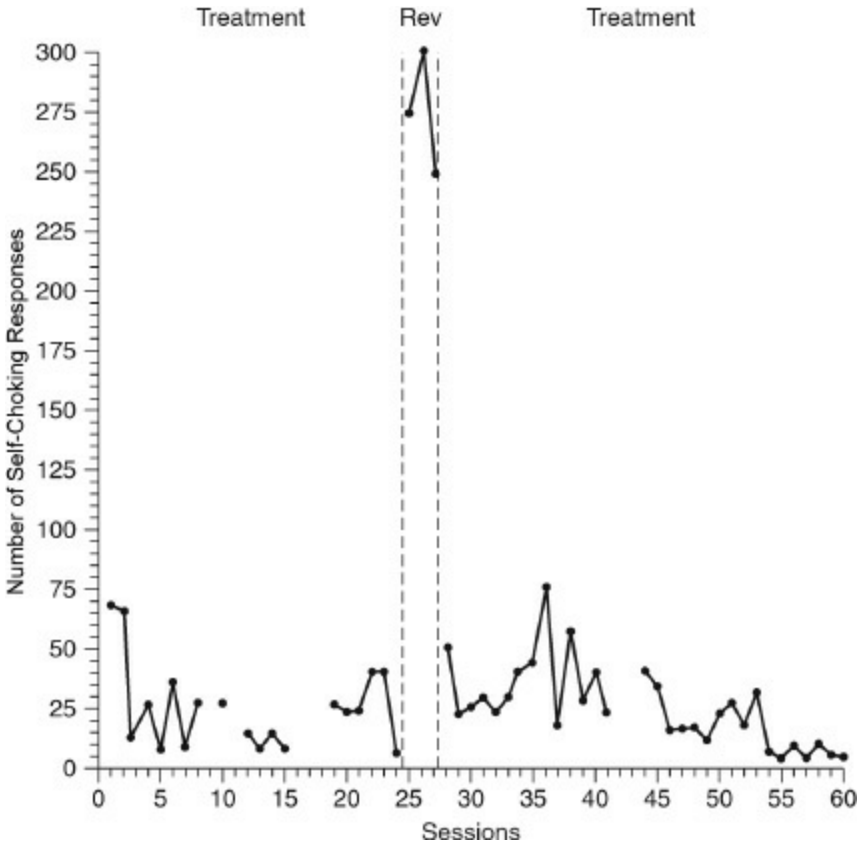


Figure 9.3 B-A-B design with brief withdrawal.

Source: Murphey, R. J., Ruprecht, M. J., Baggio, P., & Nunes, D. L. (1979). The use of mild punishment in combination with reinforcement of alternate behaviors to reduce the self-injurious behavior of a profoundly retarded individual. *American Association for the Education of the Severely-Profoundly Handicapped Review*, *4*, 187–195.

The key distinction between reversal and withdrawal designs is that when the reversal design is used, researchers (a) withdraw or remove the intervention from one behavior and (b) simultaneously apply it to an incompatible behavior. The withdrawal design, on the other hand, involves simply removing the intervention during the third condition of the design (A_2). An easy way to distinguish the two designs may be to associate reversal designs with differential reinforcement of an incompatible behavior (DRI) and the withdrawal design with extinction (e.g., systematic ignoring of a single attention getting behavior). A true reversal design is a powerful demonstration of experimental control because it includes three potential opportunities for demonstrating the effect of the independent variable on two incompatible behaviors, though there are few "true" reversal designs in the literature.

The distinction between the reversal design (A-B-A'-B) and the withdrawal design (A-B-A-B) is small, but is warranted given the procedural differences relative to the third condition (A' or A_2). It is therefore recommended that the A-B-A-B design notation be restricted to those time series designs in which A_2 procedures are identical to A_1 procedures and when the independent variable is withdrawn. It is our recommendation that an A-B-A'-B design be referred to as a "reversal design" only when (a) the first and

third conditions of the study are **not** procedurally identical and (b) the independent variable is present in the third condition, applied to a different behavior than in the second and fourth conditions.

Procedural Guidelines

When using a reversal design, adhere to the following guidelines:

1. Identify and define a reversible target behavior.
2. Select a sensitive, reliable, valid, and feasible data collection system and pilot the system and your behavior definitions.
3. Determine a priori frequency of reliability and fidelity data collection (e.g., 33% of sessions), and conduct data collection for the duration of the study.
4. Collect continuous baseline data (A) on target behaviors for a minimum of 3 consecutive days or until data are stable.
5. Introduce Intervention (B) only after data stability has been established in the initial baseline (A) condition.
6. Collect continuous data during Intervention (B) on target behaviors for a minimum of 3 consecutive days or until data are stable, and continue to monitor non-target behaviors on a regular schedule.
7. After a stable data pattern occurs under the intervention (B) condition, *reverse* the intervention contingencies (e.g., apply reinforcement to an incompatible behavior).
8. After a stable data pattern occurs under the A' (reversal) condition, re-introduce the intervention and continue measurement until a stable data pattern emerges.
9. Replicate with similar participants.

Applied Example 9–3: A-B-A'-B Design

Lopez, K., Dewey, A., Barton, E. E., & Hemmeter, M. L. (2017). The use of descriptive praise to increase diversity during easel painting. *Infants and Young Children*, *30*, 133–146.

Lopez, Dewey, Barton, and Hemmeter (2017) investigated the effectiveness of descriptive praise on increasing the diversity of art activities for four preschool children (Albert, Brice, Cora, and Dana). The study was conducted in a university-affiliated preschool in an inclusive classroom. Participants were 44 to 47 months old and did not have diagnosed disabilities. Inclusion criteria included that children: demonstrated limited diversity during art activities, possessed adequate motor and language skills to meaningfully engage in art activities, were able to identify colors and forms, and demonstrated in pre-intervention assessment that they engaged in behaviors at a higher rate when an adult provided descriptive praise.

The primary dependent variable was the number of diverse acts during painting. Researchers calculated diversity by adding the number of different forms (e.g., circular lines; straight lines) used, brush switches, and color switches. Data were collected using event recording in which data collectors tallied the number of different forms the child used, the number of times the child used a new color, and the number of times the child switched to a different brush during a single painting activity. IOA data were collected for at least 20% of sessions during the course of the study, and ranged from 50% to 100% agreement across participants and conditions.

The independent variable was the implementer's descriptive praise statements in the context of four (one for each participant) true reversal designs (A-B-A'-B; Study 1 only). During baseline (A), participants painted freely and implementers provided general praise statements. During painting with descriptive praise (B_1 and B_2), implementers provided descriptive praise statements when the children used different colors, forms, or brushes (e.g., "I love how you are using blue and green paints!"). In the reversal condition (A'), implementers *reversed* the praise statements used in B, and provided descriptive praise statements for *sameness* (i.e., using the same colors, forms, and brushes), rather than diversity. Data were collected for 5 minutes or until children indicated they were finished. Procedural fidelity data were collected for 100% of sessions using a checklist (fidelity=100%) and tallying the number of implementer praise statements during the sessions. As planned, *descriptive* praise statements were 0 across participants in baseline conditions; descriptive praise ranged from 10–35 statements during B and A' conditions.

Figure 9.4 shows the number of the four participants' diverse acts ("creativity score") during each painting session. During the initial baseline condition, Albert's

diversity score ranged from 21 to 33, Brice's from 20 to 32, Dana's from 26 to 28, and Cora's was variable with an increasing trend. Upon introduction of the first intervention condition (B_1, descriptive praise for diverse painting acts), diversity score initially decreased for three participants (Albert, Brice, Cora) then data were variable and generally consistent with baseline levels. Dana's diversity score increased slightly in level compared to baseline. After the reversal was introduced (A', descriptive praise for sameness), levels decreased for all participants as compared to both the initial baseline and intervention conditions. After the re-introduction of descriptive praise for diversity (B_2), all participants' diversity scores immediately increased to levels higher than B_1. Results of this study suggest that changes in teacher praise may influence preschool children's diversity during art activities.

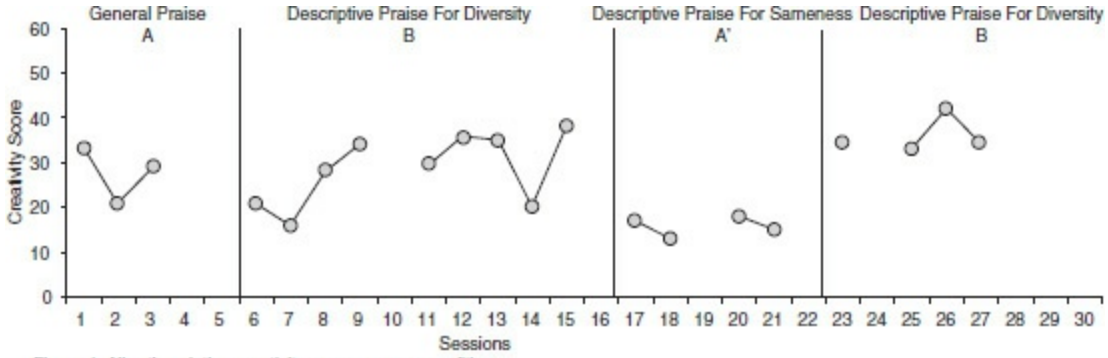


Figure 1. Albert's painting creativity scores across conditions.

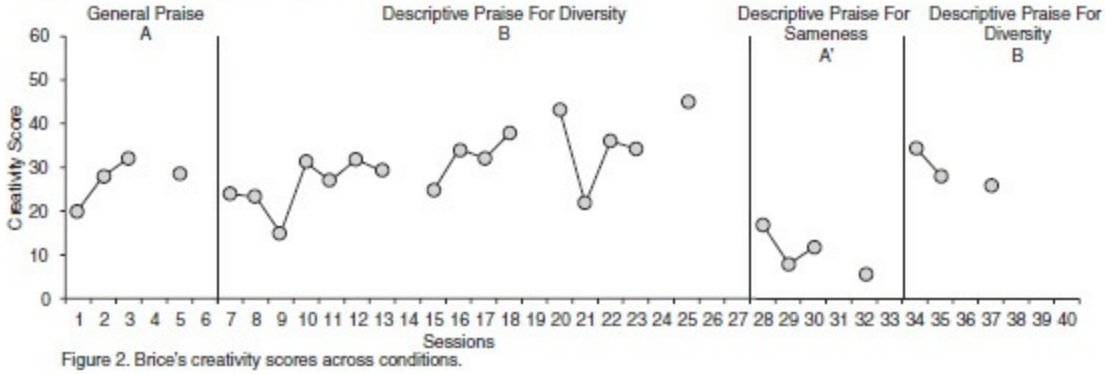


Figure 2. Brice's creativity scores across conditions.

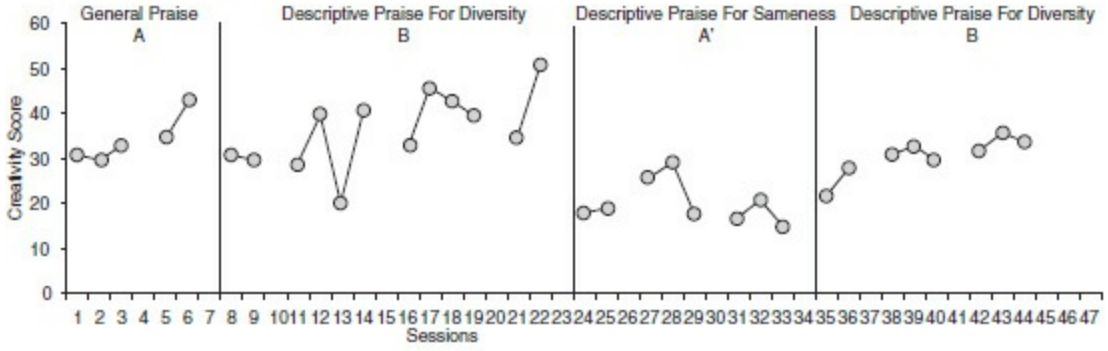


Figure 3. Cora's creativity scores across conditions

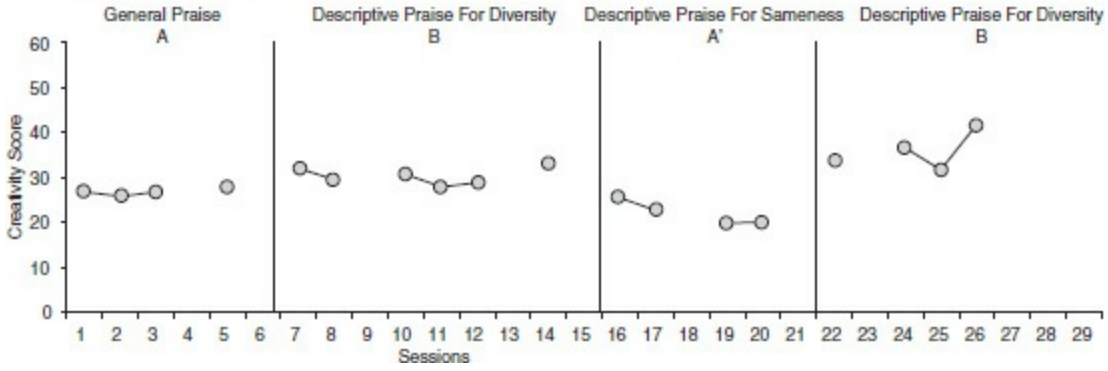


Figure 9.4 Graphs for four participants in the context of a true reversal design.

Source: Lopez, K., Dewey, A., Barton, E. E., & Hemmeter, M. L. (2017). The use of descriptive praise to increase diversity during easel painting. *Infants and Young Children, 30,* 133–146.

Summary

This chapter has described and exemplified the basic A-B-A-B design as well as common variations. In spite of some educators' and clinician's reluctance to employ an A-B-A-B design because of the brief withdrawal (or reversal) requirement, it continues to be a convincing and straightforward evaluation paradigm for evaluating experimental control. Its primary advantage, when compared to abbreviated forms of the design (A-B, A-B-A, B-A-B), is that it provides two replications of intervention effectiveness with the same research participant and the same behavior under similar stimulus conditions. This also is an advantage of the A-B-A-B design over the more popular multiple baseline and multiple probe designs.

Appendix 9.1

Visual Analysis for A-B-A-B Withdrawal Design

Adequate design	**Examples**: A-B-A-B, A-B-A-C-A-C **Non-Examples**: A-B-A-C, A-B-C-A, A-B-A
Visual analysis considerations specific to design	None
Common and potentially problematic data patterns	• **Behavior does not reverse in A_2**. If this occurs, you have potentially chosen a non-reversible behavior, but you cannot rule out history as an explanation for behavior change, even if the initial behavior change occurred concurrently with intervention onset. • **Behavior does not *fully* reverse in A_2**. If this occurs, your confidence in the presence of a functional relation decreases because of a lack of *consistency* in data patterns between A_1 and A_2. Determination of a functional relation can still be made if all other criteria below are met. • **Delayed change across conditions**. A delayed change is less problematic if (a) you continue conditions until data are stable, (b) a delay was predicted a priori, (c) the delay occurs in both intervention conditions, and (d) the latency and magnitude of the delay are consistent. • **Small magnitude changes**. Small changes are not problematic if data patterns are consistent for similar conditions (e.g., behavior changes were small for both A_1 B_1 and A_2 B_2) and if between-condition level change exceeds within-condition variability (e.g., no overlap is present). Small magnitude changes are potentially problematic if agreement data are discrepant (e.g., data from a second observer might suggest no change occurred; assessed via visual analysis of plotted data from both observers). • **Highly variable data in one or more condition**. Variable data are less problematic if between-condition level change exceeds within-condition variability (e.g., no overlap), or if changes in variability predictably change across conditions (e.g., high variability in baseline followed by low variability during intervention). Variability is problematic if there is a high percentage of overlapping data points or variability otherwise precludes making a decision regarding

	behavior change. • **Therapeutic trends in baseline conditions**. Therapeutic trends are not problematic if (a) a large and abrupt change in level coincides with implementation of the intervention condition, and (b) A_2 results in contra-therapeutic behavior change.
Convincing Functional Relation	• Behavior patterns in A_1 and A_2 are similar • Behavior patterns in B_1 and B_2 are similar • Changes from A_1 B_1 and A_2 B_2 are similarly therapeutic • Changes from B_1 A_2 are contra-therapeutic • All changes are abrupt and concurrent with condition changes • Overlap is minimal • Variability and trends in any condition do not preclude ability to identify between-condition changes

References

Ahearn, W. H., Clark, K. M., & MacDonald, R. P. F. (2007). Assessing and treating vocal stereotypy in children with autism. *Journal of Applied Behavior Analysis*, *40*, 263–275.

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.

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

Birnbrauer, J. S., Peterson, C. R., & Solnick, J. V. (1974). Design and interpretation of studies of single subjects. *American Journal of Mental Deficiency*, *79*, 191–203.

Bruzek, J. L., & Thompson, R. H. (2007). Antecedent effects of observing peer play. *Journal of Applied Analysis*, *40*, 327–331.

Bryan, L. C., & Gast, D. L. (2000). Teaching on-task and on-schedule behaviors to high functioning children with autism via picture schedules. *Journal of Autism and Developmental Disabilities*, *30*, 553–567.

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

Carnine, D. W. (1976). Effects of two teacher-presentation rates on off-task behavior, answering correctly, and participation. *Journal of Applied Behavior Analysis*, *9*, 199–206.

Cooper, J. O., Heron, T. E., & Heward, W. L. (2007). *Applied behavior analysis* (2nd ed.). Upper Saddle River, NJ: Pearson: Merrill Prentice Hall.

Falcomata, T. S., Roane, H. S., Hovanetz, A. N., Kettering, T. L., & Keeney, K. M. (2004). An evaluation of response cost in the treatment of inappropriate vocalizations maintained by automatic reinforcement. *Journal of Applied Behavior Analysis*, *37*, 83–87.

Field, C., Nash, H. M., Handwerk, M. L., & Friman, P. C. (2004). A modification of the token economy for nonresponsive youth in family-style residential care. *Behavior Modification*, *28*, 438–457.

Gibson, J., Pennington, R. C., Stenhoff, D., & Hopper, J. (2009). Using desktop videoconferencing to deliver interventions to a preschool student with autism. *Topics in Early Childhood Special Education*, *29*, 214–225.

Glass, G. V., Willson, V. L., & Gottman, L. J. (1975). *Design and analysis of time-series experiments*. Boulder, CO: Colorado Associated University Press.

Goetz, E. M., & Baer, D. M. (1973). Social control of form diversity and the emergence of new forms in children's block building. *Journal of Applied Behavior Analysis*, *6*, 209–217.

Gresham, F. M., Van, M. B., & Cook, C. R. (2006). Social skills training for teaching replacement behaviors: Remediating acquisition deficits in at-risk students.

Behavioral Disorders, *31*, 363–377.

Hagopian, L. P., Kuhn, S. A., Long, E. S., & Rush, K. S. (2005). Schedule thinning following communication training: Using competing stimuli to enhance tolerance to decrements in reinforcer density. *Journal of Applied Behavior Analysis*, *38*, 177–193.

Horner, R., Carr, E., 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.

Kennedy, C., & Souza, G. (1995). Functional analysis and treatment of eye poking. *Journal of Applied Behavior Analysis*, *28*, 27–37.

Kern, L., Starosta, K., & Adelaman, B. E. (2006). Reducing pica by teaching children to exchange inedible items for edibles. *Behavior Modification*, *30*, 135–158.

Kuhn, S., Lerman, D., Vorndran, C., & Addison, L. (2006). Analysis of factors that affect responding in a two- response chain in children with developmental disabilities. *Journal of Applied Behavior Analysis*, *39*, 263–280.

Leitenberg, H. (1973). The use of single-case methodology in psychotherapy research. *Journal of Abnormal Psychology*, *82*, 87–101.

Locchetta, B. M., Barton, E. E., & Kaiser, A. (2017). Using family-style dining to increase social interactions in young children. *Topics in Early Childhood Special Education*, *37*, 54–64.

Lopez, K., Dewey, A., Barton, E. E., & Hemmeter, M. L. (2017). The use of descriptive praise to increase diversity during easel painting. *Infants and Young Children*, *30*, 133–246.

Mechling, L. C., & Gast, D. L. (1997). Combination audio/visual self-prompting system for teaching chained tasks to students with intellectual disabilities. *Education and Training in Mental Retardation and Developmental Disabilities*, *32*, 138–153.

Murphey, R. J., Ruprecht, M. J., Baggio, P., & Nunes, D. L. (1979). The use of mild punishment in combination with reinforcement of alternate behaviors to reduce the self-injurious behavior of a profoundly retarded individual. *AAESPH Review*, *4*, 187–195.

Posavac, H. D., Sheridan, S. M., & Posavac, S. S. (1999). A cueing procedure to control impulsivity in children with attention deficit hyperactivity disorder. *Behavior Modification*, *23*, 234–253.

Rusch, F. R., & Kazdin, A. E. (1981). Toward methodology of withdrawal designs for the assessment of response maintenance. *Journal of Applied Behavior Analysis*, *14*, 131–140.

Schilling, D. L., & Schwartz, I. S. (2004). Alternative seating for young children with autism spectrum disorder: Effects on classroom behavior. *Journal of Autism and Developmental Disorders*, *34*, 423–432.

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). *The role of between-case effect size in conducting, interpreting, and summarizing single-case research.* Washington, DC: National Center for Education Research, Institute of Education Sciences, U. S. Department of Education. Retrieved from

http://files.eric.ed.gov/fulltext/ED562991.pdf

Spriggs, A. D., Gast, D. L., & Ayres, K. M. (2007). Using picture activity schedule books to increase on-schedule and on-task behaviors. *Education and Training in Developmental Disabilities*, *42*, 209–223.

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

Theodore, L. A., Bray, M. A., & Kehle, T. J., & Jenson, W. R. (2001). Randomization of group contingencies and reinforcers to reduce classroom disruptive behavior. *Journal of School Psychology*, *39*, 267–277.

Zimmerman, K. N., Ledford, J. R., & Barton, E. E. (2017). Using visual activity schedules for young children with challenging behavior. *Journal of Early Intervention*, *39*, 339–358. doi: 10.1177/1053815117725693

10
Multiple Baseline and Multiple Probe Designs

David L. Gast, Blair P. Lloyd, and Jennifer R. Ledford

Important Terms

time-lagged designs, continuous measurement, intermittent measurement, concurrent, functionally independent, functionally similar, behavioral covariation, inconsistent intervention effects, probe session, probe condition, conditions variation, days variation, nonconcurrent

Practitioners have been subjected to increasing pressure from consumers, professional

organizations, legislatures, and courts to account for their intervention practices. They have been asked to provide objective, data-based responses to such questions as: "When should intervention programs be maintained, modified, or replaced?" and "Can a student or client's progress be attributed to identifiable instructional strategies?" Within the framework of SCD, there is a class of designs well-suited for evaluating and demonstrating accountability in clinical and educational settings, namely **time-lagged designs**. There are two widely used variations of time-lagged designs: multiple baseline (MB) and multiple probe (MP) designs. (The third and lesser-used variation is the changing criterion design; see Chapter 12). Both MB and MP designs involve assessing multiple A-B comparisons by implementing A to B condition changes at three or more different points in time for three or more targets rather than introducing and withdrawing the intervention with a single target (as with A-B-A-B designs). Both MB and MP designs are flexible (i.e., the learner's behavior controls the pace and choice of programming procedures); are rigorous in their evaluation of threats to internal validity; and are practical for practitioners who want their research efforts to be compatible with their intervention activities. In this chapter, we describe MB and MP designs and their use by applied researchers investigating the effectiveness of a wide range of interventions in educational and clinical settings. We discuss how baseline logic applies to this class of designs, and how threats to internal validity are evaluated. We then present guidelines for their use, discussing advantages and limitations of variations of both designs.

Baseline Logic in MB and MP Designs

Baer, Wolf, and Risley (1968) introduced MB designs to behavioral researchers in their seminal article describing applied behavior analysis. It was 10 years later that Horner and Baer (1978) described a variation of the MB design they termed "multiple probe technique." Both designs are based on the same baseline logic for evaluating threats to internal validity and demonstrating experimental control. Procedurally, MB and MP designs differ in one way: the frequency with which *pre-intervention* data are collected. Whereas MB designs require a plan for the continuous measurement of all targets prior to the introduction of the independent variable, the plan for MP designs is to collect data intermittently prior to the introduction of the intervention. **Continuous measurement** refers to the planned implementation and data collection during each opportunity or session; for example, if a study is to occur on Mondays, Wednesdays, and Fridays, continuous measurement would imply that data were collected on each of these three days every week. **Intermittent measurement** refers to the planned absence of data collection during some opportunities or sessions. In the example above, intermittent measurement would imply that data were *not* collected on some Mondays, Wednesdays, and Fridays during the course of the study. During intervention conditions, continuous measurement is always advised, regardless of design. In MP designs, intermittent measurement is allowable during pre-intervention conditions.

Choosing to measure continuously or intermittently influences experimental rigor, likely threats to internal validity, and practicality of the two designs. Both designs are well-suited to the practical requirements of applied research in that they (a) lend themselves to program efficacy measures, (b) do not require a withdrawal of intervention, and (c) are easy to conceptualize and implement.

There are three principal variations or types of MB and MP designs:

1. Across several *behaviors* or *behavior sets* of a single individual. For example, a researcher might assess the effects of a paraprofessional training program on improving use of behavior specific praise (Behavior 1), responsive interactions (Behavior 2), and providing reinforcement for requests (Behavior 3).
2. Across several *contexts* or stimulus conditions (e.g., settings, adults, arrangements, formats, materials). For example, a researcher might assess the effects of a paraprofessional training program on improving instructional behaviors during independent work (Setting 1), lunch (Setting 2), and after-school care (Setting 3).
3. Across several *participants* (i.e., individuals or groups of individuals). For example, a researcher might assess the effects of a paraprofessional training program on improving use of behavior specific praise for three different paraprofessionals.

As with A-B-A-B designs, multiple baseline designs are used to compare baseline (A)

and intervention (B) conditions. However, as mentioned above, there is no withdrawal of the intervention. Instead, the researcher replicates the A-B comparison with several behaviors, contexts, or participants. These additional replications are typically plotted separately and presented together as a tiered figure. That is, the first A-B comparison (1st tier) is presented atop the second comparison (2nd tier), and the 3rd and remaining tiers are placed below those. These replications are not simple sequential duplications, however. Instead, behaviors across tiers are measured **concurrently**—that is, data collection begins at the same time for all 3+ tiers. The time-lagged procedure applies only to the temporal disparity regarding when the *intervention condition* begins.

Because of the similarities of MB and MP designs, we discuss them together before discussing specific guidelines, advantages, and limitations associated with each. For ease of understanding, we use the term *baseline*, rather than *probe*, to refer to the pre-intervention condition (i.e., "A"). In distinguishing these two designs it is important to understand that it is only the *planned frequency* with which data are collected prior to introducing the independent variable that differentiates the two designs. Missed opportunities for data collection (e.g., participant absences) in the context of an MB design do not constitute an MP design.

Internal Validity

Studies have adequate internal validity when all likely threats are controlled for, and experimental control is demonstrated when adequate internal validity is present and when behavior change occurs *when and only when* the intervention is introduced to each target tier, for at least three tiers with concurrent start points. Despite their widespread use, some threats to internal validity are particularly likely when these designs are used. Common threats to internal validity are described below and shown in Table 10.1. Design-specific guidelines for visual analysis are available in Appendix 10.1.

Table 10.1 Common Threats to Internal Validity, and Methods to Detect and Control for Threats

	Multiple Baseline and Multiple Probe Designs			
	Likelihood	**Detect**	**Control**	**Report**
History	Highly likely due to prolonged condition lengths; particularly problematic in MP designs due to intermittent measurement	Visual analysis: An abrupt change in data that is not concurrent with a condition change	Continue condition until data are stable; do not change conditions in a tier if a history effect is present in that tier *or* a different tier	Describe anecdotally known conditions that may have attributed to non-experimental behavior change (e.g., illness)
Maturation	May be likely especially for later tiers with protracted BL conditions	Visual analysis: Shallow trend during BL conditions	Use a different design (A-B-A-B); use for behaviors unlikely to gradually improve without intervention; collect baseline data more frequently (MP only)	Describe possibility of maturation threat if there is a therapeutic trend in BL
Instrumentation	Similar likelihood to other designs	Visual analysis: Differences between observers, particularly if one is blind	Use blind observers; carefully formulate and pilot definitions and recording systems; train observers to a criterion; have discrepancy discussions	Describe all reliability procedures and results; explicitly say whether observers were blind; describe reasons for low agreement
Procedural Infidelity	Likely immediately after condition changes	Formative analysis of direct observational recording of fidelity data	Train implementers to criterion; re-train if necessary; provide supports to implementers such as reminder checklists; ensure implementers understand value of withdrawal	Describe all fidelity procedures and results, including training, supports, and re-training
Testing	Highly likely (MB designs only); less likely for MP designs	Visual analysis: Deteriorating or therapeutic trends in BL conditions, especially for later tiers	Use an MP design; design non-aversive BL conditions; continue condition until data are stable for all tiers; collect data less frequently (MP design only)	Describe likely testing threats and solutions used
Attrition Bias	May be likely due to extended duration of MP and MB designs, particularly in "across participants" variations because one remains in baseline conditions for extended periods of time *and* when they are not randomly assigned to tiers (MB or MP across participants only)	Author report	Use an MP design; clearly describe to participants likelihood of extended baseline; design non-aversive BL conditions; randomly assign participants to tiers (across participants only)	Describe attrition in written report and report all data from all participants, even if design was not completed; describe methods for choosing participants who were included.
Sampling Bias	Likely when multiple available participants meet inclusion criteria but only some are included	Reliant on author description	Randomly choose from all available participants who are eligible for participation or include all eligible participants	Describe number of participants who met criteria given your constraints (e.g., in a specific setting); if all not chosen

Note: MB=multiple baseline, MP=multiple probe, BL=baseline.

	Multiple Baseline and Multiple Probe Designs			
	Likelihood	**Detect**	**Control**	**Report**
Adaptation	Likely when observations are apparent	Participant behavior changes over time during BL conditions	Continue BL until data are stable	Describe anecdotal evidence that BL change was due to adaptation; discuss degree to which later BL data are potentially more representative of "typical" behavior
Hawthorne Effect	Likely when participants are sensitive to perceived desirable behaviors	Participant behavior is inconsistent with expectations when study begins	Use covert measurement; do not implement intervention condition if BL data do not indicate need; continue data collection to determine whether effect is temporary and change conditions only when behavior are stable	Describe anecdotal evidence that BL change was due to Hawthorne effect; discuss degree to which later BL data are potentially more representative of "typical" behavior
Multitreatment Interference	Likely in form of carryover effects when participants cannot identify differences between conditions	Visual analysis: Delayed change in behavior when new condition is implemented	Continue all conditions until data are stable; do not intervene in later tiers until data stability are achieved during intervention for earlier tiers	Discuss potential presence of carryover effects when presenting results of visual analysis; describe extent to which delays were expected and consistent
Instability	Likely to be a problem in multi-tier designs because data analysis occurs simultaneously for multiple tiers	Visual analysis: Changes in y value of data that reduce ability to predict value of next data point given no condition change *and* variability in non-treated tiers that align with condition changes in other tiers	Change conditions only after data are stable in all tiers	Describe degree to which data instability within conditions impacted conclusions due to uncertainty regarding between-condition changes; describe rules used for changing conditions
Covariation	Likely when behaviors in each tier are not functionally independent	Visual analysis: Changes in the y value of untreated tiers that correspond with treatment initiation in previous tiers	Choose behaviors, contexts, or participants unlikely to change unless intervention is directly applied	Describe degree to which covariation between tiers impacted conclusions; describe rules used for changing conditions
Inconsistent Effects	Likely when behaviors in each tier are not functionally similar	Visual analysis: Changes in tiers are different in magnitude or behavior does not change in some tiers	Choose behaviors, contexts, or participants likely to respond to the same intervention	Describe degree to which inconsistent effects between tiers impacted conclusions; describe rules used for changing conditions

Note: MB=multiple baseline, MP=multiple probe, BL=baseline.

There are no design-specific concerns when detecting and controlling for instrumentation and fidelity threats; typical procedures for detecting and controlling for these threats should be used (see Chapter 1 and Table 10.1).

History threats are controlled for in MP and MB designs when (a) within-condition data are stable and (b) consistent between-condition differences are demonstrated. If a potential history threat occurs *in any tier*, you should avoid changing conditions *in all tiers* until data are stable. In MP designs, history threats are particularly problematic because detecting these threats is more difficult with intermittent data collection.

Several threats are likely in MB and MP designs because they tend to be longer in duration (e.g., include about twice as many sessions, on average, than A-B-A-B designs; Ledford, Severini, Zimmerman, & Barton, 2017). One example is maturation threats; these are especially problematic for behaviors, contexts, or participants assigned to later tiers. Maturation threats can be minimized by using MB designs only for behaviors that are unlikely to gradually improve in the absence of intervention. If maturation effects are likely, consider using a different design. Similarly, testing and attrition threats are more likely in later tiers of MB and MP designs due to the extended nature of the baseline condition. We will discuss testing threats specific to MB and MP variations later in the chapter. Finally, attrition bias is more likely for studies that are longer in duration, and especially when baseline conditions are extended; thus, this threat can be problematic in MB and MP designs.

In addition, the need to analyze three or more sets of data collectively increases the likelihood of some additional threats to internal validity, including data instability. This threat is critical because data in *all tiers* should be stable before condition changes occur; instability in any tier may impact experimental control. Finally, sampling bias is more likely in some MB and MP variations (described later in the chapter) due to the tiered nature of the design.

Threats Specific to MB and MP Designs

To demonstrate experimental control using MB and MP designs, you must make two predictions prior to initiating your research. First, you must identify behaviors (or contexts or participants) that are **functionally independent**. When behaviors are functionally independent, the introduction of the independent variable to one tier (behavior, context, or participant) will not bring about a change in other untreated tiers of the design. The behaviors, contexts, and participants should also be **functionally similar**. When behaviors are functionally similar, the independent variable is likely to have the same or similar effect on each tier. Should either of these predictions be incorrect, experimental control may be lost. When behaviors are not functionally independent, **behavioral covariation** may occur, in tiers not yet exposed to the independent variable, resulting in an ambiguous demonstration of effect. That is, when you begin intervention in the first tier, behavior changes in two or more tiers. Two questions are left unanswered: 1) Was the intervention effective in the first tier, with

effects generalizing to the unexposed tiers? (response generalization) or 2) Was the intervention ineffective, with covarying effects due instead to history, maturation, instrumentation, or testing confounding? In the second case (lack of functional similarity), there may be therapeutic effects in one or more tiers, but no effects in others. Again, you are left with an unconvincing demonstration of experimental control, with the intervention appearing to work in one or a few instances, but not in others (**inconsistent intervention effects**). One possibility is that you chose dissimilar behaviors; the alternative explanation is that your intervention was ineffective and the tier(s) showing positive effects were the result of a history or instrumentation threat rather than your intervention. Figure 10.1 shows MB and MP designs without behavioral covariation or inconsistent effects. Figure 10.2 shows MB designs with behavioral covariation across unexposed tiers (left) and inconsistent effects (right). Strategies to improve predictions and therefore minimize these risks are addressed as each design is described later in the chapter.

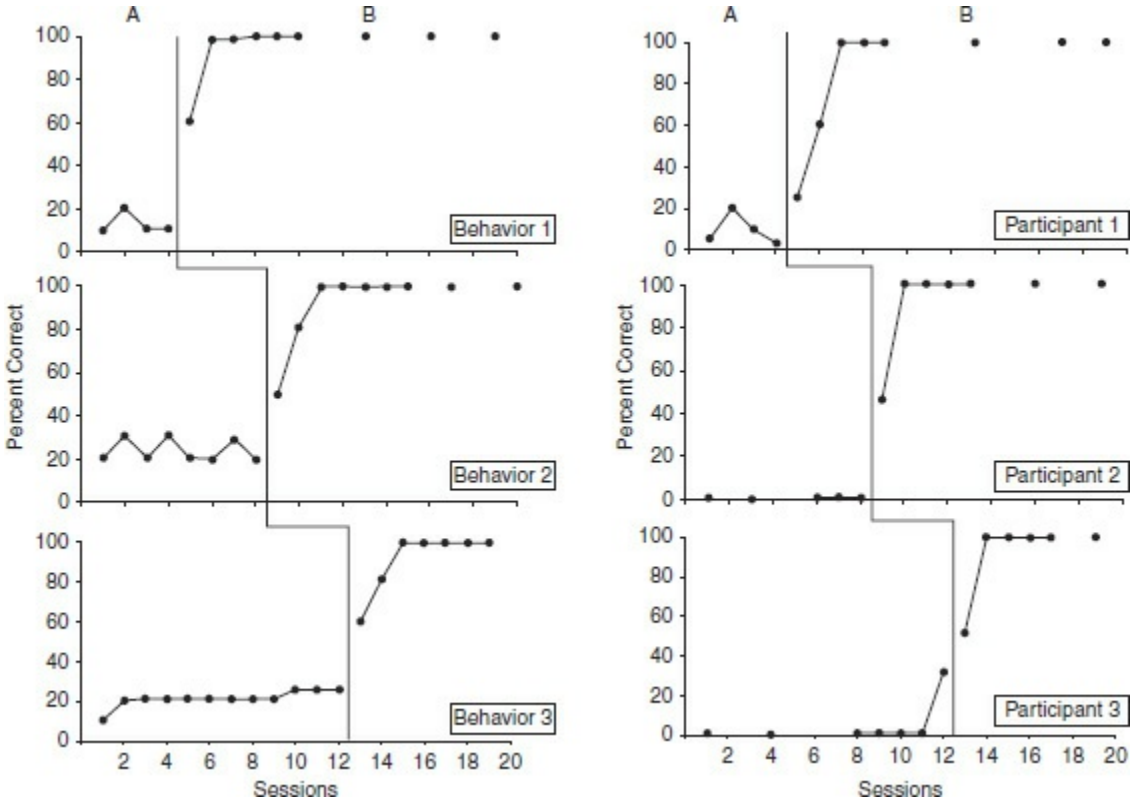


Figure 10.1 Hypothetical data within a multiple baseline (left) and multiple probe (right) design.

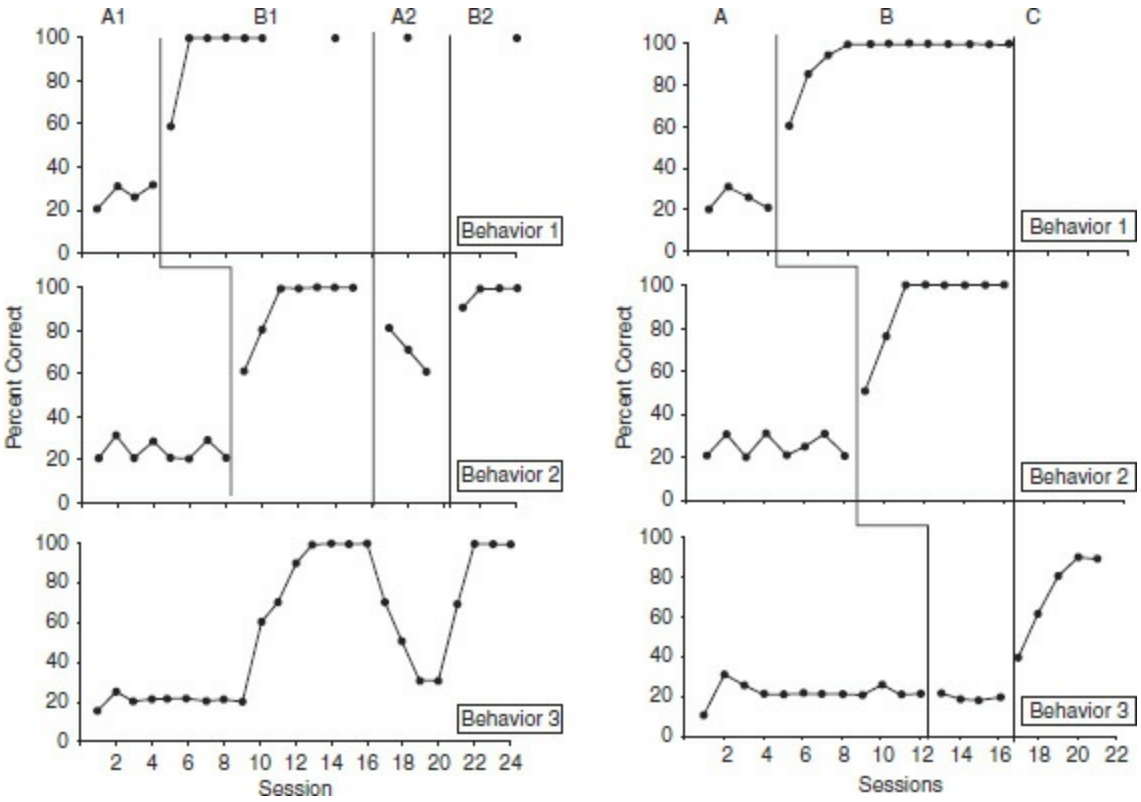


Figure 10.2 Hypothetical data for a multiple baseline design across behaviors demonstrating behavioral covariation of Behavior 3, after the intervention is sequentially applied to Behaviors 1 and 2, followed by withdrawal and reintroduction of B (left pane). Hypothetical data for a multiple baseline design across behaviors demonstrating a failure of intervention effects for Behavior 3, requiring the addition of a new intervention (left pane).

During baseline conditions, you should assess for stability in level and trend prior to introducing the intervention to the first tier. Similarly, you should introduce the intervention to the second tier only after a therapeutic change is demonstrated in the first tier *and* baseline data in all tiers remain stable. This process should be repeated for all remaining tiers. Generally, a functional relation is established when data show an immediate change in level and/or trend direction following introduction of the intervention to each tier, and not before (see Figure 10.1). If the change in behavior is delayed, the demonstration of experimental control is less clear (Lieberman, Yoder, Reichow, & Wolery, 2010). A simply stated logic to evaluate experimental control is, "where intervention is applied, change occurs; where it is not, change does not occur" (Horner & Baer, 1978, p. 189).

A believable demonstration of a functional relation between intervention and behavior change occurs when the effect is replicated across tiers (Baer et al., 1968). Determination of the appropriate number of replications required for believability is complicated by such factors as trend and level stability of the data series as well as the rate and magnitude of change upon each sequential introduction of the intervention (Kratochwill, 1978). However, provided there are reliable replications of effect, three or four tiers are generally sufficient (Barlow & Hersen, 1984; Kazdin & Kopel, 1975; Tawney & Gast, 1984; Wolf & Risley, 1971).

Given the differences between MB and MP designs, why would a researcher choose one instead of the other? MB designs have the benefit of continuous measurement prior to and during intervention, thereby allowing day-to-day data analyses and decisions. Moreover, continuous data collection allows for close visual analysis of potential threats to internal validity such as maturation and instrumentation. As for MP designs, intermittent baseline measurement means data analysis is also intermittent, limiting the ability to identify potential threats. However, some threats are more likely with MB designs than MP designs (see Table 10.1), including testing threats. In addition, extended baselines with continuous measurement may be less desirable from a practical standpoint (e.g., participants may find these sessions aversive). Sometimes both types of designs are possible given resource constraints and researcher goals. If so, we offer the following guidance for choosing MB or MP designs:

1. When *testing threats* are more likely, choose an MP design. This includes most situations in which baseline conditions consist of adult-directed trials to complete a specific task.
2. When *data instability* is more likely, choose an MB design. This includes most free operant behaviors, which tend to be more variable than trial-based behaviors.
3. If neither threat is likely, choose an MB design because continuous measurement generally allows for closer inspection of potential threats than intermittent measurement.

MP Designs

Probe Terminology

MP designs do not require *continuous* measurement of all behaviors, conditions, or participants prior to the introduction of the independent variable, as is the case with MB designs. Probe trials may occur once daily or several probe trials may be clustered and presented over a short period of time in what is referred to as a **probe session**. Several probe sessions may in turn be conducted over 3 or more consecutive days using pre-intervention procedures during what is referred to as a *probe condition*. A **probe condition** differs from a baseline condition *only* in that probe conditions do not occur for the entire duration of pre-intervention for each tier. The frequency with which data are collected after criterion has no bearing on whether a design is an MB or MP design. With both designs you may choose to monitor performance continuously or intermittently following the intervention condition (i.e., the type of design, MB or MP, is based solely on the frequency of data collection *prior* to introducing the intervention).

Variations

There are two primary variations of the MP design: (a) one in which data are collected periodically for *single sessions*, and over a minimum of 3 days immediately prior to introduction of the independent variable, which we refer to as the **days variation** or the **MP design (days)**; and (b) one in which data are collected for 3 or more *consecutive sessions*, which we refer to as the **conditions variation** or the **MP design (conditions)**. In other words, in the *days* variation (see Figure 10.1), probe sessions occur intermittently as single measurement occasions. In the *conditions* variation, probe sessions occur intermittently, but in clusters of sessions that comprise a condition. Both variations of the MP design require data to be collected on *all* tiers at the start of the study, ideally all on the first session, but certainly by the third session, regardless of the design type (i.e., across behaviors, conditions, participants). The guidelines for MP designs are identical to those for MB designs except for the frequency with which pre-intervention data are collected. The right and left panels of Figure 10.3 show data from an MP design (*conditions*), presented in two different formats—one highlighting each probe condition and one highlighting the time-lagged introduction; either graph is acceptable. Note that these variations (days and conditions) are not typically identified in research reports (e.g., authors do not report which variation is used) although it should be clear based on the graphic presentation of data. We discuss the two types separately here because the procedures for use are somewhat different.

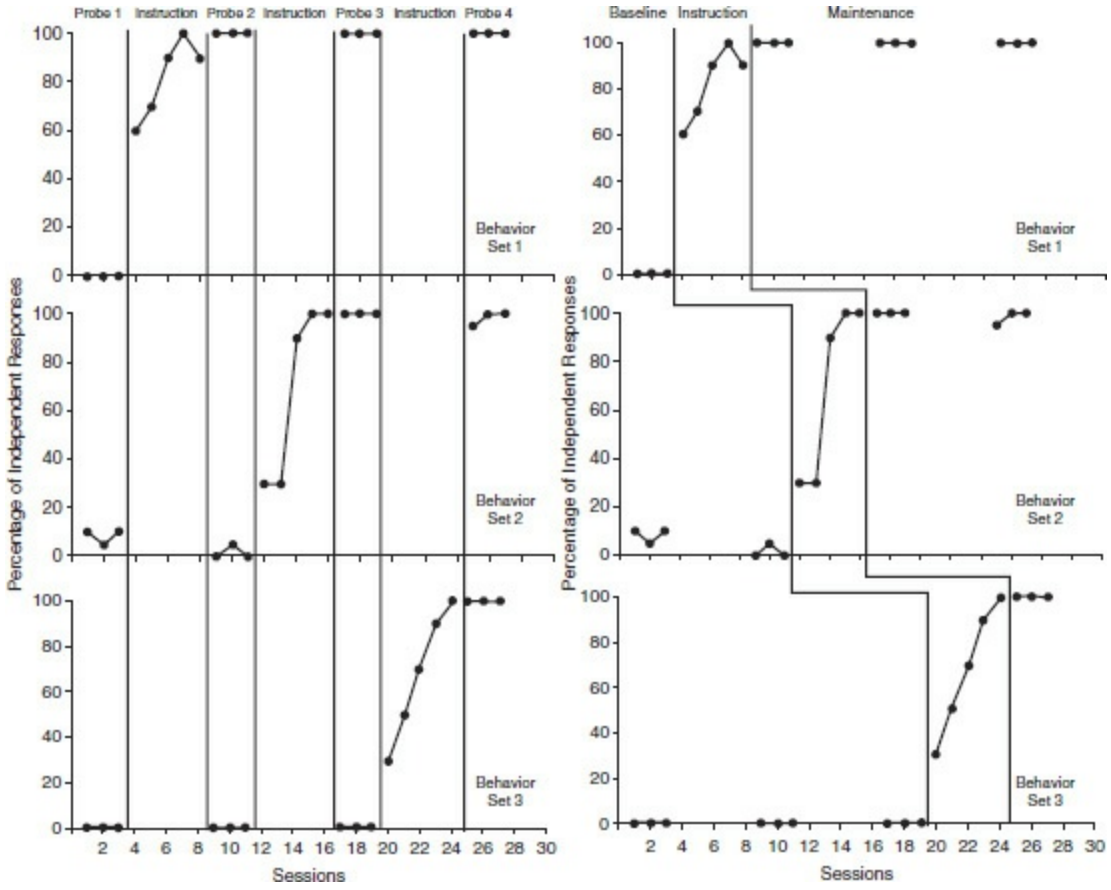


Figure 10.3 Two variations for presenting the same data within a multiple probe design (conditions) across behaviors. On the left (the traditional presentation), probe conditions are separated with vertical lines. On the right, the same data are only separated by the time-lagged introduction of intervention (e.g., for Tier 3, Probe 1, Probe 2, and Probe 3 are not separated because they are all pre-intervention probes).

MP (days)

The MP design (*days*) was first described by Horner and Baer (1978). Horner and Baer, and others (Cooper, 1981; Murphey & Bryan, 1980; Tawney & Gast, 1984) have recommended that intermittent probe data be collected as an alternative to "unnecessary" continuous baseline measures (i.e., when testing threats are likely and change in behavior is not). In these cases an MP design can serve as a practical alternative to an MB design. In addition to the guidelines presented earlier in the chapter, two additional steps are required when using MP designs:

1. Determine, a priori, how often data will be collected prior to the intervention. Data should be collected at least once every five days. Data should also be collected continuously for at least three sessions immediately prior to intervention.
2. When the intervention is introduced to the first tier, continue collecting data on other tiers at the previously determined frequency, but preferably also immediately after intervention implementation is initiated in previous tiers to assess potential

covariation. If variability exists, data should be collected more frequently.

MP (conditions)

The MP design (*conditions*) differs from the MP design (*days*) in terms of when pre-intervention data are collected. In MP design (*conditions*), a series of *consecutive probe sessions* (or observational days) are introduced at scheduled times over the course of the study. This particular MP design variation is well-suited for practitioner-paced instruction when a number of stimuli or behaviors are grouped together and taught across three or more tiers. As shown in Figure 10.3, researchers initially assess all behaviors in a probe condition, labeled Probe 1 (again, probe conditions are synonymous with baseline conditions except that they do not occur continuously). During Probe 1 all behaviors being tested are intermixed and presented to a participant in a single session, after which participant's responses are separated, according to the tier to which they were assigned, and the percentage correct for each tier during that session is graphed. A minimum of 3 consecutive sessions over 2 days (or until data are stable) should be conducted before introducing the independent variable to behaviors assigned to the first tier. Once criterion is reached on the first tier, a second probe condition, Probe 2, is conducted and is procedurally identical to Probe 1. This alternating sequence of probe condition and intervention condition, staggered across tiers, continues until the independent variable has been introduced in all tiers. Typically there is a final probe condition after all tiers have reached criterion. It should be noted that later probe conditions serve as both a baseline condition (for behaviors not yet introduced to the intervention) and a maintenance condition (for tiers in which criterion has been reached). For example, Probe 2 in Figure 10.3 serves as a maintenance condition for behaviors assigned to Tier 1 and as baseline data for behaviors assigned to Tiers 2 and 3.

MB and MP Designs Across Behaviors

In the sections that follow, we discuss each MB and MP design separately. First we discuss MB and MP designs across behaviors, then across contexts, and finally across participants. Though these designs are based on the same baseline logic, each has different advantages and disadvantages that you should consider prior to designing your study or evaluating the study of another researcher.

MB and MP designs across behaviors are both widely used; MB designs across behaviors are typically used to assess treatments designed to improve desirable behaviors (e.g., behaviors that should increase in level during treatment conditions; Ledford et al., 2017), and are more appropriate for free-operant than trial-based behaviors. MP designs across behaviors are also typically used to assess treatments designed to improve desirable behaviors (Ledford et al., 2017), and are more appropriate for academic or other non-reversible, trial-based behaviors. When MP designs are used to assess interventions to improve non-reversible, trial-based, behaviors, *sets of behaviors* should be targeted rather than single behaviors. For example, if teaching a child to name letter sounds, you might assign four letters to the first tier, four different letters to the second tier, and four additional letters to the third tier. Thus, the skill (naming letter sounds) is the same across tiers, but the actual behaviors (specific letter sounds) are different, and are taught at least two at a time. This is done for practical rather than experimental purposes; intermixing targets allows you to ensure the child is attending to relevant stimulus features (Doyle, Wolery, Ault, & Gast, 1989; Grow, Carr, Kodak, Jostad, & Kisamore, 2011). Table 10.2 and Table 10.3 summarize several applied research studies in which an MB or MP design across behaviors, respectively, was used to evaluate experimental control.

Table 10.2 *Studies Using Multiple Baseline Designs Across Behaviors*

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
Ganz, J. B., Heath, A. K., Lund, E. M., Carmago, S. P. H., Rispoli, M. J., Boles, M., & Plaisance, L. (2012). Effects of peer-mediated implementation of visual scripts in middle school. *Behavior Modification*, *36*, 378–398.	Number: 1 Sex: F Age: 15 Disability/Diagnosis: ID, autism, SI	Setting: Middle-school cooking lab Arrangement: Individual	Peer-prompted script intervention	Percentage of communicative responses (questions, praise, requests for help)
Hanley, N. M., & Tiger, J. H. (2012). Teaching coin discrimination to children with visual impairments. *Journal of Applied Behavior Analysis*, *45*, 167–172.	Number: 2 Sex: 1 M, 1 F Age: 6–8 Disability/Diagnosis: VI (2), DD (1)	Setting: Empty therapy room or classroom at state school for children with visual impairments Arrangement: Individual	Errorless training procedure for teaching coin relations	Percentage of trials with correct responses
Johnson, J., McDonnell, J., Holwarth, V., & Hunter, K. (2004). The efficacy of embedded instruction for students with developmental disabilities enrolled in general education classes. *Journal of Positive Behavior Interventions*, *6*, 214–227.	Number: 3 Sex: 1 M, 2 F Age: 7–9 Disability/diagnosis: Moderate ID (2), autism (1)	Setting: General education Arrangement: Individual	Embedded instruction	Percentage correct and rate of acquisition
Marckel, J. M., Neef, N. A., & Ferreri, S. J. (2006). A preliminary analysis of teaching improvisation with the picture exchange communication system to children with autism. *Journal of Applied Behavior Analysis*, *39*, 109–115.	Number: 2 Sex: M Age: 4–5 Disability/Diagnosis: Autism	Setting: Clinic Arrangement: Individual	Individual PECS instruction with prompt fading	Number of independent requests with improvisation
Westerlund, D., Granucci, E. A., Gamache, P., Clark, H. B. (2006). Effects of peer mentors on work-related performance of adolescents with behavioral and/or learning disabilities. *Journal of Positive Behavior Interventions*, *8*, 244–251.	Number: 4 Sex: F Age: 16–18 Disability/Diagnosis: Emotional or learning disabilities	Setting: Vocational training site Arrangement: Individual	Peer mentor training	Percent of steps of hairstyling routine performed correctly
Youmans, G., Youmans, S. R., & Hancock, A. B. (2011). Script training treatment for adults with apraxia of speech. *American Journal of Speech—Language Pathology*, *20*, 23–37.	Number: 3 Sex: 1 M, 2 F Age: 40–81 Disability/Diagnosis: Apraxia of speech	Setting: Clinic Arrangement: Individual	Script training procedure	Percentage of script words produced correctly, errors, and words per minute

Note: ID=intellectual disability, SI=speech impairment, VI=visual impairment, DD=developmental delay, M=male, F=female

Table 10.3 *Studies Using Multiple Probe Designs Across Behaviors*

Reference	Participants	Setting/Arrangement	Independent Variable	Dependent Variable
Alberto, P. A., Fredrick, L., Hughes, M., McIntosh, L., Cihak, D. (2007). Components of visual literacy: Teaching logos. *Focus on Autism and Developmental Disabilities, 22*, 234–243.	Number: 6 Sex: 3 M, 3 F Age: 9–14 Disability/ Diagnosis: Moderate to severe ID	Setting: Self-contained classroom Arrangement: Individual	CTD instruction	Percent correct (naming logos)
Jimenez, B. A., Browder, D. M., Spooner, F., & DiBiase, W. (2012). Inclusive inquiry science using peer-mediated embedded instruction for students with moderate intellectual disability. *Exceptional Children, 78*, 301–317.	Number: 5 Sex: 3 M, 2 F Age: 11–14 Disability/ Diagnosis: Moderate ID	Setting: General education science classroom Arrangement: Small groups of 4–5 students	Peer-mediated time-delay instruction and use of a knowledge chart	Number of correct science responses

Reference	Participants	Setting/Arrangement	Independent Variable	Dependent Variable
Mechling, L., Ayres, K. M., Purrazzella, K., & Purrazzella, K. (2012). Evaluation of the performance of fine and gross motor skills within multi—step tasks by adults with moderate intellectual disability when using video models. *Journal of Developmental and Physical Disabilities*, *24*, 469–486.	Number: 4 Sex: M Age: 29–35 Disability/ Diagnosis: Down syndrome and Moderate ID	Setting: Classroom in post school Compensatory Education Program for adults with disabilities Arrangement: Individual	Video modeling procedure	Percent correct responses (steps from task analysis for daily living task)
Werts, M. G., Caldwell, N. K., & Wolery, M. (2003). Instructive feedback: Effects of a presentation variable. *The Journal of Special Education*, *37*, 124–133.	Number: 4 Sex: M Age: 11 Disability/ Diagnosis: LD (3), Mild ID (1)	Setting: Self-contained classroom Arrangement: Small group	CTD instruction	Percent correct responses (naming words)
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 Education*, *22*, 14–25.	Number: 3 Sex: M Age: 5–7 Disability/ Diagnosis: Speech delay (1), behavior problems (1), ADHD (1)	Setting: Inclusive summer camp Arrangement: Individual (embedded in large group)	CTD instruction	Percent correct responses (naming words or multiplication facts)
Yanardag, M., Akmanoglu, N., & Yilmaz, I. (2013). The effectiveness of video prompting on teaching aquatic play skills for children with autism. *Disability & Rehabilitation*, *35*, 47–56.	Number: 3 Sex: 2 M, 1 F Age: 6–8 Disability/ Diagnosis: Autism	Setting: Indoor swimming pool Arrangement: Individual	Video prompting procedure	Percentage of correct steps from task analysis per skill

Note: ID=intellectual disability, LD=learning disability, ADHD=attention deficit hyperactivity disorder, M=male, F=female, CTD=constant time delay

Procedural Guidelines

When using an MB or MP design across behaviors, adhere to the following guidelines:

1. Identify and define a minimum of three similar yet functionally independent behaviors, or sets of behaviors, emitted by one individual.
2. Select a sensitive, reliable, valid, and feasible data collection system and pilot the system and your behavior definitions.
3. Prior to the start of the study, identify a criterion for introduction of the intervention. For the initial introduction, stability in all tiers is an appropriate, conservative criterion. For remaining tiers, you may set a criterion level (e.g., 90% correct or better for 3 consecutive sessions) or a visual analysis criterion (e.g.,

following a clear change in level, with at least 3 consecutive sessions with no overlapping data with baseline; see Chapter 8).

4. Prior to the start of the study, identify the method by which you will assign interventions to tiers. For some interventions, there is a reasonable therapeutic sequence (cf. Roberts, Kaiser, Wolfe, Bryant, & Spidalieri, 2014); for others, it is reasonable to randomly assign behaviors to tiers (see Chapter 13 for more information about randomization).
5. Determine a priori frequency of reliability and fidelity data collection (e.g., 33% of sessions for each condition), and conduct data collection for the duration of the study.
6. Concurrently collect baseline (or probe) data for all tiers.
7. When data in all tiers are stable, intervene on behaviors assigned to the first tier, while monitoring other behaviors under the pre-intervention condition. The remaining steps are dependent on whether you have chosen to use an MB design, or the conditions or days variation of the MP design, as outlined below.

MB	*MP (days)*	*MP (conditions)*
8. When data in the 1st tier have met your criterion (usually behavior change indicated via visual analysis) and data in all tiers are stable, begin intervention in the 2nd tier.	8. When data in the 1st tier approach your criterion, ensure that you have three consecutive days of data collected for behaviors assigned to the 2nd tier.	8. When data in the 1st tier have reached your criterion (usually an a priori mastery criterion), discontinue instruction and begin a second probe condition.
9. When data in the 2nd tier have met your criterion and data in all tiers are stable (see above), begin intervention in the 3rd tier.	9. When data are stable in all tiers, begin intervention for behaviors assigned to the 2nd tier.	9. When data in all tiers are stable in the probe condition, begin intervention for behaviors assigned to the 2nd tier.
10. Repeat steps 8 and 9 for remaining tiers.	10. Repeat steps 8 and 9 for remaining tiers.	10. Repeat steps 8 and 9 for remaining tiers.

Internal Validity

Behavioral covariation is somewhat likely in the across behaviors variation of MB and MP designs; care should be taken to select targets that are independent and functionally similar. For example, if you teach a participant to name four letter sounds assigned to Tier 1, doing so is unlikely to lead to their learning the four sounds assigned to Tier 2

and the four sounds assigned to Tier 3. However, if you teach a participant to read words in Tier 1, and those words are similar to the words in Tier 2, some covariation might occur. This may be due to baseline levels of letter sounds knowledge by the participant or specific instruction during the intervention. To assess for likely covariation, you may review previous research on related dependent variables. If previous studies show covariation, you should choose a different design variation. You can also test for potential covariation by teaching behavior sets sequentially to non-participants to determine whether those pilot participants learn from previous instruction.

With social behaviors, the prediction of independence becomes more problematic. For example, beyond certain thresholds, some behaviors may occasion others, as when a person is taught an appropriate greeting behavior (e.g., saying hello to a familiar adult) that, in turn, may increase the number of social opportunities for the participant, perhaps resulting in changes in other desirable behaviors. In such a case, an MB or MP design across social behaviors may result in response generalization, precluding a demonstration of experimental control due to behavioral covariation across the three target behaviors. When there is concern about behavioral covariation, consider identifying more than the minimum number of behaviors (i.e., three), using a different research design, or combining research designs (e.g., MB design across behaviors with an MP design across participants; Cronin & Cuvo, 1979; see Chapter 12 for a discussion of combination designs). Similarly, the deceleration of some behaviors may result in an acceleration of other inappropriate behaviors if the net result, otherwise, would be a lower overall frequency of reinforcement (Bandura, 1969). For example, a strategy intended to sequentially suppress spitting, hitting, and object throwing might become progressively less effective for each subsequent behavior, unless the intervention provides alternative avenues for the learner to recruit reinforcement (i.e., differential reinforcement for appropriate behaviors). On the other hand, differential reinforcement for appropriate behaviors may result in behavioral covariation if all challenging behaviors are likely to decrease simultaneously. If, however, behaviors are reversible, you may attempt to salvage experimental control by briefly withdrawing the intervention, as in an A-B-A-B design. Finally, if you select behaviors that are too topographically or operationally different (e.g., spelling words, adding numbers, reciting a poem), you risk not demonstrating experimental control due to the differences between stimuli or responses associated with each task (i.e., the behaviors would not be functionally similar).

In a multiple baseline across behaviors design, testing threats may be likely due to the extended baseline condition. If testing threats are likely, you should use a multiple probe design. However, with free-operant behaviors, it may be feasible and prudent to measure occurrence continuously because those behaviors tend to be more variable and thus continuous data collection will improve your ability to detect threats.

Advantages

MB and MP designs across behaviors offer several advantages. First, both designs permit an evaluation and demonstration of intra-participant direct replication. When multiple participants are included in a study, both intra-and inter-participant replications are possible. Second, the MP design across behaviors provides a practical means for evaluating programs designed to teach academic and functional skills that are nonreversible once acquired (e.g., spelling, self-care) and the MB design across behaviors provides a reasonable method for evaluating programs designed to improve social behaviors that are difficult to establish and would be inappropriate to reverse (e.g., greeting responses, asking questions). Third, the conditions variation of the MP design across behaviors provides a paradigm for repeatedly monitoring progress over time, a practical benefit for practitioners (e.g., later probe conditions serve as maintenance assessments for previously-treated tiers). Finally, the MP and MB designs across behaviors allow researchers to begin treatment on one behavior set after a relatively short baseline condition. This differs from MP and MB designs across participants, in which one participant experiences a much longer baseline condition.

Limitations

MB and MP designs across behaviors require adherence to specific guidelines that may present problems under some circumstances. First, for each participant a minimum of three behaviors (or sets of behaviors) must be identified, each of which is independent of the others yet responsive to the same independent variable. This may be difficult when teaching sets of behaviors that are likely to covary (e.g., academic skills that build on previously-taught behaviors). Second, all behaviors must be monitored repeatedly and concurrently, which may prove time-consuming, distracting, cumbersome, or otherwise impractical. This potential limitation is generally more problematic for MB designs due to continuous measurement in baseline. However, MB designs across behaviors do allow a child to receive instruction on some behaviors or behavior sets throughout a study, making it more practical than MB and MP designs across participants.

Applied Example 10–1: MP Design (Days) Across Behaviors

Flores, M. M., & Ganz, J. B. (2007). Effectiveness of direct instruction for teaching statement inference, use of facts, and analogies to students with developmental disabilities and reading delays. *Focus on Autism and Other Developmental Disabilities*, *22*, 244–251.

The effectiveness of Direct Instruction (DI) for teaching three reading comprehension behaviors (statement inference, use of facts, and analogies) to children with developmental disabilities was studied by Flores and Ganz (2007). The study took place at a private school, in a self-contained classroom. Participants were 5th- and 6th-grade students (age range: 10–14). Two students were diagnosed with autism, one student was diagnosed with mild ID, and one student was diagnosed with ADHD. The two students with autism had average or low-average decoding skills, but significantly below-average comprehension skills. The student with ID had decoding and comprehension skills that were significantly below average, and the student with ADHD had decoding and comprehension skills in the low-average range.

The dependent variable was the percentage of correct responses during probe, instruction, and maintenance conditions. Data were collected using an event recording procedure in which children's responses were scored as correct if they emitted an appropriate oral response to a teacher-delivered question, or incorrect if the child inappropriately responded to the question. Procedural reliability data were collected once per week (20% of sessions) using a checklist of teacher behaviors from the DI program, and fidelity was 100%. The point-by-point method was used to calculate interobserver agreement, which ranged from 96%–100% (mean = 98%).

The independent variable was the DI program that was used to teach three comprehension skills in the context of an MP design (days) across behaviors. During statement inference sessions, students were taught to respond to questions related to a statement that was read by the teacher. In DI sessions that focused on using facts, the instructor read two facts followed by a series of scenarios. Students were asked to name those facts that explained why the event happened. In analogy instructional sessions, students were asked to complete simple analogies like, "A rake is to leaves as a shovel is to what?" Sessions occurred for approximately 20 minutes per day and were conducted by one of two researchers who were not classroom teachers. Scripts included modeling the skill, "leading" as students demonstrated the skill, and asking students to perform the behavior independently. The instructor followed DI procedures of choral responding, individual responding,

and error correction. Intermittent probe sessions were conducted in a 1:1 arrangement while DI sessions were conducted daily in a small group. Maintenance sessions were conducted one month after the final instructional session.

Figure 10.4 shows the percentage of correct responses for two students evaluated within the context of an MP design (days) across behaviors. Prior to DI instruction, all students had low percentages of correct responding on each comprehension skill. Visual analysis shows that for each skill, levels of independent correct responding changed from a stable zero-celerating trend during the probe condition (which they label "Baseline") to a stable accelerating therapeutic trend immediately upon introduction of the DI condition. This effect was replicated across the three comprehension skills for each of the four students, with each student reaching 100% correct responding. All students maintained criterion-level performance on the three reading comprehension skills during their maintenance session one month later.

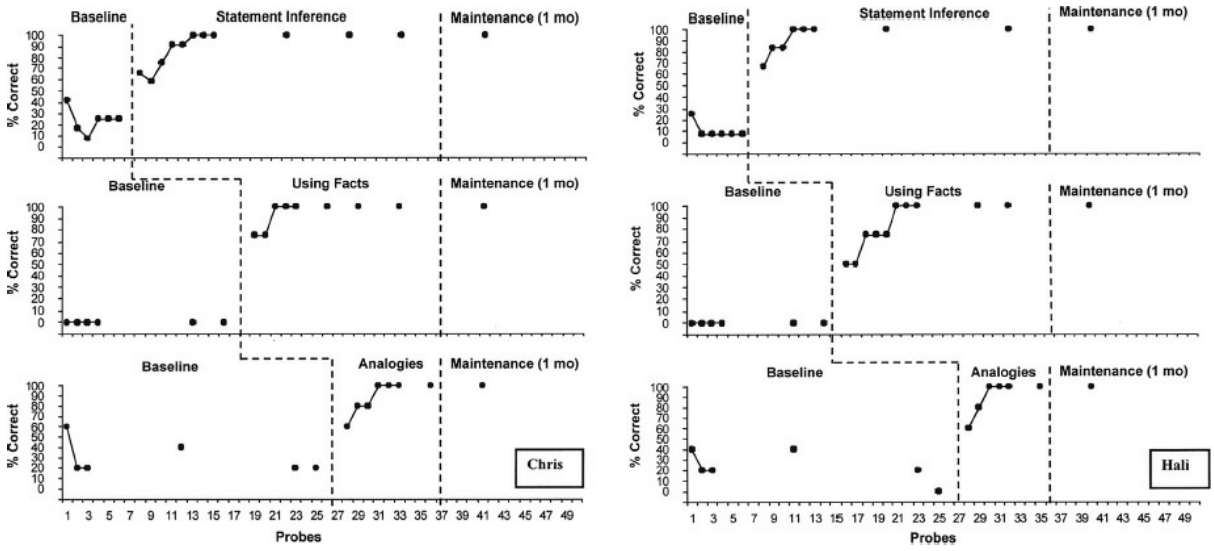


Figure 10.4 Multiple probe design (days variation) across behaviors for two participants.

Source: Flores, M. M., & Ganz, J. B. (2007). Effectiveness of direct instruction for teaching statement inference, use of facts, and analogies to students with developmental disabilities and reading delays. *Focus on Autism and Other Developmental Disabilities*, *22*, 244–251.

Third, prolonged baseline conditions can result in *testing threats to internal validity*. Either facilitative or inhibitive effects are possible depending on baseline (or probe) condition procedures. Repeated testing may result in a *facilitative effect* in which a participant's performance improves over time due to response consequences (e.g., differential reinforcement of correct and incorrect responses), observation of others, or independent research (e.g., looking something up on the internet). An *inhibitive effect* may occur due to response consequences (e.g., lack of reinforcement), fatigue due to extended session durations or number of trials, or task difficulty. Several strategies may be employed to overcome these potential outcomes of prolonged baseline or probe

sessions. First, you can positively reinforce desired behaviors during pre-intervention sessions. You may choose to (a) contingently reinforce target behaviors when performed correctly, assuming you are not interested in studying the influence of contingent reinforcement alone (Wolery, Cybriwsky, Gast, & Boyle-Gast, 1991); (b) contingently reinforce correct responses to known stimuli interspersed with target behaviors (Gast, Doyle, Wolery, Ault, & Baklarz, 1991); (c) intermittently reinforce non-target behaviors emitted between trials or between steps during a response chain task (Wall & Gast, 1999); or (d) inform study participants prior to the start of a session that a reinforcer menu will be presented immediately after the session from which they will be able choose one activity or item. Second, if sessions are too long, as indicated by a decrease in response attempts over trial presentations and time, or an increase in aberrant behavior, you may *shorten the session*. This can be done by dividing sessions into two shorter daily data collection periods, or by scheduling a break midway through the session.

MP designs have the practical benefit of not requiring extended periods of time in an assessment condition, particularly when it is highly unlikely for a participant to respond correctly prior to introduction of the independent variable. In such situations, intermittent assessments, rather than continuous assessments, will suffice in documenting behavior stability. Undetected response generalization is a potential limitation of MP designs.

Applied Example 10–2: MP Design (Conditions) Across Behaviors

Ledford, J. R., Gast, D. L., Luscre, D., & Ayres, K. M. (2008). Observational and incidental learning by children with autism during small group instruction. *Journal of Autism and Developmental Disorders*, *38*, 86–103.

In this study, acquisition of target behaviors as well as information presented as instructive feedback (incidental learning targets) and information presented to group mates (observational learning targets) was assessed for six children with autism served in a self- contained elementary public school classroom. Instruction on words commonly found on products and community signs (e.g., *poison, caution*) was delivered using a 3-second constant time delay (CTD) procedure in a small group (dyad) instructional arrangement. Six males diagnosed with autism and speech language impairments, in kindergarten through 2nd grade, participated in the study. Students were placed into dyads based on entry skills and previous reading instruction. A summary description of each of the experimental conditions is presented in Table 10.4.

The dependent variables were the percentage of target words read correctly, the percentage of target words taught to group mates that were read correctly (observational information), the percentage of associated visuals correctly identified (incidental target information; ITI), and the percentage of associated visuals assigned to group mates that were correctly identified (incidental observational information; IOI). Data were collected using an event recording procedure across all conditions. Reliability data were collected for 23%–50%) of all sessions across all conditions. Using the point-by-point method, inter-observer agreement (IOA) ranged from 92%–100% (mean: 99.4%) and procedural reliability agreement ranged from 90%–100% (mean: 99.7%).

Generalization (student's ability to read target and non-target product or community signs/pictures in novel settings) was assessed during pre- and post-test conditions. Other conditions were introduced within the context of an MP design (condition variation) across behaviors (word sets). Probe conditions were scheduled immediately prior to the introduction of the first CTD condition on a word set and immediately following a student reaching criteria on a word set. All probe condition sessions were conducted in a 1:1 arrangement, while instructional sessions were conducted in a small group arrangement (2 students and 1 teacher).Figure 10.5 displays the percentage of prompted (closed square) and unprompted correct (open triangle) responses for one student evaluated within the context of an MP design. Prior to instruction, all students identified 0% of target words. Visual analysis shows that levels of unprompted correct responses changed

from a stable trend at 0% to a therapeutic trend that increased to criterion level for all students when each word set was taught during CTD conditions. Percentage correct for words not yet introduced to CTD instruction remained stable at 0%. The mean number of sessions to criterion was 7 (range: 4–12) and the mean percentage of errors across students and word pairs was 3.6% (range: 0–10%). Each student maintained between 50% and 100% correct responding during post-instruction probes, and five of six participants maintained 100% correct responding for all target words during the final probe.

In addition to target information, all participants acquired some or all observational targets (words directly taught to their partner). All students also acquired some or all of the incidental target information (ITI) related to their target words (e.g., When presented with the picture of a yellow diamond, the student identified this as "caution" in the absence of the word).

Table 10.4 *Description of Conditions*

Generalization	Word probe	Picture probe	CTD
Point & "Look" (General Attentional Cue)	*"Look" (General Attentional Cue)*	*"Look" (General Attentional Cue)*	*"Look" (General Attentional Cue)*
	"Tell me the letters" (Specific Attentional Cue)		*"Tell me the letters" (Specific Attentional Cue)*
"What is this?"	*"What word?"*	*"What is this?"*	*"What word?"*
Unprompted Correct: Verbal Praise	*Unprompted Correct: Verbal praise, token*	*Unprompted Correct: Verbal praise, token*	*Unprompted Correct: Verbal praise, presentation of incidental information ("Right! Caution." & present picture)*
Unprompted Incorrect: Walk away from the sign	*Unprompted Incorrect: Remove written word*	*Unprompted Incorrect: Remove picture*	*Unprompted Incorrect: Remove written word ("Wait if you don't know")*

No Response: Walk away from the sign	*No Response: Remove written word*	*No Response: Remove picture*	*No response: Model prompt* *Prompted Correct: Verbal praise, presentation of incidental information ("Right! Caution" & present picture of caution sign)* *Prompted Incorrect: Ignore. Remove written word No response: Wait 3 s. Ignore & remove word*

From: Ledford, J. R., Gast, D. L., Luscre, D., & Ayres, K. M. (2008). Observational and incidental learning by children with autism during small group instruction. *Journal of Autism and Developmental Disorders*, *38*, 86–103.

Conclusions

MB and MP designs across behaviors can be used to evaluate experimental control with a wide range of interventions, in a variety of educational and clinical settings, and across many types of learners exhibiting a variety of behaviors. When compared to MB and MP designs across participants, these designs permit direct intra-participant replication, thereby increasing confidence in the findings. These designs are often the single case designs (SCDs) of choice for applied researchers when compared to A-B-A-B designs because the withdrawal of an effective intervention is not required to demonstrate experimental control but still only require a single participant. You must ensure that target behaviors are independent and functionally similar, and determine whether continuous or intermittent baseline data collection is more appropriate given the nature of the dependent variable. The potential for testing effects in prolonged baseline conditions is a limitation of MB designs; an undetected change in responding within or across tiers due to intermittent data collection is the primary limitation of MP designs. If there are practical concerns regarding the use of an MB design across behaviors, we

recommend that you consider an MP design across behaviors. Although experimental control is demonstrated for single participants when MB and MP designs across behaviors are used, we recommend that multiple participants be recruited to improve external validity.

MB and MP Designs Across Contexts

When using an MB or MP design across contexts to evaluate experimental control, you sequentially introduce the independent variable to the same behavior across several different stimulus conditions. *Stimulus conditions* can encompass the dimensions of time, instructional arrangement (individual, small group, independent), activity, setting, control agent (practitioner, parent), or composition of peer group; we refer to all of these variations as "contexts." In contrast to the MB and MP designs across behaviors, these designs require you to target a single behavior and a minimum of three different contexts in which you want the behavior to occur (or not occur, depending on the objective of the intervention). For example, contexts can range from monitoring percentage of time on-task across math, spelling, and social studies periods (across activities) to the frequency of disruptive behaviors in a classroom, cafeteria, and playground (across settings), to the number of minutes with active play across morning, lunch, and afternoon recesses (across time). MB and MP designs across contexts are typically used to evaluate interventions designed to increase reversible behaviors. Table 10.5 summarizes several applied research studies in which an MB or MP design across contexts was used to evaluate experimental control.

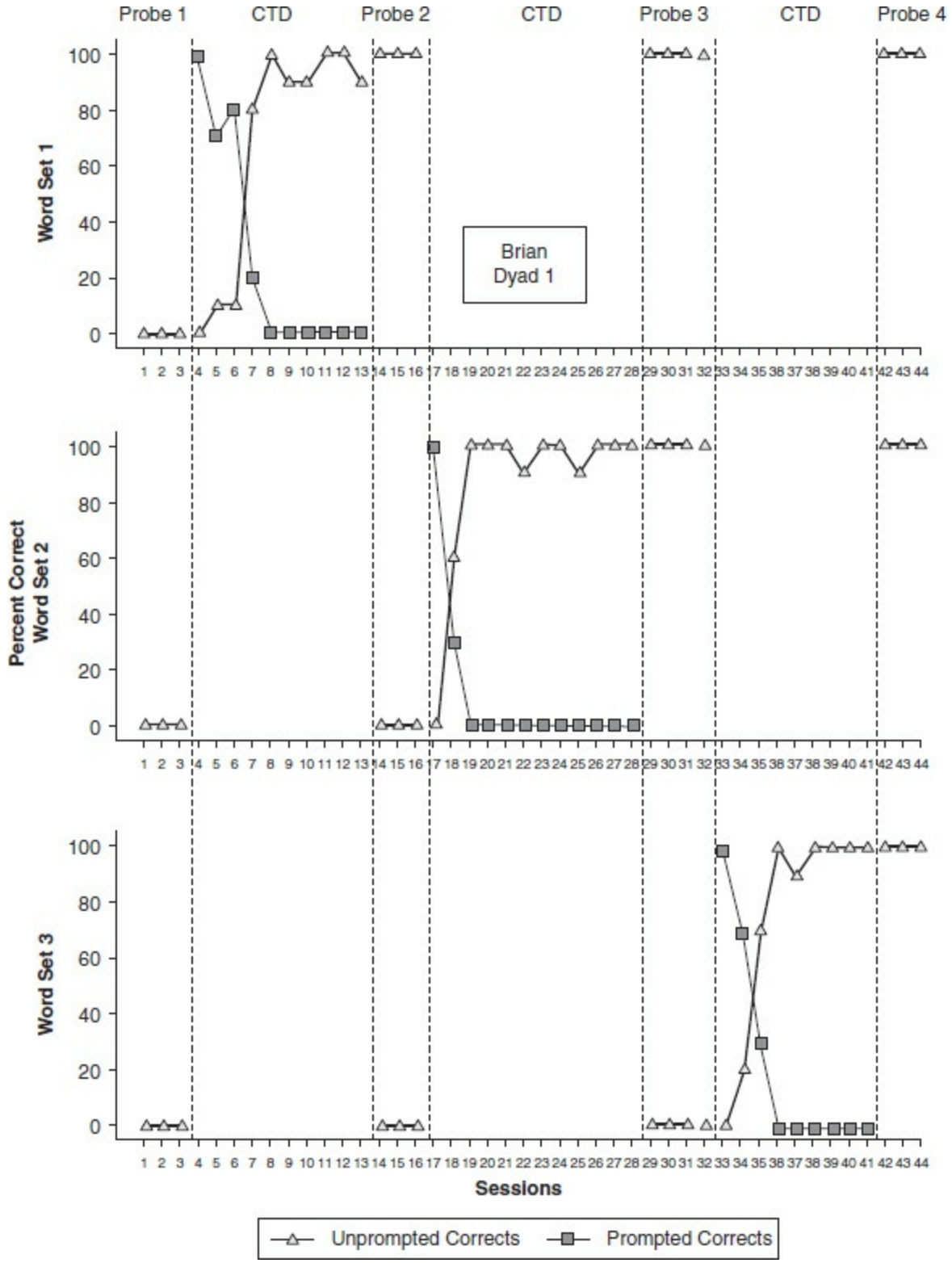


Figure 10.5 Multiple probe design (conditions variation) across behaviors for one participant.

Source: Ledford, J. R., Gast, D. L., Luscre, D., & Ayres, K. M. (2008). Observational and incidental learning by children with autism during small group instruction. *Journal of Autism and Developmental Disorders*, 38, 86–103.

Table 10.5 *Studies Using Multiple Baseline and Multiple Probe Designs Across Contexts*

Reference	Participants	Setting/Arrangement	Independent Variable	Dependent Variable
Huffman, R. W., Sainato, D. M., & Curiel, E. S. (2016). Correspondence training using social interests to increase compliance during transitions: An emerging technology. *Behavior Analysis in Practice, 9*, 25–33.	Number: 1 Sex: M Age: 6 Disability/ Diagnosis: Down syndrome	Setting: Inclusive preschool classroom Arrangement: Individual	Correspondence training	Percentage of intervals with on-task behavior
Darling, J. A., Otto, J. T., & Buckner, C. K. (2011). Reduction of rumination using a supplemental starch feeding procedure. *Behavioral Interventions, 26*, 204–213.	Number: 1 Sex: M Age: 27 Disability/ Diagnosis: Profound ID, BPD, PDD	Setting: Therapy room in residential facility Arrangement: Individual	Supplemental starch feeding procedure	Ruminations per minute
Dunlap, G., Ester, T., Langhans, S., & Fox, L (2006). Functional communication training with toddlers in home environments. *Journal of Early Intervention, 28*, 81–96.	Number: 2 Sex: F Age: 30, 33 months Disability/ Diagnosis: Language/ speech delays	Setting: Homes Arrangements: Individual, with mothers (also participants)	Functional communication training implemented by mothers	Percentage of intervals with challenging behavior
Hetzroni, O. E., & Tannous, J. (2004). Effects of a computer-based intervention program on the communicative functions of children with autism. *Journal of Autism and Developmental Disorders, 34*, 95–113.	Number: 5 Sex: 3 M, 2 F Age: 7–12 Disability/ Diagnosis: Autism	Setting: Self-contained classroom Arrangement: Individual	Interactive computer instruction program	Number of instances of echolalia, relevant and irrelevant speech, and communication initiations
Hughes, C., Golas, M., Cosgriff, J., Brigham, N., Edwards, C., & Cashen, K. (2011). Effects of a social skills intervention among high school students with intellectual disabilities and autism and their general education peers. *Research and Practice for Persons with Severe Disabilities, 36*, 46–61.	Number: 5 Sex: 3 M, 2 F Age: 16–21 Disability/ Diagnosis: ID (2), Autism, and ID (3)	Setting: Cafeteria, general education classrooms in high school Arrangement: Individual during training; Individual with peer during communication book use	Communication book use training for students with disabilities and general education peers	Percentage of intervals with (a) any social interaction, (b) participant initiation or peer response, and (c) peer initiation or participant response

Reference	Participants	Setting/Arrangement	Independent Variable	Dependent Variable
Barton, E. E., & Wolery, M. (2010). Training teachers to promote pretend play in young children with disabilities. *Exceptional Children, 77*, 85–106.	Number: 4 Sex: 2 M, 2 F Age: 30–50 months Disability/ Diagnosis: LD (1), DD (1), autism (2)	Setting: Classroom Arrangement: Individual in typical free play settings	System of least prompts	Number of pretense behaviors

Note: ID=intellectual disability, BPD=bipolar disorder, PDD=pervasive developmental disorder, LD=language delay, DD=developmental delay, M=male, F=female.

Procedural Guidelines

When using an MB or MP design across contexts, adhere to the following guidelines:

1. Identify a minimum of three similar yet functionally independent contexts for one individual.
2. Prior to the start of the study, identify a criterion for introduction of the intervention. For the initial introduction, stability in all tiers is an appropriate, conservative criterion. For remaining tiers, you may set a criterion level (e.g., 90% correct or better for 3 consecutive sessions) or a visual analysis criterion (e.g., following at clear change in level, with at least 3 consecutive sessions with no overlapping data with baseline; see Chapter 8).
3. Prior to the start of the study, identify the method by which you will assign contexts to tiers. It is generally reasonable to randomly assign contexts to tiers.
4. Determine a priori frequency of reliability and fidelity data collection (e.g., 33% of sessions), and conduct data collection for the duration of the study.
5. Concurrently collect baseline (probe) data for all contexts.
6. When data in all tiers are stable, intervene in the 1st context.

MB

7. When data in the 1st context have reached your criterion (usually behavior change indicated via visual analysis) and data in all contexts are stable, begin intervention in 2nd tier.
8. When data in the 2nd context have met your criteria (see above) and data in all tiers are stable, begin intervention in the 3rd context.
9. Repeat steps 7 and 8 for all remaining contexts.

MP (days)

7. When data in the 1st context approach your criterion, ensure that you have three consecutive days of data collected for the 2nd context.
8. When data are stable in all contexts, begin intervention in the 2nd context.
9. Repeat steps 7 and 8 for all remaining tiers.

MP (conditions)

7. When data in the 1st context have reached your criterion (usually an a priori mastery criterion), discontinue instruction and begin a second probe condition.
8. When data in all contexts are stable, begin intervention in the 2nd context.
9. Repeat steps 7 and 8 for all remaining tiers.

Internal Validity

MB and MP designs across contexts have adequate internal validity when all likely threats are controlled for, and experimental control is demonstrated when adequate internal validity is present and when behavior change occurs *when and only when* intervention is introduced to each targeted context, for at least three contexts with

concurrent start points. Threats to internal validity across variations are described in the beginning of this chapter. Two threats are noteworthy to mention in relation to the MB and MP designs across contexts. Specifically, infidelity threats and behavioral covariation may be particularly likely when these designs are used. Infidelity threats are likely when the same control agent implements the intervention *across contexts*. For example, a 1:1 aide might be tasked with implementing a token system intervention during math class, but asked to withhold the intervention during lunch time and independent work time. Once the paraprofessional has implemented the token system in one tier (particularly if there is a therapeutic behavior change), she might be likely to have low fidelity to baseline conditions (i.e., *not implementing the intervention* in the other two contexts). Of course, this can lead to the second threat: behavioral covariation. The child's behavior might change because the implementer delivers some parts of the intervention during baseline conditions. Even without infidelity, covariation may be likely. For example, decreasing a child's problem behavior and increasing engagement with the token board intervention might lead to improvements in peer relations in math class (the first tier). Consequently, peers may interact with the target participant differently in the other two contexts, or the child might generalize behaviors to those contexts. Thus, the MB and MP designs across contexts are highly susceptible to covariation; the uncertainty regarding generalization between contexts may be one reason why MB and MP designs across contexts are less widely used than the other two variations. One way to control for possible generalization across contexts is to embed an MP design across contexts with another design, such as an MP design across participants (Smith et al., 2016).

To reduce the likelihood of covariation, we recommend avoiding using contexts that are highly similar. The greater the similarity between contexts, the greater the likelihood that a participant's behavior will generalize. To avoid stimulus generalization, conduct an analysis of stimulus and response similarities across the three conditions by counting the number of shared stimulus characteristics and response variations. Knowledge of a participant's history under similar stimulus conditions also will be helpful in predicting unwanted stimulus generalization before the start of your study. For example, a student who has a reinforcement history for improving study habits (attention to task, answering questions, active participation in discussions) during reading lessons may exhibit generalized improvements during spelling and math lessons. Such generalization is more likely today than when Stokes and Baer (1977) wrote their seminal article on generalization programming because most practitioners understand the importance of using multiple exemplars during the teaching of new skills. In fact, fewer studies using MB and MP designs across contexts appear in the current applied research literature, which may be due to the increased use of general case programming procedures (Chadsey- Rusch, Drasgow, Reinoehl, Halle, & Klingenberg, 1993).

Selecting three contexts that are independent, yet similar, is at best an educated guess based on your familiarity with the generalization research literature, a participant's history with the behavior and target contexts, and the number of shared stimulus

characteristics across contexts. If there is a concern that the independent variable will have an inconsistent effect across tiers, a different research design should be used. Table 10.5 summarizes several studies in which an MB design across contexts was used to evaluate experimental control. Studies using MP designs across contexts are rare (e.g., about 1% of SCD graphs in a recent study; Ledford et al., 2017). Often, studies that include a different context also measure slightly different behaviors. Examples include an MP design used to teach email use across devices/platforms (i.e., different stimulus conditions that required somewhat different behaviors; Cihak, McMahon, Smith, Wright, & Gibbons, 2015) and an MB design across settings during which contextually-specific manual signs were taught (Miller, Collins, & Hemmeter, 2002).

Advantages

Both MB and MP designs across contexts permit an evaluation and demonstration of intra- participant direct replication. When multiple participants are included in a study, both intra-and inter-participant replications are possible. In addition, the MB and MP designs across contexts provide an experimental evaluation of interventions occurring in various contexts for the same participant, which is often required in educational and clinical settings.

Limitations

Limitations of the MB and MP designs across contexts include: (a) the challenge of identifying contexts that are functionally independent, for which there is little empirical guidance; (b) difficulty with measuring behavior across multiple contexts, which may introduce procedural complexity; (c) an increased likelihood of infidelity if a single implementer is used across contexts; and (d) the requirement to delay intervention in some contexts, which might be objectionable depending on target behaviors.

Conclusions

The MB or MP design across contexts is appropriate for evaluating the effectiveness of the same intervention across a variety of conditions, including settings, implementers, materials, instructional formats, and so on. However, you must proceed with caution because few guidelines exist for identifying conditions that are functionally independent, yet similar. This uncertainty is likely the reason that MP designs across contexts are rare. This design requires careful selection of contexts based on relevant generalization literature. If selected contexts are too similar, stimulus generalization is likely and experimental control is greatly weakened. Although experimental control is demonstrated for single participants when MB and MP designs across contexts are used, we recommend that multiple participants be recruited to improve external validity.

MB and MP Designs Across Participants

The most commonly used variation of MB and MP designs are MB and MP designs across participants. When these designs are used, the independent variable is sequentially introduced across several individuals who exhibit similar behaviors (or behavioral deficits) that occur under similar environmental conditions. The most conservative research approach is to identify individuals with similar learning histories who emit the same target behavior at similar frequencies under similar pre-intervention conditions. For example, if you were interested in assessing the effects of token reinforcement on reading rates across individual students, you may want to initially attempt to identify children of the same chronological age, with similar school backgrounds, who are currently reading at the same level in the same or a similar classroom. In subsequent investigations, and after a series of direct replications, you may choose to evaluate the generality of the intervention by identifying students who vary in one or more ways (e.g., chronological age, skill level) from students used in the initial study. In these replication attempts, the greater the differences across participants, the greater the generality of the findings. Initially, however, it is prudent to evaluate the effectiveness of your independent variable on a single target behavior emitted by participants with similar characteristics. After all, without demonstrating a functional relation, you can say little about the effectiveness (or lack thereof) of an intervention.

Applied Example 10–3: MP Design Across Contexts

Barton, E. E., & Wolery, M. (2010). Training teachers to promote pretend play in young children with disabilities. *Exceptional Children*, *77*, 85–106.

Barton and Wolery trained adult employees at an inclusive early childhood center to use the system of least prompts (SLP) and contingent imitation to improve pretend play behaviors of children with disabilities. Data were measured in the context of an MP design across contexts, with the contexts being different sets of toys. Participants were four 30–50 month old children with speech impairments, autism, or developmental delays. Implementers, who were trained to implement SLP by researchers, were paid full-time employees at an inclusive early childhood program with varying experience (3–24 years) and education (high school to graduate degrees). Toy sets included baby dolls and accessories (Toy Set 1), doll house and accessories (Toy Set 2), and kitchenware and accessories (Toy Set 3); all sets also included ambiguous items like blocks and sponges.

To assess effects of SLP on pretend play, authors used event recording to measure the number of four types of pretense behaviors (functional play with pretense, object substitution, imagining absent objects, and assigning absent attributes) emitted by the child during each 8-minute session. Reliability data were collected for at least 20% of sessions across participants, tiers, and conditions. Because event recording was used (rather than timed event recording), agreement was calculated using the gross agreement calculation (smaller count divided by larger count, multiplied by 100). Average agreement per condition per participant for unprompted pretense behaviors was 91–100%.

A secondary variable of interest was the procedural fidelity of the staff implementation of procedures during probe and instructional sessions. All staff implemented intervention with the initial set with a fidelity of 84–88% and all improved to 100% fidelity by the second or third toy set. These data, along with the implementation fidelity data for the training of adults, provides compelling evidence that the changes in adult behavior were due to training and that those changes were functionally related to increases in child pretend play behaviors.

For the initial probe condition, adults were instructed to play with children as they normally would. Then, they were provided with a 6-page manual and provided with a 45-minute training session before beginning instruction with the first toy set. After reaching criterion on each toy set, a probe condition was conducted and adults were provided training sessions for the use of SLP with the next toy set. During instructional sessions, adults imitated child behavior, praised pretend play behaviors, and used the system of least prompts to increase the number of pretend play behaviors. During the later probe conditions (e.g., all probe

conditions except the initial one), adults were instructed not to use SLP or reinforcement to encourage pretend play. In addition to instructional sessions, generalization sessions were conducted with each toy set; these sessions were conducted by a second adult who did *not* use the system of least prompts; these sessions were designed to assess the durability of responding in the absence of intervention use.

As shown in Figure 10.6, unprompted pretend play (depicted with filled triangles) was low during all pre-instruction probe conditions with the primary implementer, across all three toy sets. When instruction was initiated, there was an immediate but variable increase in unprompted correct responding across tiers. Generalization data were collected in the context of a multiple probe design, which is a considerable strength of the study since generalization data are more typically collected less frequently (e.g., pre- and post-intervention). Generalization data are less compelling in terms of functional relation conclusions, given the initially higher levels of play in probe conditions for two of three tiers and variability within probe and intervention conditions. Additional data are presented in tables (not included in this text) regarding the specific types of play behaviors across conditions as well as play diversity (e.g., number of unique play behaviors).

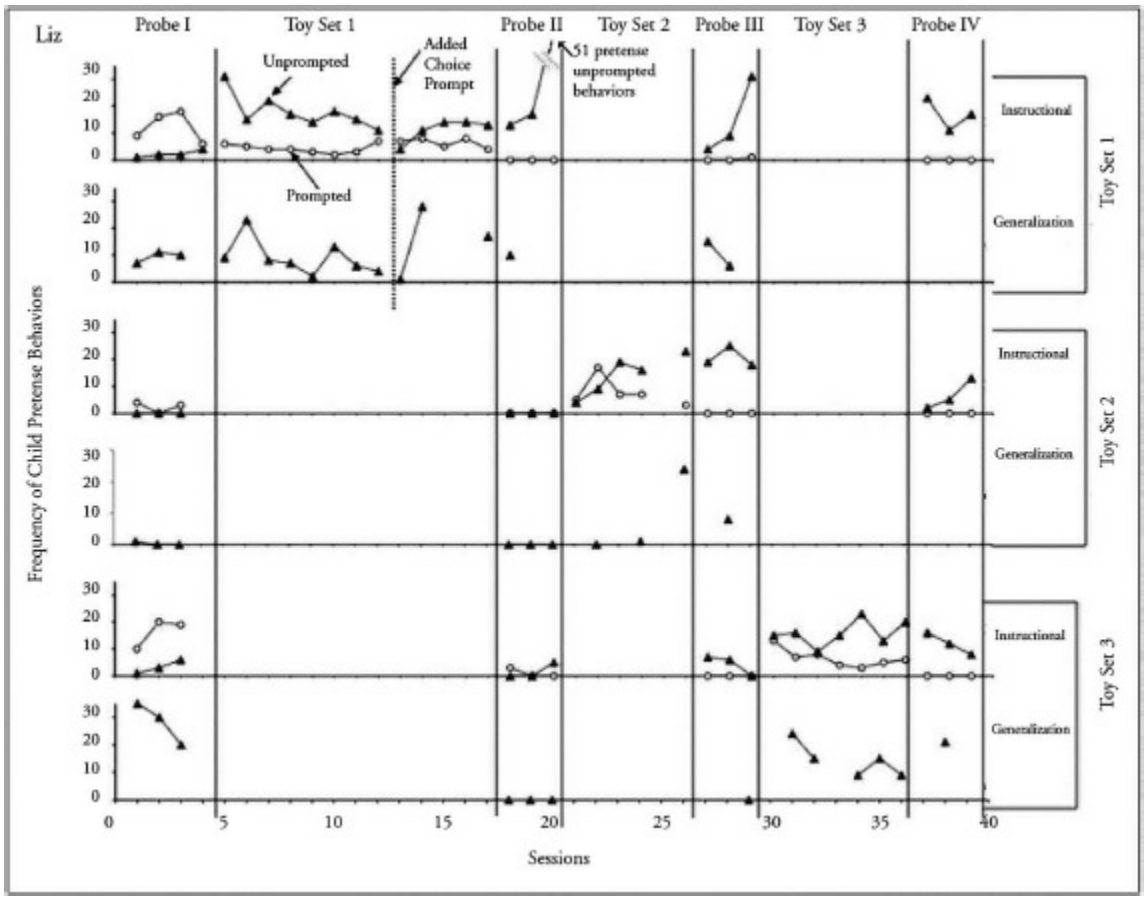


Figure 10.6 Multiple probe design (conditions variation) across contexts for one participant.

Source: Barton, E. E., & Wolery, M. (2010). Training teachers to promote pretend play in young children with disabilities. *Exceptional Children*, *77*, 85–106.

MB and MP designs across participants are well-suited for educational and clinical research when three or more individuals exhibit similar behavior excesses or deficits that require intervention. Assuming that behaviors emitted by prospective participants are not dangerous to themselves or others, it would be justifiable to introduce your intervention to one participant at a time before investing your time, and your participants' time, in an intervention that is not yet evidence-based. Identifying instructional programs and intervention strategies that are effective with several different individuals (or groups of individuals) extends the generality of findings to the extent that participants differ, which is a goal of educational and clinical research. However, although these designs are popular and practical, they are not as methodologically rigorous as MB or MP designs across behaviors or conditions. This is because there are no possibilities of intra-participant replication (unless you are using a combination design; see Chapter 12). Tables 10–16 and 10–17 summarize studies from the applied research literature that have used MB or MP designs across participants.

Table 10.6 *Studies Using Multiple Baseline Designs Across Participants*

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
Barton, E. E., Chen, C. I., Pribble, L., Pomes, M., & Kim, Y. A. (2013). Coaching preservice teachers to teach play skills to children with disabilities. *Teacher Education and Special Education, 36*, 330–349.	Number: 9 Sex: 3 M, 6 F Age: 23–27 Disability/ Diagnosis: None	Setting: Summer practicum placements Arrangement: Individual	Training and coaching for preservice teachers	Number of errors, number of correct prompts, percentage of intervals with contingent imitation
Cammilleri, A. P., Tiger, J. H., & Hanley, G. P. (2008). Developing stimulus control of young children's requests to teachers: Classwide applications of multiple schedules. *Journal of Applied Behavior Analysis, 41*, 299–303.	Number: 12 (Classroom A), 12 (Classroom B), and 10 (Classroom C) Sex: NR Age: 5–13 Disability/ Diagnosis: None	Setting: Classrooms in private elem. school Arrangement: Class-wide (large group)	Multiple schedule	Social approaches per minute
Harris, K. R., Friedlander, B. D., Saddler, B., Frizzelle, R., & Graham, S. (2005). Self-monitoring of attention versus self-monitoring of academic performance: Effects among students with ADHD in the general education classroom. *The Journal of Special Education, 39*, 145–156.	Number: 6 Sex: 4 M, 1 F Age: 3rd–5th grade Disability/ Diagnosis: ADHD	Setting: General education classroom Arrangement: Individual	Self-monitoring	Percent of intervals on task and number of correct responses (spelling)
Ingersoll, B., Lewis, E., & Kroman, E. (2007). Teaching the imitation and spontaneous use of descriptive gestures in young children with autism using a naturalistic behavioral intervention. *Journal of Autism & Developmental Disorders, 37*, 1446–1456.	Number: 5 Sex: M Age: 3–4 Disability/ Diagnosis: Autism	Setting: Clinic Arrangement: Individual	Contingent imitation, following child's lead, reinforcement	Percent of intervals during which participant imitated

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
Pisacreta, J., Tincani, M., Connell, J., & Axelrod, S. (2011). Increasing teacher use of a 1:1 praise-to-behavior correction ratio to decrease student disruption in general education classrooms. *Behavioral Interventions*, *26*, 243–260.	Number: 3 Sex: 2 M, 1 F Age: Not reported Disability/ Diagnosis: N/A	Setting: General education classrooms Arrangement: Individual/ class-wide	Modeling and performance feedback training	Ratio of praise to correction (teachers) and percentage of intervals with disruptive behavior (students)
Rakap, S. (2017). Impact of Coaching on Preservice Teachers' Use of Embedded Instruction in Inclusive Preschool Classrooms. *Journal of Teacher Education*, *68*, 125–139.	Number: 3 dyads Sex: F (T), M (C) Age: 53–60 months Disability/ Diagnosis: None (T), DD (C)	Setting: Inclusive preschool classrooms Arrangement: Typically occurring routines	Training and coaching for preservice teachers	Percentage of correct implementation (teachers), unprompted correct responding (children)

Note: ADHD=attention deficit hyperactivity disorder, M=male, F=female, NR=not reported, DD=developmental delay, T=teacher, C=children

Table 10.7 *Studies Using Multiple Probe Designs Across Participants*

Reference	Participants	Setting/Arrangement	Independent Variable	Dependent Variable
Godsey, J. R., Schuster, J. W., Lingo, A., Collins, B., & Kleinert, H. (2008). Peer-implemented time delay procedures on the acquisition of chained tasks by students with moderate and severe disabilities. *Education and Training in Developmental Disabilities*, *43*, 111–122.	Number: 4 Sex: M Age: 15–20 Disability/Diagnosis: Moderate ID	Setting: Kitchen/living area adjacent to self-contained classroom Arrangement: Small group	Peer-implemented time delay procedure	Percentage of correct steps completed independently on chained food preparation tasks
Hume, K., Plavnick, J., & Odom, S. L. (2012). Promoting task accuracy and independence in students with autism across educational setting through the use of individual work systems. *Journal of Autism and Developmental Disorders*, *42*, 2084–2099.	Number: 3 Sex: M Age: 7 Disability/Diagnosis: Autism	Setting: Self-contained classroom (intervention), general ed (generalization) Arrangement: Individual	Training on and introduction of independent work systems	Percentage of steps completed accurately

Reference	Participants	Setting/Arrangement	Independent Variable	Dependent Variable
Kelley, K. R., Bartholomew, A., & Test, D. W. (2013). Effects of the Self-Directed IEP delivered using computer-assisted instruction on student participation in educational planning meetings. *Remedial and Special Education, 34*, 67–77.	Number: 3 Sex: 1 M, 2 F Age: 15–20 Disability/Diagnosis: LD and ADHD (1), Mild ID (1), PDD (1)	Setting: Vacant classroom at private segregated school Arrangement: Individual (instruction), Small group (planning meetings)	Self-directed IEP training delivered using computer-assisted instruction (CAI) and role-playing	Student participation in educational planning meetings (number of points earned)
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.	Number: 4 Sex: 3 M, 1 F Age: 8–11 Disability/Diagnosis: Autism	Setting: Self-contained classroom	Video modeling	Percentage of steps for playing electronic gaming systems completed correctly
Taylor, P., Collins, B. C., Schuster, J. W., & Kleinert, H. (2002). Teaching laundry skills to high school students with disabilities: Generalization of targeted drills and nontargeted information. *Education and Training in Mental Retardation and Developmental Disabilities, 37*, 172–183.	Number: 4 Sex: M Age: 16–20 Disability/Diagnosis: Moderate ID	Setting: Self-contained classroom Arrangement: Individual	System of least prompts	Percent of steps completed independently (laundry)
Wright, T. S., & Wolery, M. (2014). Evaluating the effectiveness of roadside instruction in teaching youth with visual impairments street crossings. *The Journal of Special Education, 48*, 46–58.	Number: 4 Sex: 1 M, 3 F Age: 13–20 Disability/ Diagnosis: Vision impairments	Setting: Roadway intersections Arrangement: Individual	Verbal rehearsal and graduated guidance	Percentage of correct street-crossing behaviors

Note: ID=intellectual disability, LD=language delay, ADHD=attention deficit hyperactivity disorder, PDD=pervasive developmental disorder, M=male, F=female

Internal Validity

MB and MP designs across participants have adequate internal validity when all likely threats are controlled for. Experimental control is demonstrated when adequate internal validity is present and when behavior change occurs *when and only when* the intervention is introduced to each participant, for at least three participants with

concurrent start points. Threats to internal validity across variations are described in the beginning of this chapter. However, five threats are noteworthy specifically in relation to the MB and MP designs across participants variations. First, maturation may be most likely to occur in this variation of MB/MP designs than in any other SCD variation. This is due to the relatively long baseline conditions required for participants who are assigned to later tiers. In recently published studies, the average number of sessions in MB and MP designs is around 40—the final participant spends many of those sessions in the baseline condition (Ledford et al., 2017). Second, similar to MB and MP designs across behaviors, testing effects may be likely. This is true when the behavior of participants change due to the baseline condition procedures themselves. The testing threat is directly related to condition length, which is why this threat is likely for this design. Using the MP design across participants lowers the risk of testing threats, but decreases opportunities to assess for instability and covariation.

The two most concerning threats for MB and MP designs across participants are attrition bias and inconsistent effects. Attrition bias occurs when participant loss (attrition) has a high likelihood of impacting study results. This threat can be controlled for by *randomly assigning participants to tiers*. Historically, this assignment was more often based on data stability or researcher judgment; unfortunately, these procedures lead to the potential for bias. Thus, particularly for this design variation, we strongly recommend randomly assigning participants to tiers. Inconsistent effects may be more likely in this design because we have little information about what variables are related to response to intervention (cf. Eldevik et al., 2010). Thus, it may be likely that children who are similar on a number of variables will respond differently to a given intervention. For example, a review of social skills interventions for individuals with autism found relatively consistent success rates across design types, with the exception of MB designs across participants (Ledford, King, Harbin, & Zimmerman, 2016). When other designs are used (e.g., A-B-A-B, MB design across behaviors), inconsistent effects *for different participants* are informative, but do not necessarily influence experimental control *for other participants*. When experimental control is demonstrated for some participants and not others (e.g., for 2 of 3 participants in separate A-B-A-B designs), we can confidently say the intervention worked for some participants. When behavior change occurs for some participants in the context of an MB or MP across participants design, we cannot confidently attribute causality for any participants (i.e., behavior change for one or more participants might be related to history, maturation, etc.). Familiarity with dependent and independent variables and careful selection of participants with similar characteristics is critical to minimize likelihood of inconsistent effects.

Procedural Guidelines

When using an MB or MP design across participants, adhere to the following guidelines:

1. Identify a minimum of three participants who are functionally similar.
2. Prior to the start of the study, identify a criterion for introduction of the intervention. For the initial introduction, stability in all tiers (for all participants) is an appropriate, conservative criterion. For remaining tiers, you may set a criterion level (e.g., 90% correct or better for 3 consecutive sessions) or a visual analysis criterion (e.g., following at clear change in level, with at least 3 consecutive sessions with no overlapping data with baseline; see Chapter 8).
3. Prior to the start of the study, randomize participants to tiers.
4. Determine a priori frequency of reliability and fidelity data collection (e.g., 33% of sessions), and conduct data collection for the duration of the study.
5. Concurrently collect baseline (probe) data for all participants.
6. When data in all tiers are stable, intervene for the participant assigned to the first tier.

MB

7. When data for the 1st participant have reached your criterion (usually behavior change indicated via visual analysis) and data in all tiers are stable, begin intervention for the 2nd tier.
8. When data in the 2nd tier have met your criteria (see above) and data in all tiers are stable, begin intervention in the 3rd tier.
9. Repeat steps 7 and 8 for all remaining tiers.

MP (days)

7. When data for the 1st participant approach your criterion, ensure that you have three consecutive days of data collected for the participant assigned to the 2nd tier.
8. When data are stable in all tiers, begin intervention for the participant assigned to the 2nd tier.
9. Repeat steps 7 and 8 for all remaining tiers.

MP (conditions)

7. When data for the 1st participant have reached your criterion (usually an a priori mastery criterion), discontinue instruction and begin a second probe condition.
8. When data are stable for all participants, begin intervention for the participant assigned to the 2nd tier.
9. Repeat steps 7 and 8 for all remaining tiers.

Advantages

The primary advantage of MB and MP designs across participants is that they demonstrate some degree of external validity not shared by SCDs that include only a single participant. If consistent effects occur across participants, the researcher has demonstrated that intervention effects are not due to some idiosyncratic characteristic of a single participant. We do argue, however, that these inter-participant replications can occur with other designs by repeating designs including intra-participant replications

with multiple participants.

Limitations

Despite its widespread use, limitations of the MB and MP designs across participants are numerous. They include: (a) the need to identify and recruit three participants for whom the same intervention is likely to be effective for changing the same dependent variable; (b) an increased likelihood of inconsistent effects, leading to loss of experimental control; (c) complexity of simultaneously measuring behaviors for three participants; (d) ethical and experimental concerns regarding extended baseline conditions for participants assigned to later tiers; (e) lack of intra-participant replication; and (f) potentially high risk of testing and maturation threats due to prolonged baseline conditions for some participants.

Conclusions

In spite of considerable limitations, MB and MP designs across participants can be used to validate interventions across many types of individuals exhibiting a variety of behaviors in a variety of educational and clinical settings. As will be discussed in Chapter 12, MB and MP designs across participants can be superimposed over other MB or MP designs (behaviors or conditions) to enhance the internal and external validity of an intervention, thereby providing a powerful demonstration of experimental control.

Nonconcurrent (or Delayed) Multiple Baseline Designs

To reduce the length of baseline conditions or to increase flexibility to include new behaviors, conditions, or participants as they become available, some researchers have proposed the use of a "delayed multiple baseline design" (Watson & Workman, 1981), or **nonconcurrent** multiple baseline design (Harvey et al., 2004; Christ, 2007). This design is essentially a group of A-B designs with varying amounts of time spent in the "A" condition; beginning data collection in the first tier is *not* yoked with data collection in other tiers (Figure 10.8). As proposed in the literature (e.g., Carr, 2005; Harvey et al., 2004; Watson & Workman, 1981) the nonconcurrent or delayed multiple baseline design obviates the need for concurrent data collection across tiers.

Applied Example 10–4: MB Design Across Participants

Briere, D. E., Simonsen, B., Sugai, G., & Myers, D. (2015). Increasing new teachers' specific praise using a within-school consultation intervention. *Journal of Positive Behavior Interventions*, *17*, 50–60.

Briere, Simonsen, Sugai, and Myers (2015) studied the effects of a within-school consultation model on use of specific praise by newly certified teachers. Three teacher-mentor dyads participated in the study. The research question was: "What effects does a within-school consultative approach have on the new teachers' rates of specific praise statements during teacher-directed instruction?" (p. 51).

The dependent variable was the rate of new teachers' specific praise during 15-minute periods of directed instruction. Observers tallied the frequency of specific praise statements per minute, then calculated the rate by summing the total frequency of praise statements and dividing by the total number of minutes observed. Inter-observer agreement (IOA) data were collected for 41% of all sessions (range, 16%–50% of sessions within each condition for each teacher). Percentages of agreement were calculated by dividing the smaller frequency by the larger frequency and multiplying by 100. Mean IOA was 87% across all conditions and participants, but varied by participant and condition (e.g., M = 72% for baseline sessions with Holly; M = 99% for intervention sessions with Jill). Fidelity data were collected on three components of the consultation intervention: mentor-mentee training, consultation meetings, and teacher self-monitoring. As measured by checklists, fidelity to mentor-mentee training and consultation meeting procedures was 100% across teachers. As measured by whether each teacher used the hand-held counter during all, some, or none of the 15-minute observation periods, 88% of intervention sessions were scored as fully implemented, 7% were scored as partially implemented, and 5% were scored as not observed.

Three conditions were included in the study: baseline, intervention, and follow-up, all of which occurred during the same segment of teacher-delivered instruction in each teacher's classroom. During the baseline condition, teacher-mentor dyads met weekly but were not introduced to the within-school consultation intervention. Baseline data were collected concurrently on new teachers' specific praise during the selected 15-minute instructional period. One teacher was excluded from the study because she exceeded a pre-specified criterion of six or fewer praise statements in the 15-minute period; this left three teacher-mentor dyads (i.e., minimum number of tiers required to demonstrate a functional relation).

The order in which the consultation intervention was introduced was determined via randomly assigning dyads to tiers. During the intervention condition, a member of the research team trained the mentor using a scripted

protocol, and mentors used the same scripted protocol to train teachers on the self-monitoring procedures. Throughout intervention, new teachers self-monitored their use of specific praise statements during the selected instructional period using a hand-held counter, recorded total counts of specific praise in an Excel graphing template, and met with mentors on a weekly basis. During the follow-up condition, observers returned to each new teacher's classroom during the same instructional period and collected data on specific praise statements once a week for four weeks.

Figure 10.7 displays specific praise statements per minute during baseline, intervention, and follow-up conditions evaluated within the context of an MB design across participants (i.e., three teacher-mentor dyads). Prior to the intervention, each new teacher engaged in relatively low and stable rates of specific praise. Visual analysis shows immediate increases in rates of specific praise when the consultation intervention was introduced to each tier. Importantly, vertical analysis reveals no evidence of covariation across tiers—that is, introducing the consultation intervention to one dyad did not produce co-occurring behavior changes for other dyads still in baseline. Though rates of praise during intervention were somewhat variable, there was minimal overlap between baseline and intervention conditions. Finally, rates of specific praise in the follow-up condition were similar to rates observed during the intervention condition across tiers. Taken together, the implementation of the MB design across participants and visual analysis of data patterns within and across conditions provides evidence of a functional relation between the within-school consultation intervention and new teachers' rates of specific praise.

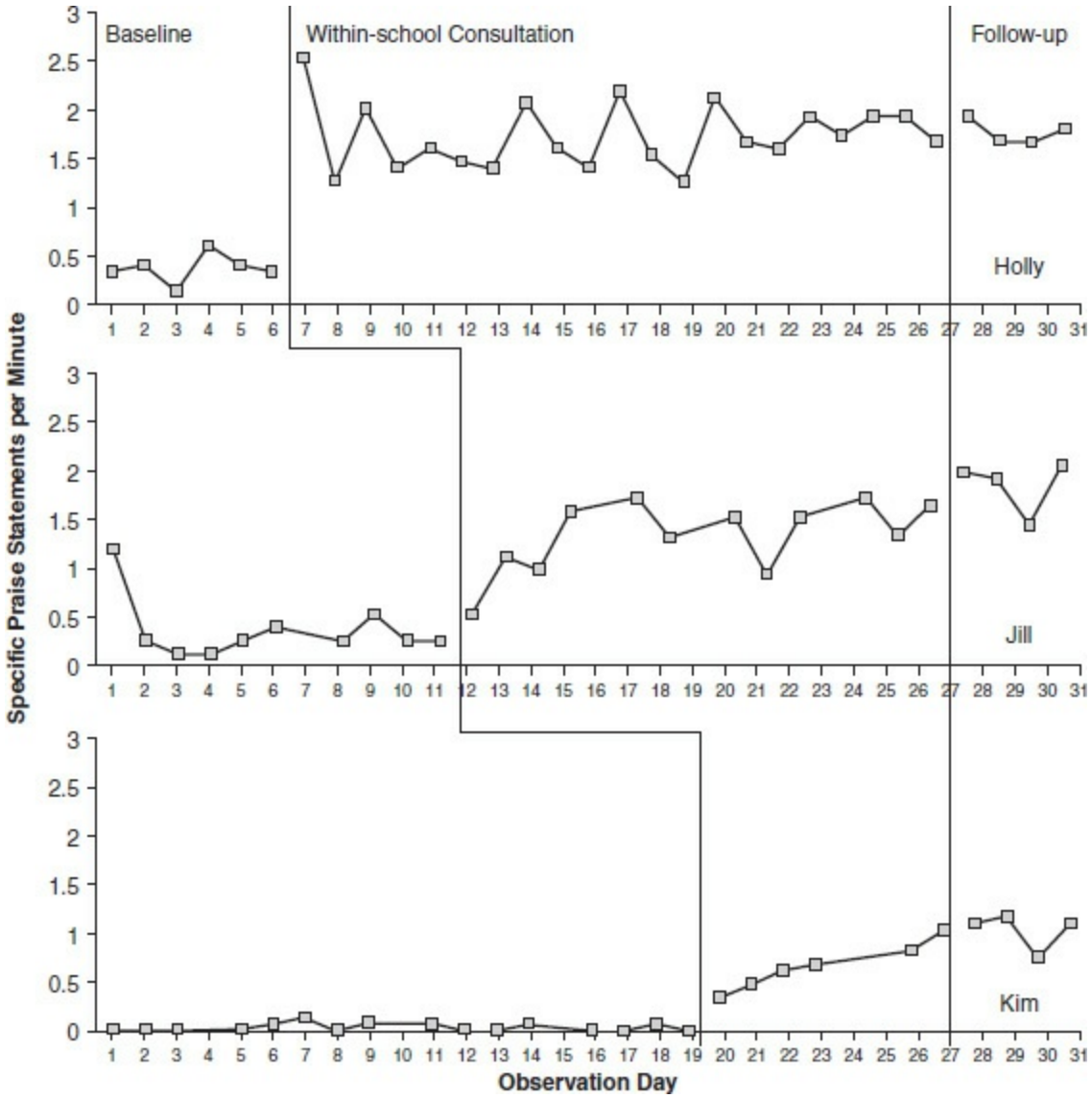


Figure 10.7 Multiple baseline across participants design.

Source: Briere, D. E., Simonsen, B., Sugai, G., & Myers, D. (2015). Increasing new teachers' specific praise using a within-school consultation intervention. *Journal of Positive Behavior Interventions, 17*, 50–60.

Assume you are interested in studying the effects of a school-wide positive behavior support strategy to reduce student office discipline referrals. You have identified three middle schools in three different school districts that have agreed to participate in your study. Consistent with nonconcurrent multiple baseline design guidelines, in year one you collect baseline data at one of the three schools; no baseline data are collected at the other two schools in the first year. Data are collected repeatedly (e.g., weekly) on the number of office discipline referrals, and when baseline data are stable the intervention is introduced school-wide. Data continue to be collected repeatedly across weeks until the end of the school year. In the second year of the research project, you collect baseline data at the second school for a longer period of time than at the first school (e.g., 6 weeks rather than 3 weeks), or until baseline data are stable, after which you implement the same intervention as you did at the first school, monitoring office referrals until the end

of the school year. Like in the first year of the project, no baseline data are collected at the third school, and you may or may not choose to collect maintenance data at the first school. This sequence of conditions is replicated at the third school in the third year with the important exception that the baseline condition must exceed the length of the baseline condition at the second school (e.g., 9 weeks rather than 6 weeks), or until data are stable. In other words, the nonconcurrent multiple baseline design requires that the same independent variable be implemented across tiers, that the same dependent variable be repeatedly measured, and that each subsequent tier's baseline condition be longer than preceding tiers. The assumption by those who advocate its use is that by requiring longer baseline conditions across tiers (e.g., 3 weeks, 6 weeks, 9 weeks) maturation threats to internal validity are adequately evaluated (i.e., the mere passage of time will not influence the dependent variable). This assumption is predicated on the position that organizations, clinics, or individuals do not change over time; or if they do change, they change in a predictable way. However, without baseline or probe data to substantiate these assumptions, it would be unwise to rule out maturation as a threat to internal validity.

Advocates of this design recognize that, "A primary limitation of the nonconcurrent multiple baseline design is the inability to identify history effects that may be coincidental with the application of a prescribed intervention, or occur at another time during the analysis." (Harvey et al., 2004, p. 274) History threats, as well as maturation threats, are major concerns to applied researchers who use multiple baseline and multiple probe designs. Confidence that maturation and history threats are under control is based on observing (a) an immediate change in the dependent variable upon introduction of the independent variable, and (b) baseline (or probe) condition levels remaining stable while other tiers are exposed to the intervention. Without the latter you cannot conclude, with confidence, that the intervention alone is responsible for observed behavior changes since baseline (or probe) data are not concurrently collected on all tiers from the start of the investigation. Only through repeated measurement across all tiers from the start of a study can you be confident that maturation and history threats are not influencing observed outcomes.

The advantage of the nonconcurrent multiple baseline design is strictly practical, not experimental. The design allows researchers to add participants who exhibit rare behaviors as they become available by implementing a series of A-B designs. In a similar vein, researchers interested in studying organization policies and their effect on behavior may not have the resources to collect frequent and repeated measures across three or more schools, hospitals, or clinics. In studying rare cases when only one case is referred every 6 months or once a year, you should consider all other SCD options before settling on an A-B design, regardless of your intent to lengthen the baseline condition over previous clients. In the case of studying organization policies, it is unclear why periodic probe data could not be collected intermittently, though infrequently, to dispel concerns regarding maturation and history threats to internal validity. Multiple probe designs are well suited for such research questions since pre-intervention data are collected

intermittently.

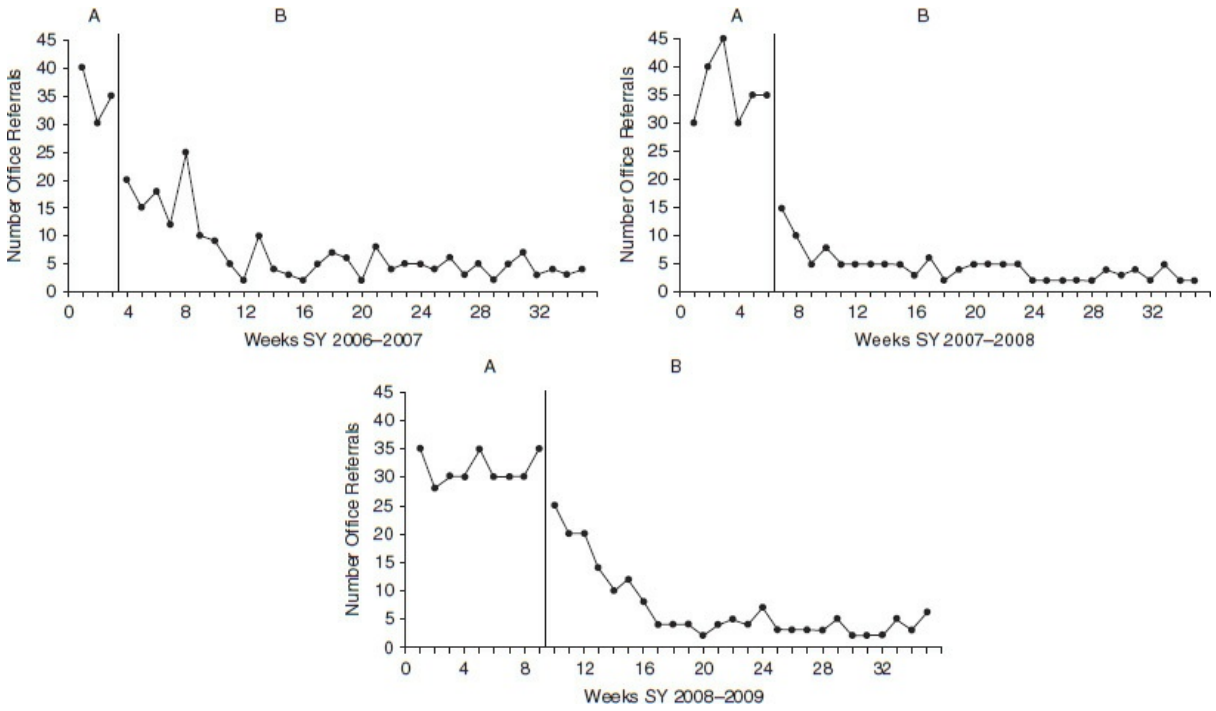


Figure 10.8 Prototype of a nonconcurrent multiple baseline design with appropriate visual display. Graphed in this way, readers are unlikely to assume concurrent measurement across tiers, analyzing data as three A-B designs (correct).

A word of caution is in order for consumers of research and those who might consider using nonconcurrent multiple baseline designs. Figure 10.8 shows the most appropriate and least misleading way to graphically present data generated in the context of a nonconcurrent multiple baseline design. Data on each tier show when baseline data collection was initiated over the course of three school years, clearly showing that data were not collected concurrently across tiers. Figure 10.9 shows an alternative graph format plotting the same data as shown in Figure 10.8; however, the first week of data collection on each tier is aligned with the first week noted along the abscissa giving the visual appearance that baseline data were collected concurrently across tiers. Though dates may be added along the abscissa of each tier, it is highly likely that readers will identify the design as a "true" multiple baseline design, rather than a nonconcurrent multiple baseline design, and incorrectly visually analyze findings without attention to possible maturation and history threats that were not evaluated. We believe, as do others (Carr, 2005) that a graphic representation of this type is deceptive and should be avoided; we recommend a graphic format similar to that in Figure 10.8.

Although the nonconcurrent multiple baseline design may have more flexibility than traditional multiple baseline and multiple probe designs, it does not, and cannot, provide as convincing a demonstration of experimental control because it fails to concurrently evaluate dependent variable levels in baseline conditions. The visual analysis of such data is limited to a simple A-B design, with all its shortcomings, followed by a series of

A-B replications across data series. In spite of its limitations, it continues to be used to evaluate clinical programs that address a wide range of behaviors of individuals, including bladder control (Duker, Averink, & Melein, 2001), sleep disturbance and sleepwalking (France & Hudson, 1990; Frank, Spirito, Stark, Owens-Stively, 1997), instruction compliance (Everett, Olmi, Edwards, & Tingstrom, 2005) and expressive communication (Hanser & Erickson, 2007; Tincani, Crozier, & Alazetta, 2006; Lancioni, O'Reilly, Oliva & Coppa, 2001). In reviewing these and other nonconcurrent multiple baseline design studies, it is important to attend to their compliance with the "baselines of different lengths" guideline and use of a graphic display format that accurately shows their attempts to evaluate threats to internal validity. Some researchers who have recognized the limitations of the nonconcurrent multiple baseline design have combined it with other SCDs, such as the A-B-A-B design (e.g., Freeman, 2006; Tiger, Hanley, & Hernandez, 2006) or multiple baseline design (Schindler & Horner, 2005), while others have developed variations of the design (e.g., longer baseline conditions on earlier tiers and systematically shorter baseline conditions on later tiers; Barry & Singer, 2001). In sum, the nonconcurrent multiple baseline design should be considered only as a last resort when more stringent SCDs cannot be used.

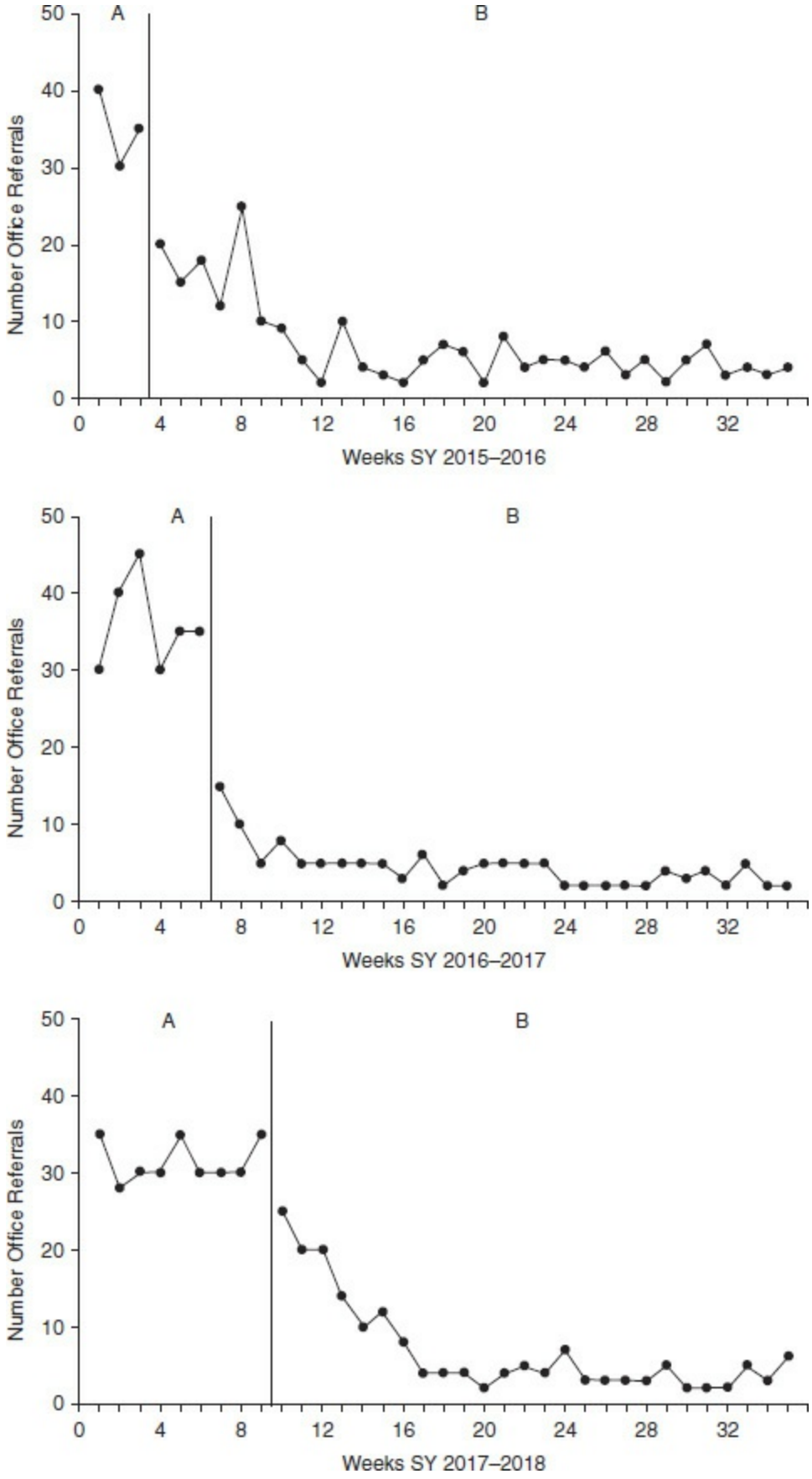


Figure 10.9 Prototype of a nonconcurrent multiple baseline design with inappropriate visual display. Graphed in this way, readers are likely to assume concurrent measurement across tiers (incorrect).

Summary

In this chapter, we provided an overview of the three types of MB and MP designs: across behaviors, across contexts, and across participants. Each of these designs has the distinct advantage of not having to return to pre-intervention conditions to evaluate experimental control. MP designs, in contrast to MB designs, are advantageous when continuous baseline measures are unnecessary, impractical, or reactive. To demonstrate experimental control with MB and MP designs, you will need to systematically introduce the independent variable to each tier in a time-lagged manner. If, upon introduction of the independent variable, there is (a) an immediate change in level or trend of the dependent variable, while there is (b) no change in level or trend in those data series not exposed to the independent variable, and (c) this effect is replicated across three or more tiers, experimental control has been demonstrated. Intra-participant direct replication increases confidence that the independent variable was responsible for observed changes, as demonstrated in MB and MP designs across behaviors or conditions. Intra-participant direct replication is not evaluated with MB and MP designs across participants. Inter-participant direct replication increases the generality of the findings if the effects of the independent variable are repeated across three or more participants in a study. Within a single investigation, the extent to which participants differ will determine the degree to which generality has been extended. As with all SCD studies, regardless of the particular experimental design, confidence in research findings increases with systematic replications—that is, when other researchers, in other settings, with different participants, studying similar or different behaviors, replicate the effects of the same independent variable.

Appendix 10.1

Visual Analysis for Multiple Baseline and Multiple Probe Designs

Adequate design	**Examples**: Three-tiered design with concurrent baseline measurement and three separate start points **Non-examples**: Two-tiered designs, three-tiered designs with two separate start points, non-concurrent designs
Visual analysis considerations specific to design	• **Concurrent baseline start points**. Before conducting visual analysis, you should ensure that start points for baseline conditions are concurrent (e.g., start at the same time). • **Sufficiently separate intervention start points**. For time-lagged implementation to be sufficient for controlling for threats to internal validity, intervention start points must occur *after* behavior change is demonstrated in the previous tier. • **Vertical analysis**. To determine whether behavior change occurs *when and only when* intervention is introduced, inspect each tier before and after intervention is introduced for all tiers. For example, if behavior change occurs in Tier 2 when intervention is introduced in Tier 1, experimental control is compromised. • **Sufficient pre-intervention data (MP)**. To adequately conduct vertical analysis, MP designs should include at least one data point per tier in *all tiers* (a) before intervention begins in any tier and (b) after intervention begins in each tier. In addition, each tier should include data collection immediately preceding intervention implementation.
Common and potentially problematic data patterns	• **Covariation across tiers**. Changes in untreated tiers concurrent with treatment initiation for others tiers indicate the tiers are not independent, and experimental control is compromised. • **Small magnitude changes**. Small magnitude changes are not problematic if data patterns are consistent for all tiers and if between-condition level change exceeds within-condition variability (e.g., no overlap is present across adjacent conditions within tiers). Small magnitude changes are potentially problematic if agreement data are discrepant (e.g., data from a second observer might suggest no change occurred; assessed via visual analysis of plotted data from both observers). • **Highly variable data in one or more condition**. Variable data are less problematic if changes in level are

	above and beyond variability (e.g., no overlap), or if changes in variability predictably change across conditions (e.g., high variability in baseline followed by low variability during intervention). Variability is problematic if there is a high percentage of overlapping data points across adjacent conditions within tiers or variability otherwise precludes making a decision regarding behavior change. • **Therapeutic trends in baseline conditions**. Therapeutic trends are not problematic if a large and abrupt change in level coincides with implementation of the intervention condition.
Convincing Functional Relation	• Consistent changes between A and B conditions for all tiers • Changes are abrupt and concurrent with condition changes • No change occurs in baseline conditions for untreated tiers concurrent with treatment initiation for others tiers • Overlap is minimal • Variability and trends in any condition do not preclude ability to identify between-condition changes.

References

Alberto, P. A., Fredrick, L., Hughes, M., McIntosh, L., Cihak, D. (2007). Components of visual literacy: Teaching logos. *Focus on Autism and Developmental Disabilities*, *22*, 234–243.

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.

Bandura, A. (1969). *Principles of behavior modification*. New York, NY: Holt, Rinehart, & Winston.

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

Barry, L. M., & Singer, G. H. S. (2001). A family in crisis: Replacing the aggressive behavior of a child with autism toward an infant sibling. *Journal of Positive Behavior Interventions, 3,* 28–38.

Barton, E. E., & Wolery, M. (2010). Training teachers to promote pretend play in young children with disabilities. *Exceptional Children*, *77*, 85–106.

Briere, D. E., Simonsen, B., Sugai, G., & Myers, D. (2015). Increasing new teachers' specific praise using a within- school consultation intervention. *Journal of Positive Behavior Interventions*, *17*, 50–60.

Cammilleri, A. P., Tiger, J. H., & Hanley, G. P. (2008). Developing stimulus control of young children's requests to teachers: Classwide applications of multiple schedules. *Journal of Applied Behavior Analysis*, *41*, 299–303.

Carr, J. E. (2005). Recommendations for reporting multiple-baseline designs across participants. *Behavioral Interventions, 20,* 219–224.

Chadsey-Rusch, J., Drasgow, E., Reinoehl, B., Halle, J., & Collett-Klingenberg, L. (1993). Using general case instruction to teach spontaneous and generalized requests for assistance to learners with severe disabilities. *Journal of the Association for Persons With Severe Handicaps*, *18*, 177–187.

Christ, T. J. (2007). Experimental control and threats to internal validity of concurrent and nonconcurrent multiple baseline designs. *Psychology in the Schools, 44,* 451–459.

Cihak, D. F., McMahon, D., Smith, C. C., Wright, R., & Gibbons, M. M. (2015). Teaching individuals with intellectual disability to email across multiple device platforms. *Research in Developmental Disabilities*, *36*, 645–656.

Cooper, J. O. (1981). *Measuring behavior* (2nd ed.). Columbus, OH: Charles E. Merrill.

Cronin, K. A., & Cuvo, A. J. (1979). Teaching mending skills to mentally retarded adolescents. *Journal of Applied Behavior Analysis*, *12*, 401–406.

Darling, J. A., Otto, J. T., & Buckner, C. K. (2011). Reduction of rumination using a supplemental starch feeding procedure. *Behavioral Interventions*, *26*, 204–213.

Doyle, P. M., Wolery, M., Ault, M. J., & Gast, D. L. (1989). Establishing conditional discriminations: Concurrent versus isolation-intermix instruction. *Research in*

Developmental Disabilities, *10*, 349–362.

Duker, P. C., Averink, M., & Melein, L. (2001). Response restriction as a method to establish diurnal bladder control. *American Journal on Mental Retardation, 106,* 209–215.

Dunlap, G., Ester, T., Langhans, S., & Fox, L. (2006). Functional communication training with toddlers in home environments. *Journal of Early Intervention*, *28*, 81–96.

Eldevik, S., Hastings, R. P., Hughes, J. C., Jahr, E., Eikeseth, S., & Cross, S. (2010). Using participant data to extend the evidence base for intensive behavioral intervention for children with autism. *American Journal of Intellectual and Developmental Disabilities*, *115*, 381–405.

Everett, G. E., Olmi, D. J., Edwards, R. P., & Tingstrom, D. H. (2005). The contributions of eye contact and contingent praise to effective instruction delivery in compliance training. *Education & Treatment of Children, 28,* 48–62.

Flores, M. M., & Ganz, J. B. (2007). Effectiveness of direct instruction for teaching statement inference, use of facts, and analogies to students with developmental disabilities and reading delays. *Focus on Autism and Other Developmental Disabilities*, *22*, 244–251.

France, K. G., & Hudson, S. M. (1990). Behavior management of infant sleep disturbance. *Journal of Applied Behavior Analysis, 23, 91*–98.

Frank, N. C., Spirito, A., Stark, L., & Owens-Stively, J. (1997). The use of scheduled awakenings to eliminate childhood sleepwalking. *Journal of Pediatric Psychology, 22,* 345–353.

Freeman, K. A. (2006). Treating bedtime resistance with the bedtime pass: A systematic replication and component analysis with 3-year-olds. *Journal of Applied Behavior Analysis, 39*, 423–428.

Ganz, J. B., Heath, A. K., Lund, E. M., Carmago, S. P. H., Rispoli, M. J., Boles, M., & Plaisance, L. (2012). Effects of peer-mediated implementation of visual scripts in middle school. *Behavior Modification*, *36*, 378–398.

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

Godsey, J. R., Schuster, J. W., Lingo, A. S., Collins, B. C., & Kleinert, H. L. (2008). Peer-implemented time delay procedures on the acquisition of chained tasks by students with moderate and severe disabilities. *Education and Training in Developmental Disabilities*, *43*, 111–122.

Grow, L. L., Carr, J. E., Kodak, T. M., Jostad, C. M., & Kisamore, A. N. (2011). A comparison of methods for teaching receptive labeling to children with autism spectrum disorders. *Journal of Applied Behavior Analysis*, *44*, 475–498.

Hanley, N. M., & Tiger, J. H. (2012). Teaching coin discrimination to children with visual impairments. *Journal of Applied Behavior Analysis*, *45*, 167–172.

Hanser, G. A., & Erickson, K. A. (2007). Integrated word identification and communication instruction for students with complex communication needs:

Preliminary results. *Focus on Autism and Other Developmental Disabilities, 22,* 268–278.

Harris, K. R., Friedlander, B. D., Saddler, B., Frizzelle, R., & Graham, S. (2005). Self-monitoring of attention versus self-monitoring of academic performance: Effects among students with ADHD in the general education classroom. *The Journal of Special Education*, *39*, 145–156.

Harvey, M. T., May, M. E., & Kennedy, C. H. (2004). Nonconcurrent multiple baseline designs and the evaluation of educational systems. *Journal of Behavioral Education, 13*, 267–276.

Hetzroni, O. E., & Tannous, J. (2004). Effects of a computer-based intervention program on the communicative functions of children with autism. *Journal of Autism and Developmental Disorders*, *34*, 95–113.

Horner, R. D., & Baer, D. M. (1978). Multiple probe technique: A variation of the multiple baseline design. *Journal of Applied Behavior Analysis*, *11*, 189–196.

Huffman, R. W., Sainato, D. M., & Curiel, E. S. (2016). Correspondence training using social interests to increase compliance during transitions: An emerging technology. *Behavior Analysis in Practice*, *9*, 25–33.

Hughes, C., Golas, M., Cosgriff, J., Brigham, N., Edwards, C., & Cashen, K. (2011). Effects of a social skills intervention among high school students with intellectual disabilities and autism and their general education peers. *Research and Practice for Persons With Severe Disabilities*, *36*, 46–61.

Hume, K., Plavnick, J., & Odom, S. L. (2012). Promoting task accuracy and independence in students with autism across educational setting through the use of individual work systems. *Journal of Autism and Developmental Disorders*, *42*, 2084–2099.

Ingersoll, B., Lewis, E., & Kroman, E. (2007). Teaching the imitation and spontaneous use of descriptive gestures in young children with autism using a naturalistic behavioral intervention. *Journal of Autism and Developmental Disorders, 37,* 1446–1456.

Jimenez, B. A., Browder, D. M., Spooner, F., & DiBiase, W. (2012). Inclusive inquiry science using peer-mediated embedded instruction for students with moderate intellectual disability. *Exceptional Children*, *78*, 301–317.

Johnson, J., McDonnell, J., Holwarth, V., & Hunter, K. (2004). The efficacy of embedded instruction for students with developmental disabilities enrolled in general education classes. *Journal of Positive Behavior Interventions*, *6*, 214–227.

Kazdin, A. E., & Kopel, S. A. (1975). On resolving ambiguities of the multiple-baseline design: Problems and recommendations. *Behavior Therapy*, *6*, 601–608.

Kelley, K. R., Bartholomew, A., & Test, D. W. (2013). Effects of the self-directed IEP delivered using computer- assisted instruction on student participation in educational planning meetings. *Remedial and Special Education*, *34*, 67–77.

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

Lancioni, G. E., O'Reilly, M. F., Oliva, D., & Coppa, M. M. (2001). A microswitch for vocalization response to foster environmental control in children with multiple

disabilities. *Journal of Intellectual Disability Research, 45*, 271–275.

Ledford, J. R., Gast, D. L., Luscre, D., & Ayres, K. M. (2008). Observational and incidental learning by children with autism during small group instruction. *Journal of Autism and Developmental Disorders*, *38*, 86–104.

Ledford, J. R., Severini, K. E., Zimmerman, K. N., & Barton, E. E. (2017). Data and graph characteristics in recent single case design research. *Manuscript in Preparation.*

Lieberman, R. G., Yoder, P. J., Reichow, B., & Wolery, M. (2010). Visual analysis of multiple baseline across participants graphs when change is delayed. *School Psychology Quarterly*, *25*, 28–44.

Marckel, J. M., Neef, N. A., & Ferreri, S. J. (2006). A preliminary analysis of teaching improvisation with the picture exchange communication system to children with autism. *Journal of Applied Behavior Analysis*, *39*, 109–115.

Mechling, L., Ayres, K. M., Purrazzella, K., & Purrazzella, K. (2012). Evaluation of the performance of fine and gross motor skills within multi-step tasks by adults with moderate intellectual disability when using video models. *Journal of Developmental and Physical Disabilities*, *24*, 469–486.

Miller, C., Collins, B. C., & Hemmeter, M. L. (2002). Using a naturalistic time delay procedure to teach nonverbal adolescents with moderate-to-severe mental disabilities to initiate manual signs. *Journal of Developmental and Physical Disabilities*, *14*, 247–261.

Murphey, R. J., & Bryan, A. J. (1980). Multiple-baseline and multiple-probe designs: Practical alternatives for special education assessment and evaluation. *The Journal of Special Education*, *4*, 325–335.

Pisacreta, J., Tincani, M., Connell, J. E., & Axelrod, S. (2011). Increasing teachers' use of a 1:1 praise-to-behavior correction ratio to decrease student disruption in general education classrooms. *Behavioral Interventions*, *26*, 243–260.

Rakap, S. (2017). Impact of Coaching on Preservice Teachers' Use of Embedded Instruction in Inclusive Preschool Classrooms. *Journal of Teacher Education*, *68*, 125–139.

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*, 1851–1869.

Schindler, H. R., & Horner, R. H. (2005). Generalized reduction of problem behavior of young children with autism: Building trans-situational interventions. *American Journal on Mental Retardation*, *110*, 36–47.

Smith, K. A., Ayres, K. A., Alexander, J., Ledford, J. R., Shepley, C., & Shepley, S. B. (2016). Initiation and generalization of self-instructional skills in adolescents with autism and intellectual disability. *Journal of Autism and Developmental Disorders*, *46*, 1196–1209.

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.

Stokes, T., & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis*, *10*, 349–367.

Tawney, J. W., & Gast, D. L. (1984). *Single subject research in special education.* New York, NY: Merrill.

Taylor, P., Collins, B. C., Schuster, J. W., & Kleinert, H. (2002). Teaching laundry skills to high school students with disabilities: Generalization of targeted drills and nontargeted information. *Education and Training in Mental Retardation and Developmental Disabilities*, *37*, 172–183.

Tiger, J. H., Hanley, G. P., & Hernandez, E. (2006). An evaluation of the value of choice with preschool children. *Journal of Applied Behavior Analysis*, *39*(1), 1–16.

Tincani, M., Crozier, S., & Alazetta, L. (2006). The Picture Exchange Communication System: Effects on manding and speech development for school-aged children with autism. *Education and Training in Developmental Disabilities*, 177–184.

Wall, M., & Gast, D. (1999). Acquisition of incidental information during instruction for a response chain skill. *Research in Developmental Disabilities*, *20*, 31–50.

Watson, P. J., & Workman, E. A. (1981). The non-concurrent multiple baseline across individuals design: An extension of the traditional multiple baseline design. *Journal of Behavior Therapy and Experimental Psychiatry, 12*, 257–259.

Werts, M. G., Caldwell, N. K., & Wolery, M. (2003). Instructive feedback: Effects of a presentation variable. *The Journal of Special Education*, *37*, 124–133.

Westerlund, D., Granucci, E. A., Gamache, P., Clark, H. B. (2006). Effects of peer mentors on work-related performance of adolescents with behavioral and/or learning disabilities. *Journal of Positive Behavior Interventions*, *8*, 244–251.

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 Education*, *22*, 14–25.

Wolery, M., Cybriwsky, C., Gast, D., & Boyle-Gast, K. (1991). Use of constant time delay and attentional responses with adolescents. *Exceptional Children*, *57*, 462–474.

Wolf, M. M., & Risley, T. R. (1971). Reinforcement: Applied research. In R. Glaser (Ed.), *The nature of reinforcement* (pp. 310–325). New York, NY: Academic Press.

Wright, T. S., & Wolery, M. (2014). Evaluating the effectiveness of roadside instruction in teaching youth with visual impairments street crossings. *The Journal of Special Education*, *48*, 46–58.

Yanardag, M., Akmanoglu, N., & Yilmaz, I. (2013). The effectiveness of video prompting on teaching aquatic play skills for children with autism. *Disability & Rehabilitation*, *35*, 47–56.

Youmans, G., Youmans, S. R., & Hancock, A. B. (2011). Script training treatment for adults with apraxia of speech. *American Journal of Speech-Language Pathology*, *20*, 23–37.

11
Comparative Designs

Mark Wolery, David L. Gast, and Jennifer R. Ledford

Important Terms

multitreatment interference, sequence effects, carryover effects, rapid alternation effects, nonreversibility of effects, separation of treatments; multitreatment design, alternating treatments design, multielement design, adapted alternating treatments design, parallel treatments design

The previous two chapters have focused primarily on demonstration designs; that is, designs that allow researchers to demonstrate that an intervention is effective for

changing a behavior of interest. In this chapter, we will focus on several types of comparison designs; that is, designs that allow researchers to compare two different interventions to determine which is more effective (or efficient) for changing a behavior of interest. First, we caution that researchers themselves decide on the use of the terms "intervention" and "baseline", and sometimes whether a condition should be considered "baseline" or "intervention" is ambiguous. For example, in one study (Chazin, Ledford, Barton, & Osborne, 2017), one condition included as a control condition was an antecedent teacher attention condition; authors included this comparison to ensure that teacher attention alone did not alter child engagement during a subsequent large group activity. However, in other contexts, this type of condition was the intervention of interest (McComas, Thompson, & Johnson, 2003). Luckily, the rationale of single case designs (SCDs) is similar regardless of whether one or more of the adjacent conditions being compared are therapeutic in nature.

When conducting comparative studies, you must select a design to answer the research question(s), but often more than one design could be used. Some of the issues to consider are: (a) whether behaviors being studied are reversible, (b) time available for conducting the study, (c) number of accessible participants, and (d) likely threats. In Table 11.1 the comparative designs are listed by whether they are relatively fast (require little time); by the type of behavior being studied, reversible or non-reversible; and by the method of condition ordering used. Similar to designs for demonstration studies (Horner et al., 2005), some comparative SCDs use independent variables (IVs) that are rapidly alternated (alternating treatments, adapted alternating treatments), others use IVs that are slowly alternated (e.g., over several sessions; multitreatment designs), and one uses rapidly alternating treatments with time-lagged introduction of IVs across sets of behaviors (parallel treatments design). Combination designs (e.g., alternating treatments plus A-B-A-B designs; Chapter 12) can also be used to make comparisons between IVs.

Table 11.1 *Comparative Designs Categorized by the Speed of Comparison, Type of Behavior, and Method of Condition Ordering*

Design	Speed of Comparison	Type of Behavior	Condition Ordering
Multitreatment	Slow	Reversible	Sequential Introduction and Withdrawal
ATD/M-ED	Fast	Reversible	Rapid Iterative Alternation
AATD	Fast	Non-reversible	Rapid Iterative Alternation
PTD	Slow	Non-reversible	Rapid Iterative Alternation + Time-Lagged Implementation

Note: This table is adapted from one developed by John W. Schuster of the University of

Kentucky.

Types of Comparative Studies

James Johnston, a noted behavioral scholar, said comparative studies are "the bane of the applied literature. They often lead to (1) inappropriate inferences, (2) with poor generality, (3) based on improper evidence, (4) gathered in the support of the wrong questions, thus (5) wasting the field's limited experimental resources" (1988, p. 2). He argued comparative studies were not done to understand principles of behavior in nature, but to see which intervention "wins." He criticized such research for having unfair and meaningless comparisons of procedures. His criticisms are more applicable to some studies than others, but you should heed them when doing comparative studies. Comparative studies often focus on one of the following endeavors.

Comparison of Competing Interventions

When faced with the same problems or issues, different investigators study different interventions. One may focus on intervention "B" and study it several times through systematic replications, and another may do the same with intervention "C." As a result, two or more effective interventions may be identified for improving similar behaviors for similar participants. The question is, "Which of those effective interventions will result in more efficient learning or more rapid deceleration of the challenging behavior?" The goal of such studies is to determine which intervention is superior and should be recommended to practitioners for use. As Johnston (1988) indicated, this is a direct attempt to see which intervention "wins" (cf. Addison et al., 2012).

In such cases, developers of each intervention ideally join together, plan, and conduct the comparison study. When they do not, the researcher must consider their perspective when planning the study to ensure a fair test of each intervention. The interventions must be used as the respective developers have described, should use dependent measures the developers find appropriate, and should involve participants and settings similar to those in the original research. If you consider these issues, then the developers are more likely to view the comparison as fair. You may need to contact the developers to get precise information about the procedures to ensure a meaningful comparison. Use of multiple dependent variables is recommended. For example, when comparing two instructional strategies, the efficiency measures may include the number of trials or sessions to criterion, number of minutes of instruction to criterion, number and percentage of errors to criterion, percentage correct during maintenance (follow-up) sessions, and degree of generalization. You also should have sensitive and appropriate measures of procedural fidelity to document the interventions were used as planned (Billingsley, White, & Munson, 1980; Fiske, 2008; Vollmer, Sloman, & Pipkin, 2008). One study does not settle which of two effective procedures is superior; multiple studies are

needed. If you design a comparison study in which you might be biased in favor of one of the interventions, use of blind coders considerably increases the value of the study (see Chapter 5). Also note that "superiority" is only established under conditions tested.

Comparison of an Innovation to an Established Intervention

Sometimes an intervention has a solid research foundation, is recommended widely, and used frequently. You may develop and study an innovation and want to compare it to the established intervention. The goal would be to determine whether the innovation or established intervention results in better outcomes. Two requirements are important in planning such studies. First, the established intervention must be used as its developers recommend and be applied to behaviors, participants, and contexts similar to those in the original research. Second, the innovation should be sufficiently well studied so an effective and refined form of it can be used. This is accomplished through a series of demonstration studies. You do not want to compare an innovation to an established practice prematurely, because an otherwise useful intervention may be discarded because it was not sufficiently refined before the comparison was attempted. Careful measurement of how the compared procedures are used should occur (Billingsley et al., 1980), and having multiple dependent measures is recommended.

Comparisons to Refine Interventions

Some comparative studies are not of two different interventions but evaluate variations of the same intervention to develop and refine it. The variations may include parametric questions, such as whether using more or less of a procedure results in differential behavior changes. For example, would using 5 or 10 trials per behavior per session result in differential rates of learning; or would having every-day versus every-other-day sessions result in more rapid learning. Other variations may focus on component analyses; for example, adding or deleting a given component (part, element) of an intervention package that may result in differential responding. Studies also can focus on procedural fidelity issues; specifically regarding whether using a procedure with high or low fidelity on some dimension results in different behavioral patterns (cf. Groskreutz, Groskreutz, & Higbee, 2011; Holcombe, Wolery, & Snyder, 1994). The goal of these studies is to identify the most powerful and efficient form of the intervention. When planning such studies, you should compare a form of the intervention that was effective in previous (demonstration) studies to a variation of that form. These studies help refine conclusions about when, under what situations, and for whom different variants of an intervention are recommended.

Comparisons to Understand Interactions

Sometimes research focuses on whether two or more interventions are more or less effective given a couple contextual variables. Contextual variables can be categorized on at least four dimensions: (a) physical space and materials, (b) social structure, (c) temporal structure, and (d) instructional characteristics. Variables of these dimensions are shown in Table 11.2. For example, you might compare:

- Individual and group contingencies on the frequency of students' comments when they are seated in rows versus groups at tables (structuring of physical space as a contextual variable)
- The use of preferred versus non-preferred materials on children's social interactions when one or three peers are present (social structure as a contextual variable)
- Interspersing easy and difficult tasks versus difficult tasks only on students' engagement during academic time when the previous activity was active or passive (temporal structure as a contextual variable)
- Self-monitoring with self-reinforcement versus teacher-reinforcement differentially influences the accuracy of tasks completed during seat work when children or teacher choose the order for doing assignments (instructional structure as a contextual variable)

The goal of such studies is to discern whether one intervention produces differential patterns of responding under varying conditions. Ideally, interventions will be effective across contextual variables (Hains & Baer, 1989), but knowing whether an intervention's effectiveness is influenced by contextual variables is important qualifying information that allows for more accurate recommendations about the situations under which interventions are likely to produce desirable effects and may hold implications about behavior-environment interactions. The requirements for such studies are (a) interventions should have solid research support; (b) some evidence, logic, or experience should suggest the contextual variable might influence performance; and (c) the interventions and contextual variables must be under the researcher's control. This last requirement is often difficult. Note that interactions between two variations (e.g., the effects of within-activity and across-activity choices during small group and individual sessions) requires a more complex comparison design than answering questions about choice and arrangement separately.

Table 11.2 *Examples of Dimensions of Contextual Variables*

Physical Dimensions—place, furnishings, and materials (inanimate entitie)s
• How the space is organized
• Size of space, furnishings, and materials
• Rules of access to space, furnishings, and materials
• Usual use of space, furnishings, and materials
• Regular and irregular variations in the space, furnishings, and materials
• Participants' learning history with similar or the specific space, furnishings, and

materials

Social Dimensions—others (adults, peers) in the study's context (animate entities)

- Demographic characteristics of those individuals
- Number of individuals in context by type
- The social organization in the context
- Usual patterns of interaction between individuals in the context
- Participants' history of interaction with the individuals in the context

Temporal Dimensions—the schedule and organization of events/activities in the context

- Order of events (i.e., activities, routines) in the day
- Predictability of the order of events
- Variations within and across events
- Novelty and familiarity of those events and their order
- Length of events
- Nature of expectations within events
- Practices related to transitions between events
- Participants' control of the order of events

Instructional Dimensions—Methods used to transmit knowledge in context

- Usual organization of instructional interactions
- Variation of the instructional interactions
- Social or instructional partner (participants, peers, adults)
- Practices used to ensure motivation
- Practices used to ensure attention to instructional stimuli
- Practices used to ensure social and deportment behavior in the context

Comparison of Popular and Research-Based Interventions

Despite current contexts in which practitioners are expected (and often required) to use evidence- based practices, many non-evidence based practices are widely used (Tostanoski, Lang, Raulston, Carnett, & Davis, 2014). Although we generally suggest comparing two *research-based* interventions (i.e., two interventions shown to be effective with demonstration designs), when a widely used intervention with little or no research support is suggested for use it is prudent to compare the effects of the untested interventions with an intervention supported by research. Several examples of studies designed to assess the effectiveness of two different interventions, one research-based and one widely used exist in the recent literature. These studies either use a sequential demonstration approach (e.g., demonstrate the widely used intervention is ineffective relative to baseline, then demonstrate the research-based intervention is effective relative to baseline; Cox, Gast, Luscre, & Ayres, 2009) or use a comparison design to compare each intervention to baseline simultaneously (Zimmerman, Ledford, & Severini, 2017). Given widely used interventions may drain private and public resources while potentially failing to improve outcomes, we consider comparing widely used or recommended practices and evidence-based interventions an appropriate use of SCD.

Internal Validity

Nothing about comparative studies make them immune from threats to internal validity faced by demonstration studies (e.g., history, maturation, instrumentation, lack of procedural integrity, attrition; Chapter 1). However, three issues deserve special note: multitreatment interference, non-reversibility of effects, and separation of treatments (Holcombe, Wolery, & Gast, 1994). Design-specific guidelines for visual analysis are available in Appendix 11.1.

Multitreatment Interference

Multitreatment interference is the influence one experimental condition has on performance under another experimental condition. Note that these effects can occur in demonstration designs; they are just more likely when the experimental conditions are both therapeutic in nature. Historically, two types were recognized: carryover effects, and sequence effects (sequential confounding; Barlow & Hayes, 1979). **Carryover effects** are the influence of one experimental condition on performance under another condition due to the nature (characteristics) of the initial condition. **Sequence effects** are the influence of one condition on another due to the ordering of experimental conditions. A third type of multitreatment interference is **rapid alternation effects**—the effects on performance due to rapidly changing (alternating) conditions (Hains & Baer, 1989). Carryover effects occur in the context of the sequence of experimental conditions. Specifically, a condition (e.g., B) can only influence subsequent performance in Condition C if the participant experienced Condition B first. Participants must experience a sequence of at least two experimental conditions before multitreatment interference (carryover, sequence, or alternation effects) is possible; logically, a condition (e.g., Condition B) cannot influence performance under another condition (e.g., Condition C) unless the participants first experienced the original condition (Condition B). Thus, Hains and Baer (1989) asserted, "there is little reason to maintain a distinction in terminology between sequence, carry-over, and alternation effects. All that is of issue are sequence effects, sometimes in faster paced sequences, sometimes is slower paced sequences" (p. 60). This assertion is valid, because differences between sequence and carryover effects are subtle and in many cases impossible to disentangle. Figure 11.1 shows hypothetical *sequence* effects (e.g., the behavior of each intervention depends on the *sequence* in which they are introduced. Figure 11.2 shows hypothetical rapid alternation effects; that intervention (B) results in behavior change only when alternated with conditions (A) and (C). Although we are not aware of any published studies which have detected these effects, some research supports theoretically that alternating conditions may result in differential behavior change (Dunlap, 1984; Milo, Mace, &

Nevin, 2010). Figure 11.3 shows potential carryover effects.

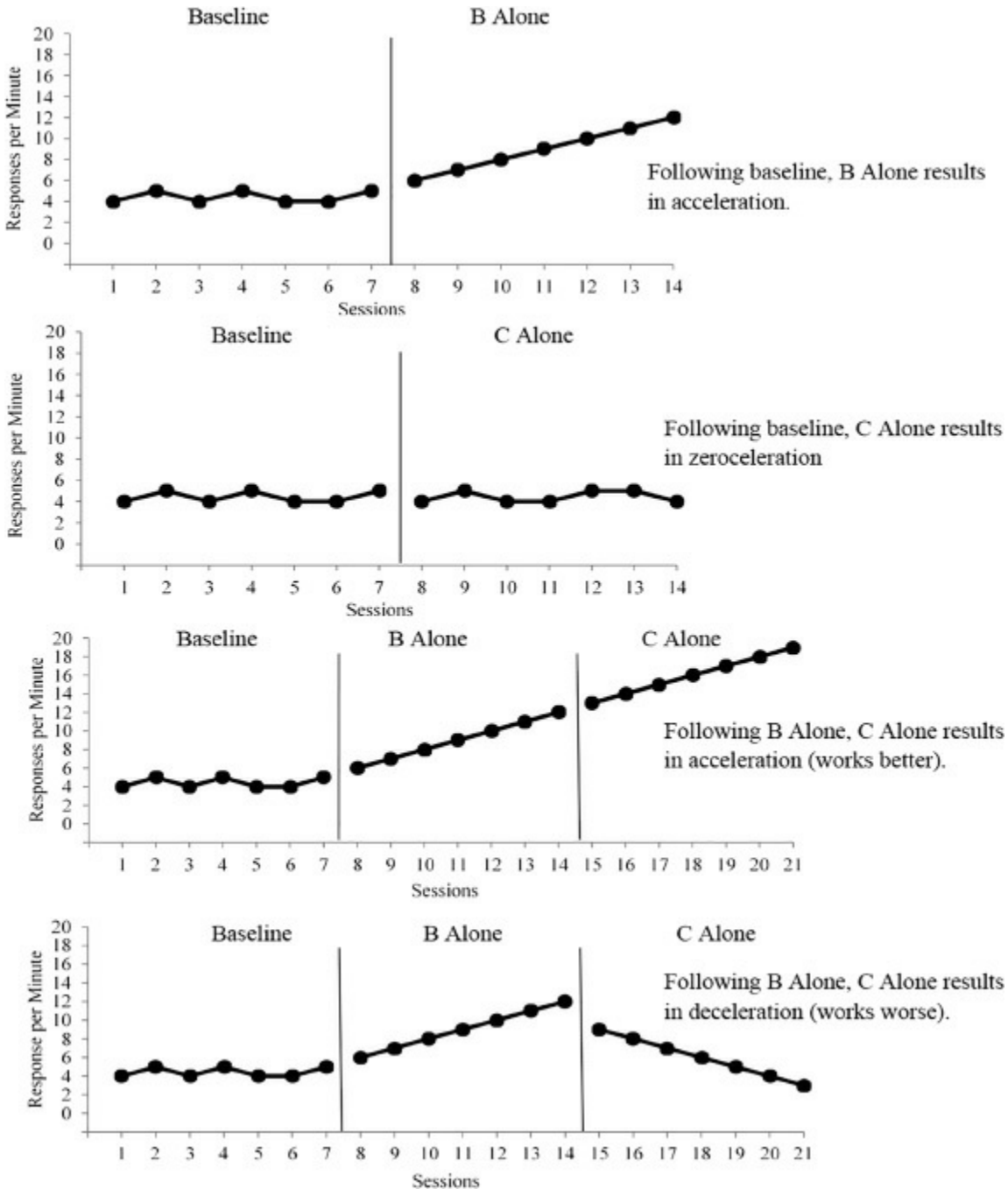


Figure 11.1 Graphs depicting potential sequence effects.

Non-reversibility of Effects

When comparing two interventions, one potential threat to internal validity is **non-reversibility of effects**; this refers to the likelihood that once behavior change occurs, it will maintain even when the condition resulting in the behavior change is removed. It is problematic when two interventions are applied to the same, non-reversible behavior (Holcombe et al., 1994). The treatment outcome for the participant is desirable, but you

have no opportunity to test the relative merits of the interventions being compared. These effects are most likely with trial-based behaviors such as academic behaviors, when generalization between conditions is likely (e.g., learning to read a word in one condition will likely result in correct performance across conditions). For this reason, special comparison designs are needed to compare the effects of two interventions on non-reversible behaviors.

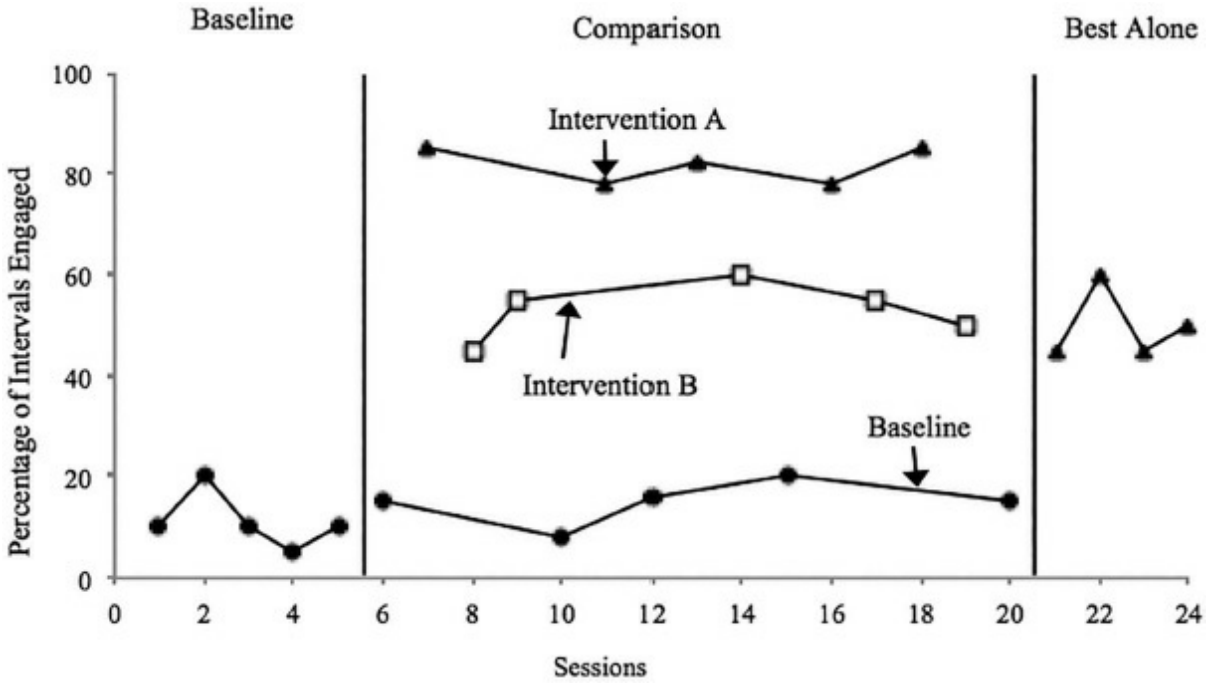


Figure 11.2 Graph depicting rapid alternation effects (i.e., Intervention A resulted in improved outcomes when it was alternated with other conditions, and less optimal outcomes when it was implemented alone).

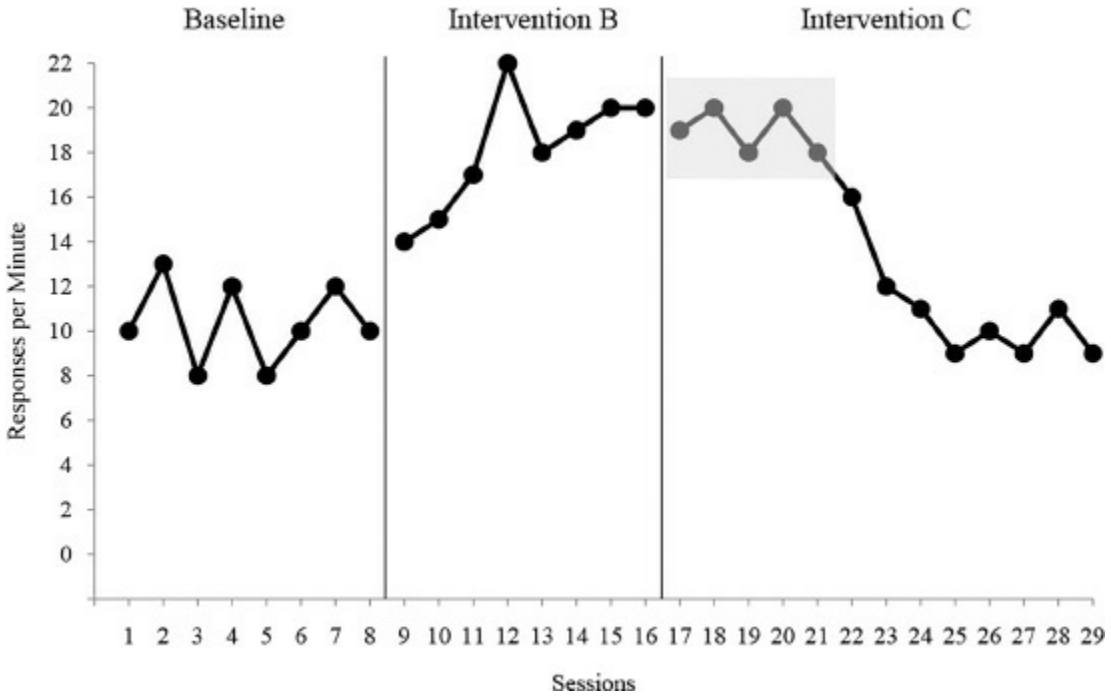


Figure 11.3 Graph depicting carryover effects.

Separation of Treatments Issue

Most comparative studies are conducted to evaluate the superiority of one intervention over other(s). You want to attribute the ultimate results to one and only one intervention. However, there is a **separation of treatments issue** when using some comparison designs: *when two or more interventions are applied to the same behavior, the ultimate levels of the behavior cannot be attributed to only one intervention.* This situation is illustrated in Figure 11.4. In the top graph, the behavior increased to 100% correct (sessions 12 and 13), but this level of responding cannot be attributed to either intervention alone. Perhaps each intervention individually would establish that level of responding, but that conclusion goes beyond the data. In the bottom graph, the behavior decelerated quickly, and Intervention B appears to be superior to Intervention C, but Intervention B may not have had the same effect if it were not alternated with Intervention C. It may have resulted in more rapid deceleration, or perhaps no deceleration at all. This inability to attribute the ultimate behavior change to one and only one intervention is known as the separation of treatments issue (Holcombe et al., 1994). This is typically not a critical threat to internal validity because researchers generally compare two interventions that have been shown to be effective in isolation but should generally be described as a limitation in a Discussion section of a research report (see Chapter 3).

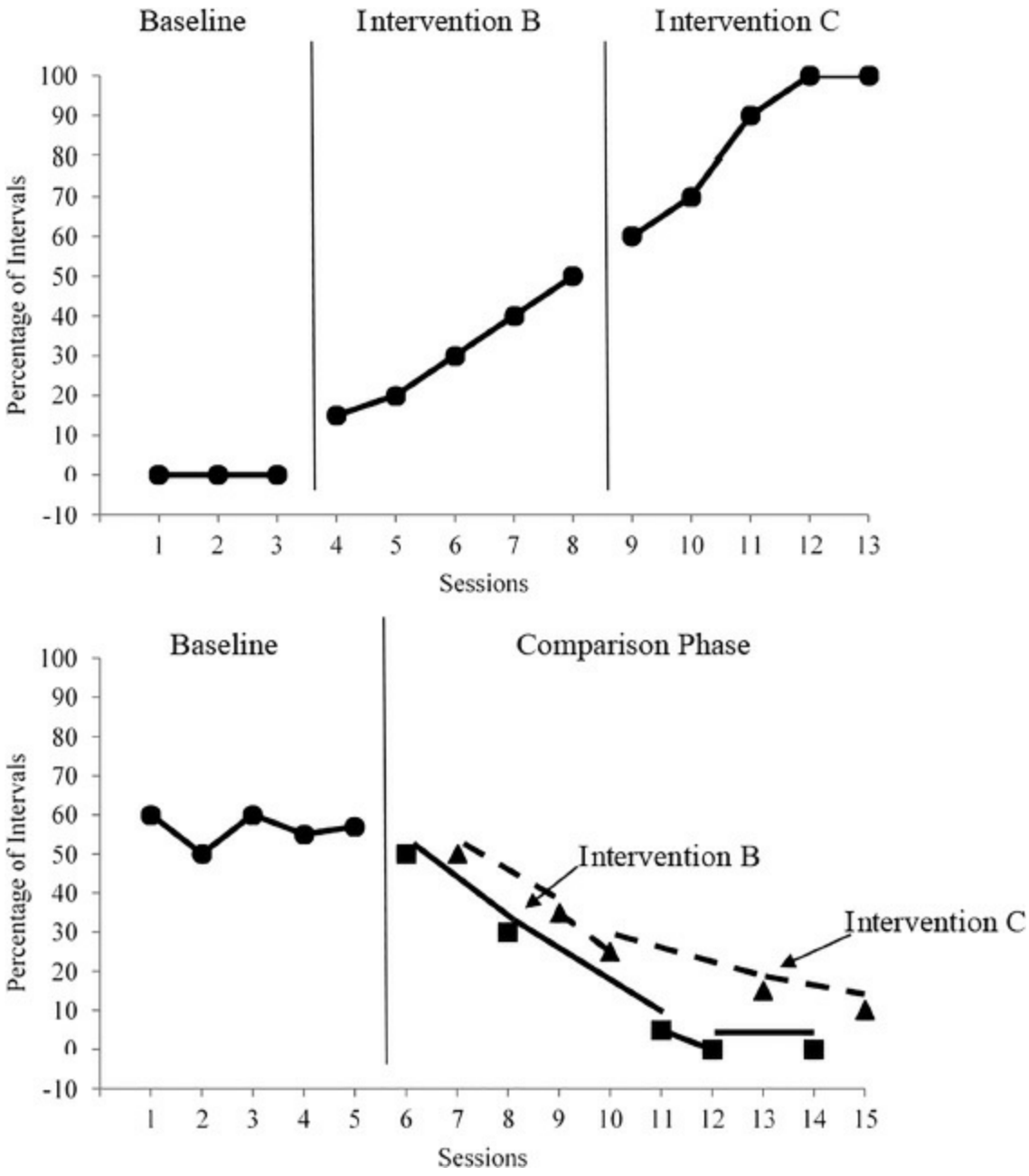


Figure 11.4 Graphs depicting separation of treatments problem.

Multitreatment Designs

Sequential introduction and withdrawal designs are flexible designs that allow for comparisons *between two treatments* in addition to comparisons between baseline and one treatment. Variations of the A-B-A-B design developed to *compare* treatments are called **multitreatment designs**. The simplest multitreatment design includes a B-C-B-C sequence. These designs are perhaps the oldest comparative SCDs (Birnbrauer, Peterson, & Solnick, 1974). Often, multitreatment designs are used when an a priori comparison is planned (e.g., your research question is "Does Intervention B produce better outcomes than Intervention C?"); however, these designs can also be extensions of A-B-A-B designs; either because the B condition did not initially show a sufficient effect (in which case, you can add a "C" condition and conduct an A-B-C-B-C design) or following the initial comparison, you would like to answer a different question (e.g., A-B-A-B-C-B-C design). In the first example above, the A-B-C-B-C design, you can demonstrate experimental control for the B C comparison. However, you *cannot* experimentally evaluate the effects of B or C in relation to the initial baseline condition. However, in the second design (A-B-A-B-C-B-C), you can answer both a demonstration question ("Does intervention B produce changes in behavior in comparison to baseline?") and a comparison questions ("Does intervention C produce superior behavior change compared to intervention B?"). The downside of this type of complex design is the long duration. With three to five data points per condition, this study would take at least 21–35 sessions to complete (and that is barring any participant absences and other practical constraints). A second way to modify an A-B-A-B design in the case of unacceptable effects for the B condition is to add a component to the intervention, and evaluate the intervention with and without that component (A-B-BC-B-BC design).

Multitreatment designs, as with withdrawal designs, require only one dependent variable, but more are recommended. When multiple behaviors are measured, you should designate a priori one behavior to use for making experimental decisions. Additional behaviors can be viewed as secondary or corollary measures. The multitreatment design can be used to study both acceleration and deceleration behaviors and should be used only with *reversible behaviors*. As with A-B-A-B designs, replication with multiple participants is recommended.

Multitreatment designs can be used with or without a baseline condition. When possible and logically reasonable, baseline conditions should be used. Having a baseline condition allows you to establish the need for the intervention. This information is needed to make generalizations about the effects of interventions to other participants (Birnbrauer, 1981). Some comparisons, however, do not have a logical baseline condition. For example, a teacher may want to compare two ways of arranging the physical space of the classroom; may want to compare child- or teacher-choice of some instructional variable (e.g., order of completing assignments); or may want to compare individual and

group contingencies. In such studies, a baseline condition may not be relevant.

The effects of SCDs are evaluated via replication (Edgar & Billingsley, 1974), and multitreatment designs are no exception. An A-B-C multitreatment design does not have sufficient replications to control for maturation and history and precludes conclusions that a functional relation exists. Similarly, despite having four conditions, an A-B-A-C design does not include three demonstrations of effect between any two conditions. At minimum, three potential demonstrations between two adjacent conditions are needed (i.e., A-B-C-B-C). This applies to each experimental manipulation in the study. For example, an A-B-C-B-C-D design has three potential demonstrations of effect between B and C (B to C, C to B, and B to C) but only one for A to B and C to D. Sufficient replications exist to draw functional conclusions about the effectiveness of B compared to C, but causal conclusions cannot be drawn regarding the effectiveness of A or D in relation to any other condition.

When comparing two or more interventions in a multitreatment design the order of conditions should be counterbalanced across participants to control for sequence effects. In terms of sequence effects, if all participants had the same order of experimental conditions (e.g., A-B-C-B-C) you could not claim intervention C would be effective without following intervention B. A stronger arrangement is to have half of the participants follow the A-B-C-B-C sequence and half of the participants follow the A-C-B-C-B sequence. If this is done, and C is clearly superior to B across all intra- and inter-participant replications, then you could conclude C is superior regardless of its sequence with B. Note that in both cases, the functional relation is demonstrated for B and C in relation to the other; not for either in relation to baseline, though a description of baseline levels is helpful in describing the extent to which behavior of all participants was similar prior to intervention implementation (Birnbrauer, 1981). Table 11.3 summarizes studies using a multitreatment design to evaluate behavior change.

Table 11.3 *Studies Using Multitreatment Designs*

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
Addison, L. R., Piazza, C. C., Patel, M. R., Bachmeyer, M. H., Rivas, K. M., Milnes, S. M., & Oddo, J. (2012). A comparison of sensory integrative and behavioral therapies as treatment for pediatric feeding disorders. *Journal of Applied Behavior Analysis*, *45*, 455–471.	Number: 2 Sex: 1 M, 1 F Age: 1–3 Disability/diagnosis: GERD, asthma, DD, dysphagia, poor oral intake	Setting: Therapy rooms Arrangement: Individual	Sensory integration therapy and escape (B) versus escape extinction and non-contingent reinforcement (C) (A-B-C-B-C)	Percentage of bites accepted and rate of inappropriate behavior per minute
Barton, E. E., Stiff, L., & Ledford, J. R. (2017). The effects of contingent reinforcement on peer imitation in a small group play context. *Journal of Early Intervention*.	Number: 10 Sex: M (3), F (7) Age: 4 Disability/diagnoses: None (7), seizure disorder (1), DD (1), Prader-Willi (1)	Setting: Inclusive early childhood center Arrangement: Small group	System of least prompts with noncontingent reinforcement versus systems of least prompts with contingent reinforcement for imitation (A-B-C-B-C-C')	Number of unprompted imitation, pretend play, and social communication behaviors

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
Freeman, K. A., & Dexter-Mazza, E. T. (2004). Using self-monitoring with an adolescent with disruptive classroom behavior. *Behavior Modification, 28*, 402–419.	Number: 1 Sex: M Age: 13 Disability/diagnosis: ADHD, conduct disorder, adjustment disorder, mathematics disorder.	Setting: Residential facility for youth with conduct problems Arrangement: Individual	Self-monitoring (B) and matching (C) (A-B-BC-B-BC)	Percent combined off-task and disruptive behavior
Hanley, G. P., Piazza, C. C., Fisher, W. W., & Maglieri, K. A. (2005). On the effectiveness of and preference for punishment and extinction components of function-based interventions. *Journal of Applied Behavior Analysis, 38*, 51–65.	Number: 2 Sex: 1 M; 1 F Age: 5–8 Disability/diagnosis: Moderate ID, ADD, ODD	Setting: Treatment rooms Arrangement: Individual	Functional communication training (B) and punishment (C) (A-B-BC-B-BC)	Aggressive responses per minute
Sanetti, L. M. H., Luiselli, J. K., & Handler, M. W. (2007). Effects of verbal and graphic performance feedback on behavior support plan implementation in a public elementary school. *Behavior Modification, 31*, 454–465.	Number: 1 group of 4 teachers (data graphed by group) Sex: F Age: Adult Disability/diagnosis: None	Setting: Public school Arrangement: Typical classroom instruction (multiple teachers and students)	Visual feedback (B) versus verbal feedback (C) for behavior support plan implementation for one student (A-B-C-B-C)	Percentage of behavior support plan components implemented as written
Torelli, J. N., Lloyd, B. P., Diekman, C. A., & Wehby, J. H. (2017). Teaching stimulus control via class-wide multiple schedules of reinforcement in public elementary school classrooms. *Journal of Positive Behavior Interventions, 19*, 14–25.	Number: 2 classrooms Sex: NR Age: 1st & 2nd grade Disability/diagnosis: 3 children in each class had undisclosed disability	Setting: Public school Arrangement: Whole-class	B and C conditions were two different multiple schedule conditions that corresponded with teacher attention (C1: A-B-C-B-C-D-FU) (C2: A-C-B-C-B-D-FU)	Rate or recruitment of teacher attention with two stimuli (light off and light on)

Note: DD=developmental delay, ADHD=attention deficit hyperactivity disorder, ADD=attention deficit disorder, ODD=oppositional defiant disorder, C1=Classroom 1, C2=Classroom 2, FU=follow up

Procedural Guidelines

Multitreatment designs are useful for answering many research questions, including comparisons of two different interventions, analyses of components of treatment packages, and parametric analyses. When using an A-B-C-B-C design, adhere to the following guidelines:

1. Identify and define a reversible target behavior.
2. Select a sensitive, reliable, valid, and feasible data collection system and pilot the system and your behavior definitions.
3. Determine a priori frequency of reliability and fidelity data collection (e.g., 33% of sessions), and conduct data collection for the duration of the study.
4. Collect continuous baseline data (A) on target behaviors for a minimum of 3 consecutive days or until data are stable.
5. Introduce Intervention (B) only after data stability has been established in the initial baseline (A) condition.
6. Collect continuous data during Intervention (B) on target behaviors for a minimum of 3 consecutive days or until data are stable, and continue to monitor non-target behaviors on a regular schedule.
7. After a stable data pattern occurs under Intervention B, withdraw Intervention B and introduce Intervention (C).
8. Collect continuous data during Intervention (C) on the target behaviors for a minimum of 3 consecutive days or until the data are stable, and continue to monitor non-target behaviors on a regular schedule.
9. Repeat Steps 6–8.
10. Replicate with similar participants; and counterbalance the order of implementing interventions across participants.

Internal Validity

Internal validity in multitreatment designs is evaluated and strengthened in manners similar to that used for A-B-A-B withdrawal designs (see Chapter 9). Experimental control is demonstrated when internal validity is adequate and change occurs *when and only when* conditions change. Typical threats to internal validity, including history, maturation, instrumentation, and procedural fidelity are similarly likely in multitreatment and withdrawal designs; typical procedures for detecting and controlling for these threats should be used.

Two threats are likely due to the comparative nature of multitreatment designs. These include *separation of treatments* and *multitreatment interference*. Because both interventions in a multitreatment design are applied to the same dependent variable, you cannot attribute the ultimate levels of behavior change to a single intervention. However, you draw conclusions about experimental (causal) relations about differences *between conditions*. Also, because interventions are applied to the same behavior, multitreatment interference is likely in the form of slow-paced sequence effects. When multitreatment interference is suspected, extending the length of a condition will help control for this effect. This is illustrated in Figure 11.3; Intervention B resulted in an increase in level and an accelerating trend. When Intervention C was used, the first few data points had values similar to those in B. By extending the C condition, the data subsequently dropped, indicating multitreatment interference may have been operating

when Intervention C was initiated.

Advantages

Multitreatment designs are flexible, making them useful for a variety of important comparisons with reversible behaviors. They can be used to compare different interventions (e.g., A-B-C-B-C); in component analyses to build interventions (A-B-BC-B-BC) or to take treatment packages apart (A-BCD-BC-BCD-BC-B-BC-B); and in studying parametric variations of an independent variable (A-B-B'-B-B'). Multitreatment designs are useful with a variety of different types of interventions, such as environmental arrangements (room arrangements, manipulations of materials), consequence-based interventions (contingencies), and antecedent-based interventions (e.g., self-monitoring, rule statements). The designs can be used when the goal is to accelerate or decelerate target behaviors. These designs only require measurement of one reversible behavior.

Applied Example 11–1: A-B-C-B-C Design

State, T. M., & Kern, L. (2012). A comparison of video feedback and in vivo self-monitoring on the social interactions of an adolescent with Asperger syndrome. *Journal of Behavioral Education*, *21*, 18–33.

In this study, the effectiveness of two interventions to decrease inappropriate noises and other inappropriate social interactions were compared using an A-B-C-B-C multitreatment design. Sessions occurred in the participant's home; additional sessions occurred at school. The participant was a 14-year-old male with Asperger syndrome. The dependent variable was the percentage of intervals including inappropriate noises and inappropriate interactions, collected via video using 15-second partial interval recording. Appropriate social interactions were measured as a corollary behavior and graphed separately (not shown). Interobserver agreement data were collected during 30% of sessions, distributed across conditions. Mean agreement was very low for appropriate interactions (72%), relatively low for social interactions (85%), and higher for inappropriate noises (91%).

During all sessions, Carl and his teacher engaged in playing a game for a 15-minute session. During baseline sessions, Carl and his teacher played a game and she was instructed to interact with Carl as usual. After this condition, a video feedback condition (B) was implemented. In this condition, Carl watched a video of each session on the following day (prior to the next session). As he watched the video, he recorded whether or not he engaged during appropriate social interactions using 15-second partial interval recording. He received points both for engaging in appropriate social interactions and for accurately self-recording data ("matching" with a second data collector). Following the first video feedback condition, an in-vivo self-monitoring condition was implemented (C). During this condition, Carl self-recorded appropriate interactions during the game using 1-minute whole interval recording, cued by a vibrating watch. The points procedure (for engaging in appropriate social interactions and accurately self-recording) was the same as in the video feedback condition. Following this condition, the video feedback (B) and in-vivo self-monitoring (C) conditions were repeated. Authors do not report collection of fidelity data. A considerable strength of the study is the assessment of acceptability by the direct consumer (Carl).

As shown in Figure 11.5, variable and increasing inappropriate noises and interactions occurred during baseline with no apparent change in level or variability during the first video feedback condition. However, when the in-vivo self-monitoring condition was implemented, a decrease in level and variability occurred, with inappropriate noises and interactions generally occurring for fewer than 20% of intervals. The re-implementation of the B condition resulted in

increased level and variability (similar to the first B condition), and the re-implementation of the C condition resulted in decreased level and variability, with near-0 levels of inappropriate behavior for the last 8 sessions. Thus, in-vivo self-monitoring consistently led to fewer intervals with behavior occurrences when compared with video feedback. Given relatively small and variable changes between conditions, the low reliability data and lack of fidelity data might reduce confidence in the functional relation conclusion.

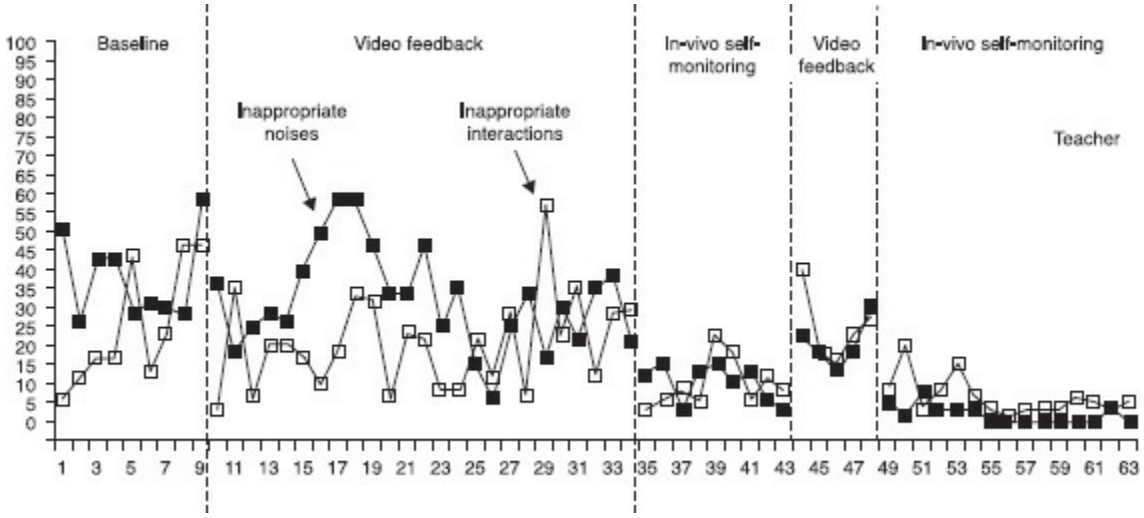


Figure 11.5 Multitreatment (A-B-C-B-C) design.

Source: State, T. M., & Kern, L. (2012). A comparison of video feedback and in vivo self-monitoring on the social interactions of an adolescent with Asperger syndrome. *Journal of Behavioral Education*, *21*, 18–33.

Limitations

Multitreatment designs can only be used when the dependent variable of interest is a reversible behavior. They are not useful for evaluating strategies to promote acquisition of new behaviors. Sequence effects are likely, and multitreatment designs do not solve the separation of treatments issue. The design also requires a long time to complete producing an increased risk of important threats to internal validity, including instrumentation (observer drift and bias), procedural infidelity, maturation, history, and attrition.

Conclusions

The multitreatment design can be used to evaluate the relative effectiveness of two interventions for changing reversible behaviors. As with the A-B-A-B design, the multitreatment design provides a convincing demonstration of differences between conditions; it is also flexible in that additional conditions can be added if needed. Although experimental control is demonstrated for single participants when

multitreatment designs are used, we recommend that multiple participants be recruited to improve external validity.

Alternating Treatments Design (ATD) and Multi-element Design (M-ED)

The **alternating treatments design** (ATD, Barlow & Hayes, 1979) and the multi-element design (M-ED; Ulman & Sulzer-Azaroff, 1975) are procedurally similar, so we discuss them together. The primary difference is that the ATD is used to compare *interventions* while the M-ED is used to *assess* factors that may be maintaining challenging behavior. When M-EDs are used, four or five conditions are used (compared) and none may be considered an intervention. Often, the terms "alternating treatments" and "multi-element" are used interchangeably. We will primarily discuss the ATD variation; because the ATD and the M-ED are experimentally identical, guidelines and internal validity considerations are the same.

The ATD uses rapid and repeated manipulation of at least two conditions (Horner et al., 2005). In other words, the compared conditions are alternated across sessions or days —thus the session sequence in an ATD might look something like this:

A-B-B-A-A-B-B-A-B-A-A-B-A-B

While the session sequence in a withdrawal design would look something like this:

A-A-A-A-B-B-B-A-A-A-B-B-B-B.

Studies using ATDs do not require an extended time, and the ATD is one of only a few designs in which you can simultaneously compare more than two conditions, making it useful to practitioners and researchers.

The ATD requires measurement of one *reversible behavior*; additional behaviors can be measured as secondary or corollary measures. One behavior should be determined to be the primary dependent variable prior to the start of the study; this behavior is used to make experimental decisions. The design is useful only for reversible behaviors. ATDs can be used to assess the effects of interventions designed to increase or decrease behavior occurrence; it is more often used to assess intended decreases compared to other designs—only about 18% of designs in special education journals identified a primary dependent variable intended to decrease with intervention, but more than 1/3 of ATDs included those behaviors (Ledford et al., 2017).

The purpose of the ATD variation is to compare two or more interventions. The simplest ATD, which is appropriate for answering comparative questions, is depicted in Figure 11.6 in which Logan, Jacobs, Gast, Murray, Diano, and Skala (1998) evaluated the effects of small group composition (typical developing peers compared to peers with disabilities) on the frequency of "happiness behaviors" (smiling, eyes open) of five primary-age children with profound multiple disabilities. Data were collected using a 10-

second partial interval recording procedure, alternated with a 5-second recording period, during small group activities (gross motor game, music, art) in which time of day, teacher behavior, activities, materials, and number of peers in the group were controlled. The independent variables, group of typical peers *versus* a group of peers with disabilities, were randomly scheduled across days with no more than two consecutive days of the same group composition. As shown in Figure 11.6, smiles/open eyes were recorded in a higher percentage of intervals for all five children when the group was comprised of typical peers rather than peers with disabilities. This simple variation of the ATD is well suited for studying similar comparative research questions; note that this comparison could also be considered baseline versus intervention, especially since the "disabilities only" group was typical for the children given their placements in self-contained classrooms.

The ATD is flexible, making it useful for many purposes. A major purpose is to assess factors maintaining challenging behaviors (M-ED variation). The intent of such assessments is to identify interventions or characteristics of interventions to treat those behaviors. Often this is done in the context of a broader functional assessment (Dunlap et al., 2006) and may use an analogue functional analysis (Iwata, Dorsey, Slifer, Bauman, & Richman, 1994). Such assessments often include four or five conditions designed to identify the motivating operation maintaining problem behavior, including contingent attention, contingent receipt of a tangible item, and escape from demands (Neef & Peterson, 2007). Other conditions, of course, can be included and individualized to the participant, behavior, and suspected maintaining variables. The conditions (independent variables) are not viewed as interventions; rather, the question is, "Which of the independent variables can be used to devise an intervention?" In general, interventions based on assessments using the M-ED variation of the ATD result in more effective interventions than when interventions are devised without such assessment (Herzinger & Campbell, 2007).

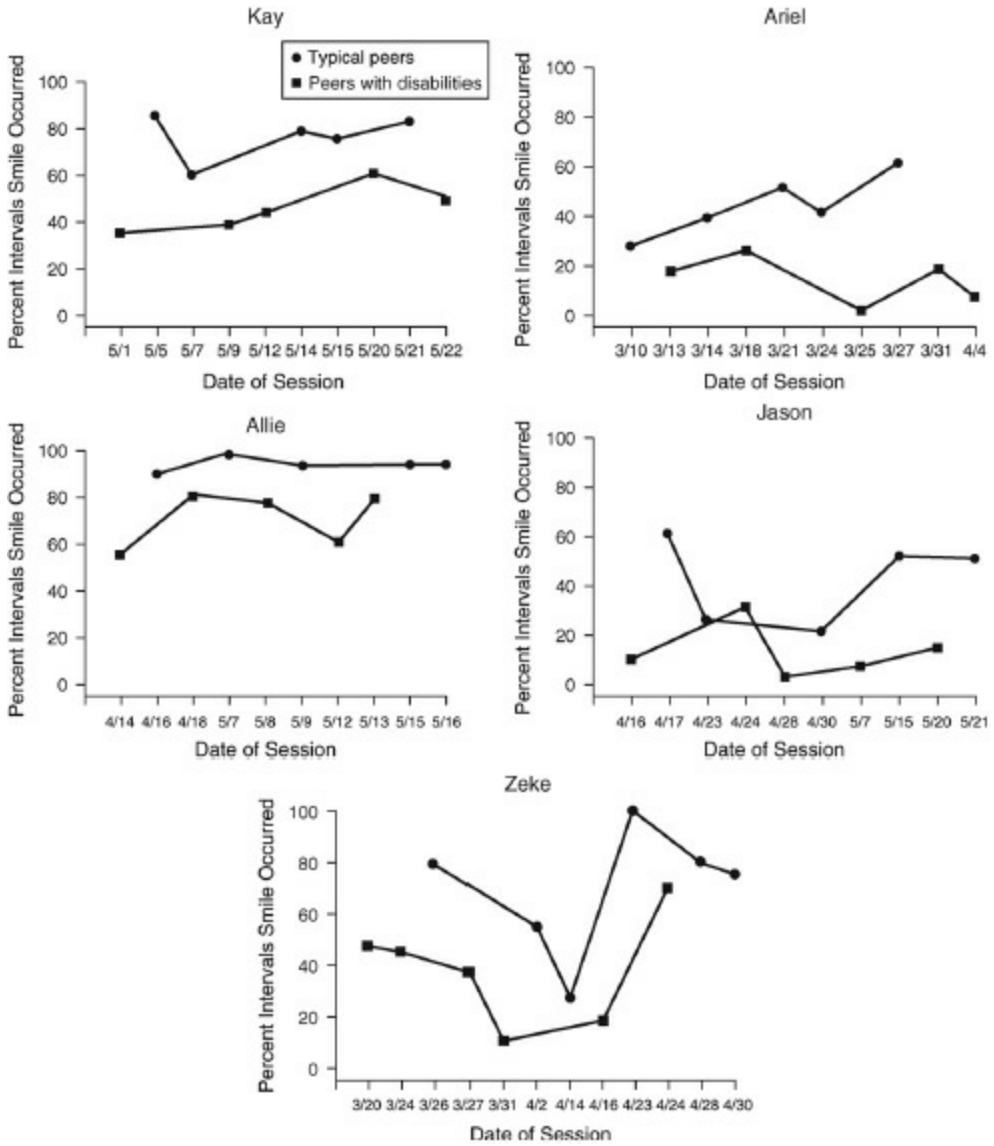


Figure 11.6 ATD designs without baseline and best alone conditions.

Conditions

Barlow and Hayes (1979) described the ATD as having four experimental "phases": Phase 1 (baseline), Phase 2 (comparison of independent variables), Phase 3 (use of superior treatment alone), and Phase 4 (follow-up). We will call all of these conditions, consistent with our usage in the rest of the book; this design is different because the conditions that are being experimentally compared are actually within a condition (the comparison condition). To avoid this potentially confusing conditions-within-condition terminology, it is acceptable use the term phase to denote the sequential conditions (baseline, comparison, best alone, follow-up). When the ATD is used, the baseline and best alone conditions are optional, but recommended. If included, the baseline condition involves repeated measurement across consecutive sessions/days under baseline procedures and it helps to describe the participant's pre-intervention performance and the need for the

intervention, although the *initial* baseline condition is not included in the experimental comparison. It is preferable for the baseline condition to remain in effect until data are stable, but stability is not a requirement because the initial baseline condition is not part of the experimental comparison.

In Figure 11.7, five variations of the ATD are shown with hypothetical data: (a) all four conditions described by Barlow and Hayes (1979), (b) a study with a baseline and comparison condition without the superior treatment alone condition, (c) a study with the comparison condition and the superior treatment alone condition, (d) a study with only the comparison condition and two interventions, and (e) a study with four interventions without a baseline or superior treatment alone condition (M-ED variation). On the left side, note that each data point occurs at a different point in time (e.g., each data point corresponds to a session). If each condition is conducted within a day, you could also graph by day (right side). We do not recommend this graphing variation because it is not possible to detect potential differences based on ordering (e.g., a child is less engaged in the third session of the day, regardless of condition, a type of maturation threat). Thus, even when multiple conditions occur per day, we recommend conserving the temporal order in the graph (as shown in the left panel). All conditions can be completed in a relatively short time period, especially when multiple sessions occur each day. The final condition, follow-up, is seldom included in ATD studies.

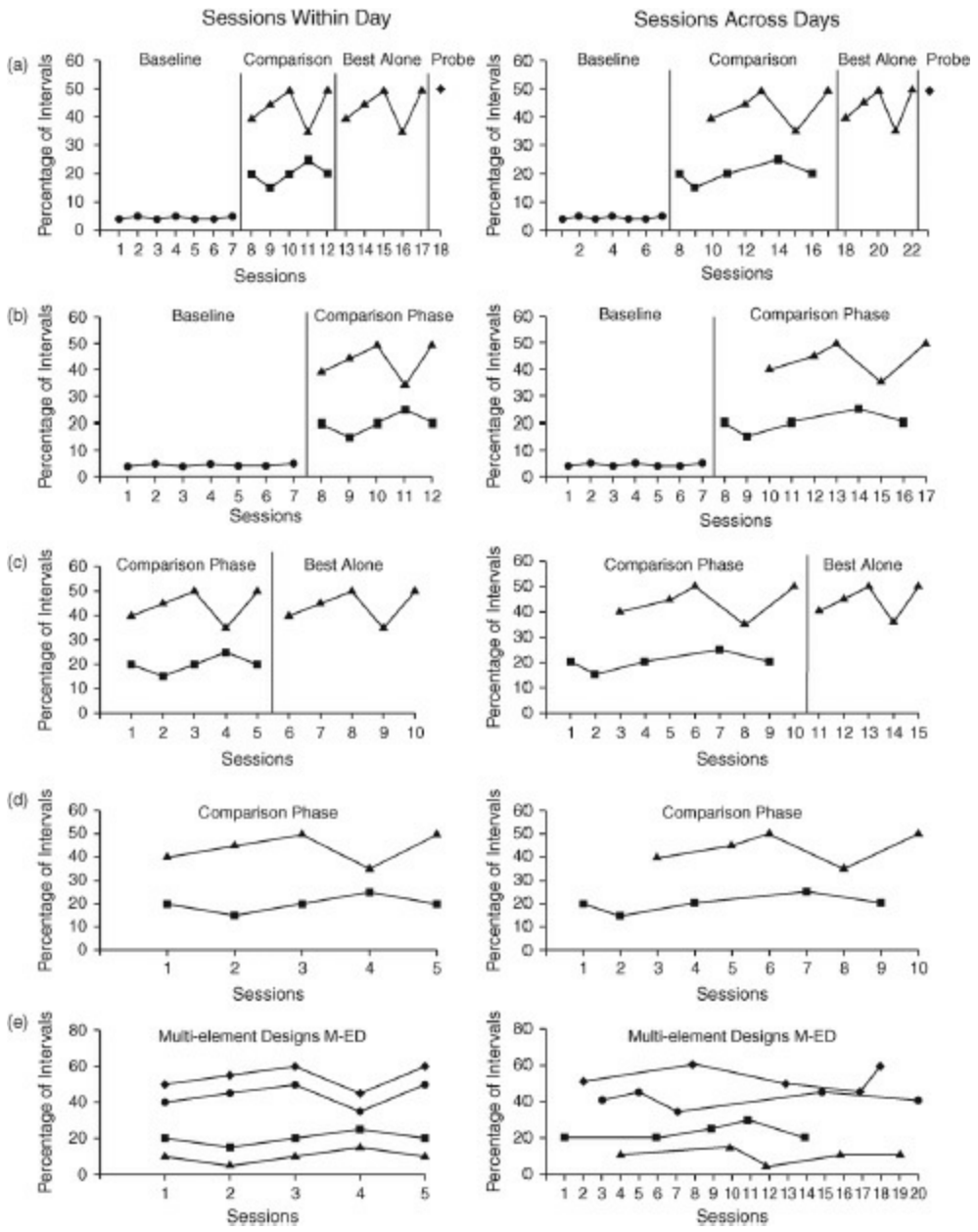


Figure 11.7 Prototypes of ATD variations with and without temporal order preserved.

In the comparison condition, data patterns of each condition are compared to one another. A conclusion that one intervention is more "effective" than other(s) is made when differentiation between data paths consistently occurs, in a therapeutic direction. As with other SCDs, evidence of an effect is based on replications of findings. With the ATD, demonstrations occur within the comparison condition. With each change of conditions in alternating sessions or days, another replication occurs (e.g., the first demonstration is between Data point 1 for Condition A and Data point 2 for Condition B). More alternations are generally better (i.e., 5 or more), but when a clear difference exists, more alternations may add relatively little value.

Because of the rapid alternation of conditions across observations, a special requirement of the ATD is that a participant must discriminate which condition is in

effect in any given session (i.e., stimulus discrimination). This requirement is less of an issue with interventions based on antecedent practices, environmental arrangements, or material modifications. It is, however, a major concern when the intervention is a consequence for a behavior—in which the behavior must occur before an intervention is used. A strategy for dealing with this is to tell participants which intervention is being used. However, with participants whose language is limited, simply telling them which intervention is in effect may not allow them to make the discrimination. With such participants researchers have used different colored lights, vests, visual representations, implementers, rooms, or other related stimuli to facilitate the discrimination as to which intervention is in effect. Table 11.4 summarizes selected studies using an ATD to evaluate behavior change.

Table 11.4 *Studies Using Alternating Treatment Designs (ATD)*

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
Chazin, K. T., Ledford, J. R., Barton, E. E., & Osborne, K. (2017). The effects of antecedent exercise on engagement during large group activities for young children. *Remedial and Special Education*. doi: 10.1177/0741932517716899	Number: 2 Sex: M (1), F (1) Age: 5 Disability/ diagnosis: None (1), Down syndrome (1)	Setting: Inclusive preschool Arrangement: Individual (intervention), large group (DV measurement)	ATD with baseline followed by three alternating conditions (continued baseline, seated activities, and exercise activities), followed by best alone for one participant	Percentage of intervals with on-task and out-of-seat behaviors; rate of challenging behavior

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
Hammond, J. L., & Hall, S. S. (2011). Functional analysis and treatment of aggressive behavior following resection of a craniopharngioma. *Developmental Medicine and Child Neurology, 53*, 369–374.	Number: 1 Sex: F Age: 6 Disability/ diagnosis: None; post-surgery aggression	Setting: Hospital Arrangement: Individual	M-ED with five alternating conditions (contingent escape, ignoring, contingent attention, contingent provision of tangible, no-demand play)	Aggressive behaviors per minute
Haydon, T., Conroy, M. A., Scott, T. M., Sindelar, P. T., Barber, B. R., & Orlando, A. (2010). A comparison of three types of opportunities to respond on student academic and social behavior. *Journal of Emotional and Behavioral Disorders, 18*, 27–40.	Number: 6 Sex: M Age: 7–8 Disability/ diagnosis: At risk for EBD	Setting: General education classroom Arrangement: Large group instruction	ATD with three alternating conditions (individual responding, choral responding, and mixed responding)	Disruptive behaviors per minute and percentage of intervals with off-task behaviors and active student responses
Ingersoll, B. (2011). The differential effect of three naturalistic language interventions on language use in children with autism. *Journal of Positive Behavior Interventions, 13*, 109–118.	Number: 2 Sex: M Age: 3 Disability/ diagnosis: Autism	Setting: Small treatment room Arrangement: Individual	ATD with three alternating conditions (responsive interactions, milieu teaching, and combined interventions)	Percentage of intervals of total language use (prompted or spontaneous requests or comments)
Kodak, T., Northup, J., & Kelley, M. E. (2007). An evaluation of the types of attention that maintain problem behavior. *Journal of Applied Behavior Analysis, 40*, 167–171.	Number: 2 Sex: 1 M; 1 F Age: 5–9 Disability/ diagnosis: ADHD (n=1), PDD-NOS (n=1)	Setting: Home or therapy room. Arrangement: Individual	M-ED with four alternating conditions: (contingent attention, demands, toy play, and alone)	Rate of problem behavior per minute
Lynch, A., Theodore, L. A., Bray, M. W., & Kehle, T. J. (2009). A comparison of group-oriented contingencies and randomized reinforcers to improve homework completion and accuracy for students with disabilities. *School Psychology Review, 38*, 307–324.	Number: 6 Sex: 2 M; 4 F Age Range: 10–11 Disability/ diagnosis: LD (n=4) or speech impairment (n=2)	Setting: Inclusive classroom Arrangement: Large group	ATD with baseline, followed by three alternating conditions (dependent, interdependent, and independent group contingencies), best alone	Percentage of students who completed homework

Note: M=male, F=female, ADHD=attention deficit hyperactivity disorder, PDD-NOS=pervasive developmental disorder, LD=learning disability, M-ED=multielement variation of the alternating treatments design.

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
Mueller, M. M., Sterling-Turner, H. E., & Moore, J. W. (2005). Towards developing a classroom-based functional analysis condition to assess escape-to-attention as a variable maintaining problem behavior. *School Psychology Review, 34*, 425–431.	Number: 1 Sex: M Age: 6 Disability/diagnosis: Autism	Settings: Self-contained classroom in a public school Arrangement: Individual	M-ED with three alternating conditions (contingent attention, contingent escape, and control), followed by an M-ED with three alternating conditions (contingent escape to attention, contingent escape, and control)	Percentage of intervals with tantrums
Neef, N. A., Cihon, T., Kettering, T., Guld, A., Axe, J. B., Itoi, M., & DeBar, R. (2007). A comparison of study session formats on attendance and quiz performance in a college course. *Journal of Behavioral Education, 16*, 235–249.	Number: 44 Sex: Not reported Age: Adult Disability/diagnosis: None	Setting: College classroom Arrangement: Group	ATD with three alternating conditions (Game study sessions, Q & A study sessions, baseline)	Mean percentage of correct quiz responses
Reinhartsen, D. R., Garfinkle, A. N., & Wolery, M. (2002). Engagement with toys in two-year-old children with autism: Teacher selection and child choice. *Journal of the Association for Persons with Severe Handicaps, 27*, 175–187.	Number: 3 Sex: M Age: 2 Disability/diagnosis: autism and DD	Setting: University-affiliated preschool Arrangement: Individual	ATD with two alternating conditions (Teacher-selected toy and child-selected toy)	Percentage of intervals of engagement and problem behavior
Simonsen, B. MacSuga, A., Fallon, L. M., & Sugai, G. (2013). The effects of self-monitoring on teachers' use of specific praise. *Journal of Positive Behavior Interventions, 15*, 5–15.	Number: 5 Sex: F Age: Adult Disability/diagnosis: None	Setting: General education (n=4) and special education (n=1) classrooms Arrangement: Typical classroom activities	ATD with baseline, followed by four alternating conditions (baseline, counting praise, tallying praise, calculating rate of praise), best alone	Rate per minute of specific praise
Travers, J. C., & Fefer, S. A. (2017). Effects of shared active surface technology on the communication and speech of two preschool children with disabilities. *Focus on Autism and Other Developmental Disabilities, 32*, 44–54.	Number: 6 Sex: 4 M, 2 F Age: 4–5 Disability/diagnosis: None (4), developmental delay (1), autism (1)	Setting: Empty inclusive classroom Arrangement: Small groups (triads)	ATD with two alternating conditions (drawing activities on active surface technology versus paper)	Percentage of intervals with social communication and nonsocial speech

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
VanDerHeyden, A. M., Snyder, P., Smith, A., Sevin, B., & Longwell, J. (2005). Effects of complete learning trials on child engagement. *Topics in Early Childhood Special Education*, *25*, 81–94.	Number: 3 Sex: M Age: 2–6 Disability/diagnosis: Moderate IDs (n=2), none (n=1)	Setting: Preschool classroom Arrangement: Individual	ATD with baseline, followed by three alternating conditions (providing a new toy, elaborating on toy play, and elaborating on toy play plus prompting)	Percentage of intervals of engagement

Note: ID=intellectual disability; M-ED=multielement variation

Procedural Guidelines

When using an ATD, adhere to the following guidelines:

1. Identify and define a reversible target behavior.
2. Select a sensitive, reliable, valid, and feasible data collection system and pilot the system and your behavior definitions.
3. Determine a priori frequency of reliability and fidelity data collection (e.g., 33% of sessions), and conduct data collection for the duration of the study.
4. Determine what rules you will use for alternating conditions (e.g., random alternation with no condition repeating until all have been conducted; random alternation with no more than two consecutive sessions in a single condition). We suggest random alternation with or without restrictions.
5. Assign condition order (e.g., randomize).
6. Collect data for the initial baseline condition, if possible, for at least 3 sessions.
7. Begin the comparison condition. Conduct at least 5 sessions in each condition, regardless of behavior patterns. If possible, collect data in a continuing baseline condition in addition to any intervention conditions.
8. After initial comparison sessions have been conducted, use visual analysis to determine whether a functional relation exists, does not exist, or whether additional data collection are needed.
9. If needed, assign session order and collect additional data.
10. Repeat step 8.
11. Replicate with similar participants.

Internal Validity

Studies have adequate internal validity when all likely threats are controlled for, and experimental control is demonstrated when adequate internal validity is present and when consistent differentiation between two conditions is present. Differentiation

(usually in level; see Chapter 8) is assessed for pairs of conditions. Thus, if an ATD is used with a continuing baseline (A), and two interventions (B and C), comparisons are made regarding differentiation between A and B, A and C, *and* B and C.

Despite their widespread use, some threats to internal validity are particularly likely to be problematic when rapid alternation designs are used; common threats to internal validity are described below and shown in Table 11.5. Because of its relatively short duration, some threats to internal validity are less likely when compared with other designs, including history, maturation, instrumentation, testing, and attrition; nonetheless, typical procedures for detecting and controlling for these threats should be used (see Chapter 1 and Table 11.5). These threats can occur in any study; the relatively short duration of ATDs simply decreases the likelihood. Despite decreased likelihood for some threats, other threats may be more likely to occur. One threat that may be more likely when this design is used is procedural infidelity. It may be more difficult to maintain fidelity to differing conditions because of the frequency of change. You can prevent this threat by conducting training to a strict criterion (including training during all condition types). You can detect infidelity by conducting frequent fidelity checks in all conditions, and re-training as needed based on formative analysis of fidelity data. You may also need to provide implementers with "cheat sheets" to remind them of the critical features of the intervention to be implemented for each session.

Table 11.5 *Common Threats to Internal Validity, and Methods to Detect and Control for Threats, for ATDs and AATDs*

	ATDs and AATDs			
	Likelihood	**Detect**	**Control**	**Report**
History	Less likely due to short study duration	Visual analysis: Abrupt change in data that is an outlier compared to remaining data in the condition	If in baseline, continue until stable; if in comparison condition, continue until differences between phases are apparent	Describe anecdotally known conditions that may have attributed to non-experimental behavior change (e.g., illness)
Maturation	Less likely due to short study duration	Change in behaviors assigned to control condition (AATD only)	Maturation is not a concern as long as one condition is consistently superior to others, even if all conditions show slight improvements (ATD); use a control set (AATD)	Report participant performance on control sets relative to sets assigned to the intervention conditions (AATD)
Instrumentation	Less likely due to short study duration	Visual analysis: Differences between observers, particularly if one is blind	Use blind observers; carefully formulate and pilot definitions and recording systems; train observers to a criterion; have discrepancy discussions	Describe all reliability procedures and results; explicitly say whether observers were blind; describe reasons for low agreement
Procedural Fidelity	Likely due to alternation that requires implementers to change their behaviors often	Formative analysis of direct observational recording of fidelity data	Train implementers to criterion; re-train if necessary; provide supports to implementers such as reminder checklists	Describe all fidelity procedures and results, including training, supports, and re-training
Testing	Less likely due to short study duration	Visual analysis: Deteriorating or therapeutic trends in BL conditions	Design non-aversive baseline conditions; measure control stimuli intermittently (AATD only)	Describe likely testing threats, if applicable

	ATDs and AATDs			
	Likelihood	**Detect**	**Control**	**Report**
Attrition Bias	Less likely due to short study duration	Author report	Clearly describe alternating nature of phases to participants to reduce confusion	Describe attrition in written report and report all data from all participants, even if design was not completed
Adaptation	Likely when observations are apparent	Participant behavior changes over time during BL conditions	Continue BL until data are stable	Describe anecdotal evidence that BL change was due to adaptation; discuss degree to which later BL data are potentially more representative of "typical" behavior
Hawthorne Effect	Likely when participants are sensitive to perceived desirable behaviors	Participant behavior is inconsistent with expectations when study begins	Use covert measurement; do not implement comparison condition if BL data do not indicate need; continue data collection to determine whether effect is temporary and change conditions only when behavior are stable	Describe anecdotal evidence that BL change was due to Hawthorne effect; discuss degree to which later BL data are potentially more representative of "typical" behavior
Multitreatment Interference	Likely in the form of alternation effects; especially when multiple sessions are conducted close in time (e.g., same day)	Not detectable via visual analysis	Use a "best alone" final condition; if data are similar, MTI is unlikely	Describe best alone phase as a control for MTI and describe the extent to which data in this phase support presence or absence of MTI
Instability	No particular likelihood associated with this design	Considerable overlap between phases due to widely ranging values in one or more phases in the comparison condition	Discontinue comparison condition only if differences between phases are apparent	Describe degree to which data instability within phases impacted conclusions in the comparison condition
Irreversibility of Behaviors	High likelihood when a non-reversible behavior is chosen in ATD only	Reported DV in ATD is non-reversible behavior	Use AATD when non-reversible behaviors are of interest	When non-reversible behaviors are used in an ATD, no conclusions regarding relative effectiveness are possible

	ATDs and AATDs			
	Likelihood	Detect	Control	Report
Unequal Behavior Difficulty	High likelihood in AATD only	Across participants, one set of behaviors is consistently learned faster, regardless of intervention assignment	Include many participants and counterbalance behavior set assignment across participants (e.g., Set 1 is assigned to Intervention A for the first participant; Set 1 is assigned to Intervention B for the second participant); conduct many comparisons within or across participants with random assignment of behavior sets to intervention	Report the extent to which one intervention was consistently superior regardless of behavior set assignment; report how behavior sets were assigned to intervention type

Another potential threat is cyclical variability, a concern when sessions are alternated systematically but not randomly (e.g., Intervention A occurs every morning and Intervention B occurs every afternoon). It may be that this systematic alternation impacts data; for example, the child may typically engage in more challenging behavior in the afternoon due to factors external to the study (fatigue, academic content, differences in social opportunities). Randomly determining session order decreases the likelihood of this threat. Researchers often choose to do restricted randomization, such that, for example, no more than two consecutive sessions occur for a single condition (Douglas, Ayres, & Langone, 2015), such that each condition is repeated once before any are repeated (Chazin et al., 2017), or such that each condition occurs a certain number of times per week (Reichow, Barton, Sewell, Good, & Wolery, 2009).

Multitreatment interference is likely in the comparison condition of the ATD in the form of rapid alternation effects (Hains & Baer, 1989). The superior treatment alone condition is designed to deal with this threat to internal validity (Barlow & Hayes, 1979). If there are changes in level, trend, or variability between the initial baseline condition and baseline data in the comparison condition, multitreatment interference is probable. This detection leaves open the possibility that one of the interventions is influencing performance in sessions of other interventions as well as the baseline sessions. A strategy for dealing with this problem is to increase the amount of time between sessions (McGonigle, Rojahn, Dixon, & Strain, 1987). For example, if multiple sessions are conducted each day and multitreatment interference is detected, conducting only one session per day is advised. The third condition, best alone, is used when one of the compared interventions produces a more therapeutic data pattern than the other(s). The intervention producing the more therapeutic data pattern is used without alternating it

with other interventions or the baseline condition.

Reversibility of behavior is critical in ATD studies. The behavior must be one that is influenced by the immediate change in conditions. If this is not the case, a totally benign intervention will seem to produce an intervention effect because the effective intervention would move the behavior to new levels with each application. If the behavior did not readily reverse, then the benign intervention's data would be at the same level as the previous session with the effective intervention.

Advantages

The ATD has three major advantages. First, it provides a rapid method for evaluating two or more interventions or two or more variations of an intervention. The benefits of this rapidity are (a) less investigator time is spent conducting the study, (b) fewer resources are used, (c) less participant time is devoted to study activities, and (d) some threats to internal validity are minimized. Second, the M-ED variation allows for efficient assessment of factors maintaining participants' problem behavior and is useful for selecting successful interventions. This assessment information is also a solid foundation for making generalizations to other non-study individuals. A third positive feature of the ATD is its flexibility. It can be used with a wide range of interventions and in several different variations (e.g., with and without initial baseline conditions, with a continuing baseline condition).

Limitations

Limitations of the ATD include: (a) It is restricted to reversible behaviors, (b) multitreatment interference can emerge from rapidly alternating interventions across sessions/days, and (c) it provides little information about the effects of an intervention from repeated and continuous use due to its relatively short duration.

Conclusions

The ATD is designed to evaluate interventions for changing reversible behaviors in a relatively short amount of time. It is one of a few designs that can be used to answer both demonstration and comparison questions (i.e., when a continuing baseline condition is used). Although experimental control is demonstrated for single participants when ATDs are used, we recommend that multiple participants be recruited to improve external validity. Additionally, unless researchers have a compelling rationale for not doing so, we recommend randomizing the order of conditions.

Adapted Alternating Treatments Design (AATD)

The adapted **alternating treatments design (AATD)** was developed to compare instructional practices with *non-reversible behaviors* (Sindelar, Rosenberg, & Wilson, 1985). The AATD is useful when comparing interventions for teaching functional, developmental, or academic behaviors. A wide range of strategies can be studied, but the purpose must be to facilitate acquisition of new behaviors. When comparing two different instructional strategies with the AATD, interventions should have been studied sufficiently with demonstration designs to document they are effective. A major use of the AATD is to compare the *efficiency* of instructional strategies. The definition of efficiency has two dimensions. First, to be efficient, a strategy must reliably produce learning (be effective). Second, to be efficient, a strategy must be superior to another strategy on an important dimension. Common dimensions of superiority include (a) rapidity, (b) extent of maintenance and generalization, (c) breadth of learning (e.g., learning two things rather than one), (d) acquisition of untrained relations, and (e) influencing future learning (Wolery, Ault, & Doyle, 1992). The most commonly-measured dimension of efficiency is the rapidity of learning. It often is assessed by comparing number of minutes of instruction, number of sessions or trials, number and percentage of errors, and number of trials or sessions to criterion.

Applied Example 11–2: ATD

Reichow, B., Barton, E. E., Sewell, J. N., Good, L., & Wolery, M. (2010). The effects of weighted vests on the engagement of children with developmental delays and autism. *Focus on Autism and Other Developmental Disabilities*, *25*, 3–11.

The purpose of this study was to assess the effects of a sensory-based treatment, weighted vests, to baseline conditions. Participants were three preschool children (4–5 years of age) who had autism or developmental delays. They wore weighted vests regularly and their teacher perceived weighted vests as effective (e.g., related to a positive outcome for the child).

In this study, an ATD with two baseline (or control) conditions was utilized: a no-vest condition and a condition in which a vest was worn without weights. Importantly, the "no weights" condition allowed for blind coding of behavior (e.g., was designed so that the observer could not detect whether the child was wearing a weighted or unweighted vest). Note that the baseline (no vest) comparison was used in an initial baseline condition *and* was continued through the comparison condition. Thus, authors were able to compare no-vest to unweighted vest, no-vest to weighted vest, and weighted and unweighted vests.

All sessions were conducted during large group activities in classrooms in a university- based early childhood center. During each school week, two weighted vest sessions, two unweighted vest sessions, and one baseline session were conducted; researchers used random assignment to determine order of sessions within the week.

Child behavior was coded via video using 10 second momentary time sampling. Three dependent variables were presented: Percentages of intervals of engagement, stereotypy, and challenging behaviors. Along with two other codes (unengaged and not visible), these codes were mutually exclusive and exhaustive (e.g., for every interval, one behavior was coded).

Interobserver agreement data were collected for at least 27% of sessions for all participants. Reliability was calculated using point-by-point agreement and agreement across conditions exceeded 90% for all participants. No fidelity data were reported.

Results for one participant are shown in Figure 11.8. Stereotypy and challenging behavior occurred at near-zero levels across baseline and comparison conditions; no functional relation is demonstrated for these behaviors. Note that the inclusion of the initial baseline condition controls for multitreatment interference threats (e.g., we can conclude confidently that these behaviors were not suppressed only due to the rapid iterative alternation of conditions). Engagement was highly variable during initial baseline conditions, ranging from about ¼ to ¾ of intervals. Likewise,

within and across conditions in the comparison condition (weighted, unweighted, and baseline), data were variable and overlapping. Thus, no functional relation was identified between the use of weighted vests as compared with baseline (no vest conditions) and no functional relation was identified between the use of weighted vests as compared with unweighted vest conditions.

Authors collected data from blind raters who used a Likert-type scale to assess engagement and stereotypy via video. Interestingly, they rated Bert's engagement higher and stereotypy lower in baseline (no vest) and unweighted conditions compared to weighted vest conditions. Authors do not offer potential reasons for the discrepancy between social validity data (which suggested weighted vests might result in negative outcomes) and experimental data (which suggested null effects). Because unweighted vests were used as a comparison, it is unlikely that the results or discrepancies were due to bias, although some measurement error is expected when MTS is used (see Chapter 5). Because adult behaviors (fidelity) were not measured, it may be that discrepancies were related to differences in content, opportunities to respond, or some other unmeasured variable. This may be particularly true since measurement occurred in the context of a typical classroom large group activity, rather than a controlled clinical context.

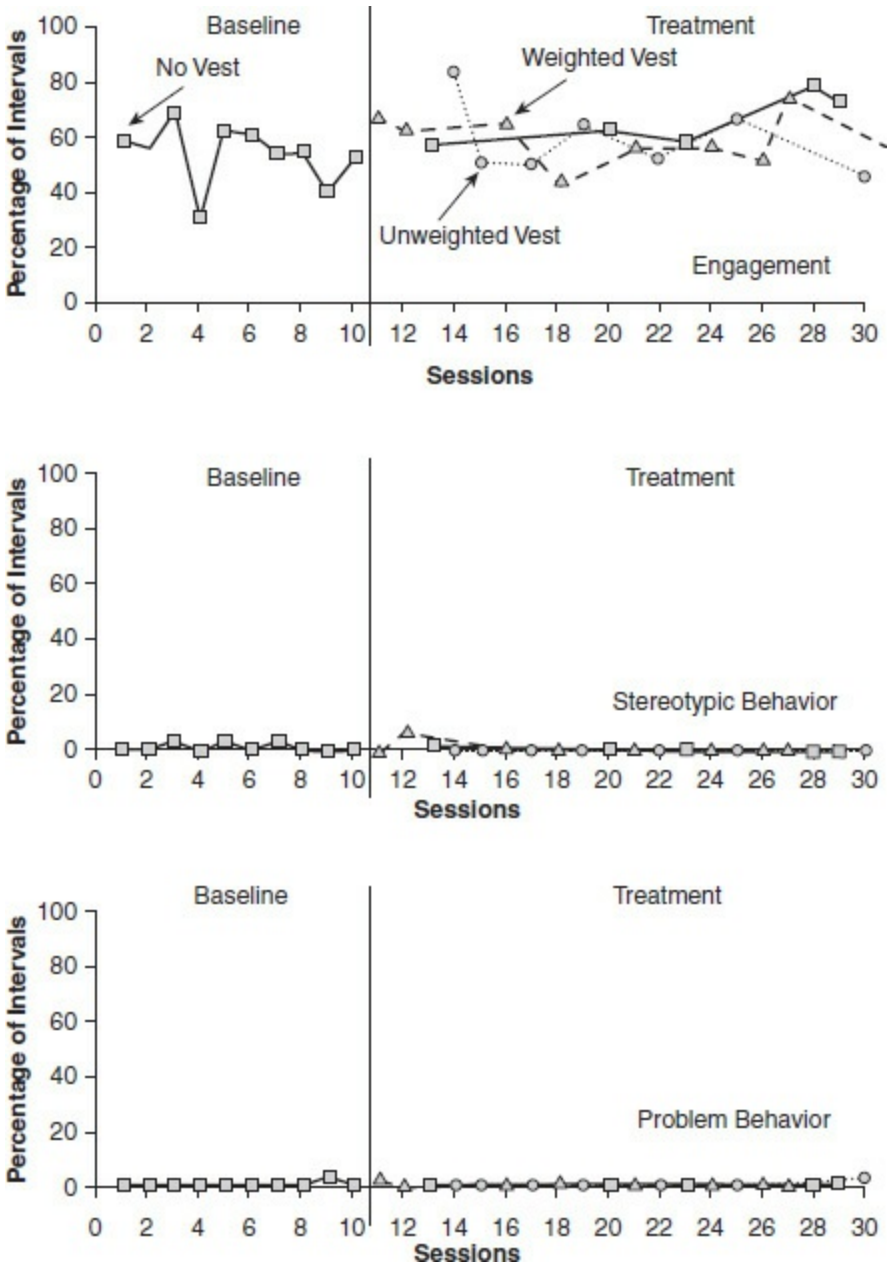


Figure 11.8 Alternating treatments design.

Source: Reichow, B., Barton, E. E., Sewell, J. N., Good, L., & Wolery, M. (2010). The effects of weighted vests on the engagement of children with developmental delays and autism. *Focus on Autism and Other Developmental Disabilities*, *25*, 3–11.

The AATD also can be used to refine an intervention, including component analyses. For example, a number of studies documented the effectiveness of a procedure called instructive feedback (Werts, Wolery, Holcombe, & Gast, 1995). It involved presenting extra non-target information in praise statements during direct instruction but not asking students to respond to that information. When this was done, students learned a great deal of the extra information. In initial studies, the extra non-target information always had been related to target behaviors. Werts, Wolery, Holcombe, and Frederick (1993) used the AATD to compare two conditions: one in which instructive feedback

information was related to target behaviors and one in which it was unrelated. The AATD also can be used to study parametric variations of an intervention. For example, Holcombe, Wolery, and Snyder (1994) used the AATD to compare two high and low levels of procedural fidelity for an instructional strategy.

When the AATD is used, *independent variables are each applied to different behavior sets or behavior chains*. This makes the AATD different from the ATD in which all interventions are applied to the same behavior. A behavior set is a collection of discrete responses (single behaviors of relatively short duration); for example, a list of 5 words to be read, 10 mathematic problems, or a list of 10 facts to be learned. Response chains are a series of behaviors that when sequenced together form a complex skill, such as completing a long division problem, putting on a garment, setting a table, making a purchase at a store, or cooking a meal. Target behaviors in AATD studies must meet five criteria. First, behaviors must be non-reversible—participants are likely to continue to perform the behaviors accurately after instruction has stopped. Second, behaviors should not be in the participants' repertoire. Third, behaviors must be independent, meaning one behavior set/chain can be acquired without influencing performance on other sets/chains. Fourth, behaviors must be functionally similar, meaning behaviors are likely to be influenced by the same environmental variables (e.g., the instructional strategies being studied). Finally, *behavior sets/chains must be of equal difficulty*. This last criterion is challenging but extremely important. Behavior sets/chains must be of equal difficulty because the instructional strategies are applied to separate behavior sets or response chains. If the behavior set taught with one strategy was easier than the behavior set taught with the other, the test of the two interventions would be unfair. Before the study, you must select behavior sets and ensure they are of equal difficulty for each individual.

Selecting Behaviors of Equal Difficulty

Several methods exist for ensuring behavior sets/chains are of equal difficulty (Romer, Billingsley, & White, 1988). A convincing method is an *experimental evaluation* of the difficulty of behavior sets/chains. This can be accomplished by teaching behaviors to non-participant individuals who are similar to participants who will be recruited for the actual study using the same intervention and the following assumption: If the behavior sets/chains are of equal difficulty, then the same procedure should require the same amount of instruction to establish criterion level responding. This is a time consuming and expensive approach, and it does not take into account variation due to a participant's learning history. Another method is to select behaviors from pools of responses for which *norms* exist. For example, reading, spelling, and other academic behaviors are often listed by grade level and often by divisions within the grade. This method is weak, because a good deal of variability in difficulty can exist even within the same grade level and the same segment of a grade level. A third method is to conduct a *logical analysis* of the difficulty of the responses and discriminations required to perform

correctly. This method is perhaps the most commonly used; you should report the dimensions on which behaviors were logically analyzed. For example, if the target behavior was reading sight words, the logical analysis would focus on (a) number of syllables in each word, (b) configuration of the words, (c) initial consonants, (d) part of speech for each word, (e) any redundant letters across words, (f) the participant's knowledge of the referent of the word, and (g) participant's ability to say each word. Yet another method is to ask *experts to rate* the difficulty of potential target behaviors. When using this method, consult with multiple experts independently and exclude the behaviors on which they disagree.

Another method is to evaluate participants' performance on *related behaviors*. For example, a study may compare two procedures in teaching preschoolers to name pictures (i.e., an expressive language task). In this case, you should select only pictures the children cannot initially name. You also should assess their ability to point to the correct picture when presented with an array of four pictures and you say, "Point to (*name of a picture)*" (i.e., a receptive language task). If children accurately and consistently point to some pictures but not others, then those to which they can point are not of equal difficulty to those to which they cannot accurately point. You should select only those pictures on which the child was correct at chance levels, or only those pictures for which the child was correct at 100% on the receptive language task. However, having pictures the child could point to when named in one set and those the child could not point to correctly would result in unequal difficulty of sets. The above methods are not mutually exclusive. Combinations of the methods should be used in the same study. Ensuring equal difficulty of behavior sets/chains is fundamental to conducting a fair comparison with the AATD. As a result, plan extra time at the beginning of the study to document carefully that the behavior sets/chains are of equal difficulty. Simply showing a participant cannot perform behaviors is not sufficient to document the behaviors are of equal difficulty.

When multiple participants are taught the same behavior sets/chains, another option is available: Behaviors can be assigned randomly or counterbalanced across strategies. This practice is highly recommended if the same behaviors are taught to two or more participants. For example, given three participants and three sets of vocabulary words, you should assign Set 1 to Intervention A for the first participant, Set 1 to Intervention B for the second participant, and Set 1 to the control condition for the third participant.

Conditions

In most cases, the AATD has three sequentially implemented conditions. The first is an initial probe (baseline) condition in which all behavior sets/chains are assessed in multiple sessions. This condition is similar to the multiple probe design (conditions variation, see Chapter 10). At least three baseline sessions are needed, but more may be necessary to ensure data are stable. Initial probe condition sessions should be about as long and contain a similar number of trials as later instructional sessions. Ideally,

behaviors of the different sets would be assessed in the same probe sessions during the initial probe condition (i.e., intermixed sets). In probe sessions, correct responses should be reinforced to avoid artificially deflating correct responding. In addition, researchers often intersperse some known behaviors and deliver reinforcers for those behaviors. As with the ATD, a continuing baseline condition is preferable. When the AATD is used, this consists of assigning one set of behaviors to a control condition, and measuring this set intermittently throughout the study, under non-instructional conditions. Thus, you should alternate instructional sessions with intermittent probes for stimuli assigned to the control set (see Procedural Guidelines below).

The second experimental condition is a comparison condition in which the instructional strategies are applied to their respective assigned behavior sets/chains in alternating sessions. *Unlike the ATD, instructional strategies are applied to separate behavior sets/chains.* These sessions may be alternated across days or both sessions can occur in a single day. All aspects other than the instructional strategies should be identical across sessions with different interventions. Examples of such variables are the instructor, reinforcers, type of materials (unless the independent variable is about type of materials), session length, and setting in which the sessions occur. Any variables that are different across sessions may separately or in combination with the instructional strategies be responsible for differences in the data. The instructional comparison condition usually continues until behaviors meet a predetermined criterion level for each intervention. A common criterion is three consecutive sessions of 100% unprompted correct responding. If one strategy produces criterion level responding before the other, then periodic review trials/sessions can be conducted for the behavior set/chain which is at criterion. When one strategy produces criterion level responding before the other, you must decide how long you will continue to use the less effective strategy if it does not also reach criterion. An acceptable guideline is 1.5 or 2 times the number of sessions it took the effective strategy to reach criterion. Intermittently during this condition, you should collect data on responding for stimuli assigned to the control set (i.e., intermittent probes).

The final study condition is a probe condition in which all behavior sets/chains are assessed, including the control set/chain. The procedures of this condition should be identical to those of the initial probe condition. This condition tests the extent to which behaviors maintain in the absence of the intervention, and in intermixed rather than separated sets. Table 11.6 shows a list of representative studies using an AATD.

Table 11.6 Studies Using Adapted Alternating Treatment Designs (AATD) With Control Conditions

Reference	Participants	Setting/Arrangement	Independent Variable	Dependent Variable
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.	Number: 12 Sex: 3 M, 9 F Age Range: 3–5	Setting: Inclusive early childhood program Arrangement: Individual	Baseline, followed by comparison of two conditions (massed and embedded instruction)	Percentage correctly answered questions
Mechling, L. C., & Ayres, K. M. (2012). A comparative study: Completion of fine motor office related tasks by high school students with autism using visual models on large and small screen sizes. *Journal of Autism and Developmental Disorders, 42,* 2364–2373.	Number: 4 Sex: M Age Range: 19–21 Disability/diagnosis: ASD	Setting: Self-contained classroom in a public school Arrangement: Individual	Baseline, followed by comparison of two conditions (video modeling on small and large screens), best alone	Percent of correctly completed tasks

Reference	Participants	Setting/Arrangement	Independent Variable	Dependent Variable
Reichow, B., & Wolery, M. (2009). Comparison of everyday and every-fourth-day probe sessions with the simultaneous prompting procedure. *Topics in Early Childhood Special Education, 29*, 79–89.	Number: 4 Sex: 1 M, 3 F Age Range: 4–5 Disability/diagnosis: None n=(3), SI (n=1)	Setting: University-affiliated preschool Arrangement: Individual	Baseline, followed by two alternating conditions (simultaneous prompting with everyday or every-fourth-day probes)	Percentage correct responses
Savaiano, M. E., Compton, D. L., Hatton, D. D., & Lloyd, B. P. (2016). Vocabulary word instruction for students who read braille. *Exceptional Children, 82*, 337–353.	Number: 3 Sex: 2 M, 1 F Age Range: 9–12 Disability/diagnosis: Blind (3), LD (2), OHI (1), ASD (1)	Setting: Special school for individuals with visual impairments Arrangement: Individual	Baseline, followed by two alternating conditions (flashcard and auditory), followed by best alone for all three sets (including control) and maintenance	Definition recall total score
Singleton, D. K., Schuster, J. W., Morse, T. E., & Collins, B. C. (1999). A comparison of antecedent prompt and test and simultaneous prompting procedures in teaching grocery words to adolescents with mental retardation. *Education and Training in Developmental Disabilities, 34*, 182–199.	Number: 4 Sex: 3 M, 1 F Age Range: 15–19 Disability/diagnosis: Moderate ID	Setting: Self-contained classroom Arrangement: Individual	Baseline followed by two alternating conditions (simultaneous prompting and antecedent prompt and test)	Percentage of correctly read words
Viel-Ruma, K., Houchins, D., & Fredrick, L. (2007). Error self-correction and spelling: Improving the spelling accuracy of secondary students with disabilities in written expression. *Journal of Behavioral Education, 16*, 291–301.	Number: 3 Sex: M Age Range: 16–18 Disability/diagnosis: LD	Setting: Resource classroom Arrangement: Small group	Three alternating conditions (baseline, traditional repeated practice, and error self-correction)	Percentage correctly spelled words

Note: ASD=autism spectrum disorder, SI=speech impairment, LD=learning disability, OHI=other health impairment, ID=intellectual disability

Procedural Guidelines

When using an AATD, adhere to the following guidelines:

1. Identify and define several nonreversible behavior sets/chains that are independent and functionally similar.
2. Select a sensitive, reliable, valid, and feasible data collection system and pilot the system and your behavior definitions.
3. Determine a priori frequency of reliability and fidelity data collection (e.g., 33% of sessions), and conduct data collection for the duration of the study.

4. Randomly assign one behavior set/chain to each treatment and to the control condition.
5. Establish a learning criterion and a criterion for stopping the comparison if one of the compared instructional strategies is not effective.
6. Determine what rules you will use for ordering conditions; unlike with ATDs, it is critical to have an equal number of sessions that are roughly evenly spaced—thus, authors typically randomly select one condition and then automatically conduct the other condition for the next session.
7. Conduct an initial probe condition by collecting data on all behavior sets/chains for a minimum of 3 sessions or until data are stable for each behavior set/chain.
8. Implement the comparison condition by applying each strategy to its respective behavior set/chain in alternating sessions. If multiple sessions are conducted each day, use counterbalancing to detect effects of time of day.
9. Collect intermittent control set data. There are two strategies for collecting these data:
    a. Conduct separate control (probe) sessions intermittently, using the same procedures as the original probe condition (e.g., no instruction).
    b. Conduct probe trials during instructional sessions by intermittently assessing stimuli assigned to the control set at the beginning or end of an instructional session.
10. When criterion level responding is reached, collect data in the final probe condition for all behavior sets/chains for at least 3 sessions or until data are stable.
11. Replicate with similar participants.

Internal Validity

Studies have adequate internal validity when all likely threats are controlled for, and experimental control is demonstrated when adequate internal validity is present and when there are differences in the rate of learning (e.g., slope, sessions to criterion) between the two interventions and no evidence of history or maturation effects (i.e., learning) for the control set. Despite their widespread use, some threats to internal validity are particularly likely to be problematic when these designs are used; common threats to internal validity are described below and shown in Table 11.5.

Like the ATD, the AATD is generally rather short in duration; thus, some threats to internal validity are less likely to occur when compared with longer-duration designs like multiple baseline designs *if a control set is included and measured intermittently*. These include history, maturation, instrumentation, testing, and attrition; nevertheless, typical procedures for detecting and controlling for these threats should be used. Of course, these threats can occur in any study; the relatively short duration and measurement of the control set just decreases the likelihood. Collecting intermittent data on the control behavior set/chain during the comparison condition increases your opportunities to detect maturation or history effects (e.g., if the control data are

increasing, it is likely that the participant has learned the behaviors in non-study contexts, which increases the likelihood he or she is also learning the experimental stimuli in those contexts rather than or in addition to the study sessions). Since no intervention is applied to the control behavior set/chain, intermittent data collection can be done during probe sessions separate from the instructional sessions. If a control condition is not included in the design, maturation and history threats cannot be detected.

The first threat that is of considerable concern due to the nature of the design is procedural fidelity. As described above for the ATD, fidelity may be difficult to maintain due to the rapid iterative alternation between conditions. As with the ATD, you can prevent infidelity by conducting training to a strict criterion (including training during all condition types). You can detect infidelity by conducting frequent fidelity checks in all conditions, and re-training as needed based on formative analysis of fidelity data. You may also need to provide implementers with "cheat sheets" to remind them of the critical features of the intervention to be implemented for each session.

Multitreatment interference is possible in AATD studies. The effects of multitreatment interference can be minimized by increasing the time between sessions of the different instructional strategies in the comparison condition. Alternating sessions by day (rather than within-day) is often adequate to minimize multitreatment interference, but it takes longer to complete the study. When sessions occur in the same day, at least 1 hour should occur between them to minimize multitreatment interference. If you are concerned about possible multitreatment interference, the time between the sessions can be lengthened (e.g., 3 hours). Maintenance of learned behaviors in the final condition (intermixed probes) also provides evidence that multitreatment interference was not instrumental in results.

The AATD solves the separation of treatments issue, because each strategy is applied to separate behavior sets/chains and is not compromised by the reversibility issue; nonreversible behaviors are selected and the independent variables are applied to separate behaviors sets. Those behaviors can move from low levels in the initial probe condition to criterion levels in the comparison condition without negatively influencing findings. Unlike demonstration designs (multiple baseline, multiple probe, A-B-A-B) and multitreatment and ATD comparative designs, the AATD presents an additional threat to internal validity: lack of equal difficulty of behavior sets/chains. If the interventions are applied to behavior sets/chains that are *not* of equal difficulty, the comparison is seriously confounded. Counterbalancing assignment (if possible) is helpful in detecting this threat. Randomly assigning sets prevents potential researcher bias (e.g., choosing to assign what is perceived as an "easier to learn" set to the preferred intervention).

Advantages

The primary advantage of the AATD is it allows you to compare independent variables for non-reversible behaviors. Further, unlike the multitreatment design and ATD, it

solves the separation of treatments issue and it is not confounded by the reversibility problem. Studies using an AATD can be completed in a relatively short time period. Given an adequate number of replications across participants, the AATD can provide useful information about the efficiency of one instructional strategy over another.

Limitations

A major limitation of the AATD is the requirement that behavior sets/chains must be of equal difficulty. Failure to establish equal difficulty will result in an unfair evaluation of the compared strategies and in spurious conclusions.

Conclusions

AATDs are appropriate and efficient designs for evaluating intervention comparisons for non-reversible behaviors; it is one of only a few designs that can do so. AATDs are used less frequently than other common designs (multiple baseline designs), but are relevant for practitioners, who are often interested in the most efficient ways to teach new skills. Although experimental control is demonstrated for single participants when AATDs are used, we recommend that multiple participants be recruited to improve external validity.

Applied Example 11–3: AATD

Cihak, D., Alberto, P. A., Taber-Doughty, T., & Gama, R. I. (2006). A comparison of static picture prompting and video prompting simulation strategies using group instructional procedures. *Focus on Autism and Other Developmental Disabilities*, *21*, 89–99.

Video and picture prompts were compared for teaching response chains to six 11-year-old boys with moderate intellectual disabilities (IQ = 38 to 51). An AATD was used to compare the effectiveness and efficiency of picture and video prompting. Group instruction occurred in each participant's special education class and data were collected in community-based instruction (CBI). Some participants attended a middle school in the Southeast and others attended a middle school in the Midwest. The inclusion criteria were: (a) age between 11 and 15, (b) cognitive abilities in the moderate intellectual disability range, (c) attended middle school, (d) no sensory deficits, (e) no prior training on target tasks, (f) participation in CBI, (g) parent permission, and (h) verbal consent. Data were collected during three conditions, (a) initial probe sessions, (b) comparison condition, and (c) one follow-up probe session.

Two response chains, withdrawing $20 from the automated teller machine (ATM) and purchasing two items with a debit card, were counterbalanced across prompting procedures. The chains were deemed equivalent based on each task analysis requiring 12 steps with similar motor responses and equally difficult based on initial baseline group performance. Initial probe session data (5 sessions) were collected for both tasks during CBI. A task direction was delivered and participants had 15 seconds to initiate the chain. A single opportunity probe method was used—if the participant did not initiate the behavior within 15 seconds, he was asked if he was finished. If the participant responded "yes" or did not respond within 1 minute, probe sessions were discontinued.

Two group instructional sessions occurred each day, one session per prompting strategy, and order within the day was randomized. For the picture prompts, pictures of each task analysis step were taken with a digital camera and copied onto a transparency. During instruction the transparencies were displayed on a screen in front of the group for 4 seconds. For the video prompts strategy, a 4 second video clip of each task analysis step was shown. CBI data collection on target behaviors was scheduled 90 minutes following instruction. One trial on each chain was conducted during each CBI session, with 15 minutes between trials. A system of least prompts strategy was used with the following hierarchy: (a) verbal prompt, (b) gesture prompt, and (c) gesture plus verbal prompt. Data were collected using event recording on the number of steps completed independently and the prompt level

needed to complete each step. Mastery was set at 100% independent responding for 3 consecutive sessions. Two weeks after each group met criterion a follow-up probe session was conducted.

Interobserver and procedural reliability data were collected simultaneously during 25% of sessions across both conditions. The point-by-point method was used to calculate IOA, and it ranged from 95–100% (Mean = 98%). Procedural fidelity data were calculated by dividing the number of observed teacher behaviors by the number of planned teacher behaviors and multiplying by 100. Procedural fidelity ranged from 96–100% (Mean = 99%).

Results for Group 1 are presented in Figure 11.9. All participants had low stable performance in the initial probe condition for both chains; all showed immediate increases in level on introduction of both interventions. Results were similar for the other three participants—acquisition occurred across behaviors and participants for both prompt types. Efficiency data for Groups 1 and 2 are shown in Table 11.7. Little difference existed in the number of sessions to criterion for the two prompting strategies, although Edgar had twice as many sessions in the video prompting instruction. In terms of number of errors to criterion, four boys had similar numbers across both chains, but Carlos and Drew had more in the video prompt condition than in the picture prompt condition.

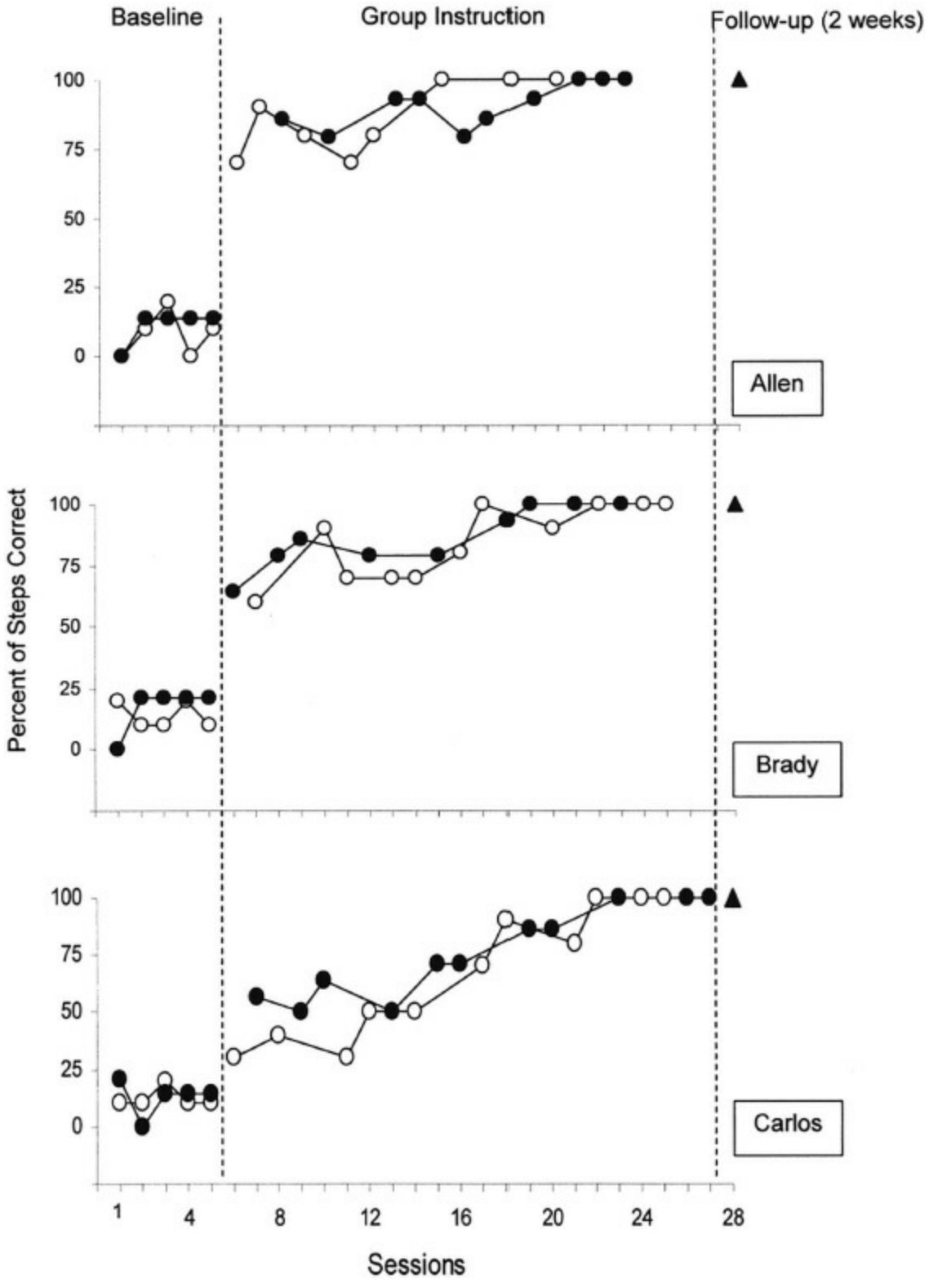


Figure 11.9 Adapted alternating treatments design.

Source: Cihak, D., Alberto, P. A., Taber-Doughty, T., & Gama, R. I. (2006). A comparison of static picture prompting and video prompting simulation strategies using group instructional procedures. *Focus on Autism and Other Developmental Disabilities, 21*, 89–99.

Table 11.7 Students' mean performance, number of errors, and number of sessions to criterion using static picture

prompts and video prompts across baseline and intervention phases.

Students' Mean Performance, Number of Errors, and Number of Sessions to Criterion Using Static Picture Prompts and Video Prompts Across Baseline and Intervention Phases

Student	Video prompting				Picture prompting			
	Baseline (%)	Instruction (%)	Errors	Sessions	Baseline (%)	Instruction (%)	Errors	Sessions
Group 1 (mean)	13.3	79.4	21.3	10	11.2	84.5	18.3	10
Allen	10.5	86.3	11	8	7.5	90.9	11	10
Brady	12.6	84.5	12	11	12.0	86.7	12	9
Carlos	16.3	67.3	41	11	14.0	75.9[a]	32	11
Group 2 (mean)	11.5	78.9	30.7	11	22.2	77.9	25	6.7
Drew	7.2	63.4	66	16	19.8	70.9[a]	52	15
Edgar	5.6	88.6	13	10	252	81.0	12	5
Frank	21.6	84.7	13	7	21.6	81.8	11	5
Overall (mean)	12.4	77.0	26.0	10.5	16.7	81.2	21.7	9.2

[a]*Wilcoxon signed-ranks matched-pairs Wtest, $p < .05$.*

From: Cihak, D., Alberto, P. A., Taber—Doughty, T., & Gama, R. I. (2006). A comparison of static picture prompting and video prompting simulation strategies using group instructional procedures. *Focus on Autism and Other Developmental Disabilities, 21*, 89–99.

Parallel Treatments Design (PTD)

Like the AATD, the **parallel treatments design (PTD)** was devised to compare instructional practices with *non-reversible behaviors* (Gast & Wolery, 1988). It can be conceptualized as two concurrently operating multiple probe designs—one instructional strategy is evaluated with one multiple probe design, and the second is evaluated with another multiple probe design—because they are concurrently-operating, you can also compare the two strategies. You can also conceptualize the designs as three time-lagged AATDs. The PTD is useful when comparing interventions for teaching functional, developmental, and academic behaviors. As with the AATD the instructional strategies being compared with the PTD are applied to separate behavior sets/chains. Target behaviors must meet the same criteria as with the AATD. Also, as with the multiple probe design, a strong *a priori* assumption should exist that the behaviors will not change until instruction occurs. Procedures for determining whether behaviors are of equal difficulty are identical to those discussed for the AATD. When the same behaviors are taught to more than one participant in a study, then you should counterbalance sets/chains across instructional strategies. With the PTD you should identify three or more behavior sets/chains for each instructional strategy being compared. Usually only two instructional strategies are compared, which means six total behavior sets/chains need to be identified. Six behavior sets/chains is a minimum, but eight sets/chains are recommended to increase intra-participant replications. Although all sets/chains should be equally difficult, at a minimum *pairs of behavior sets/chains must be equated* (e.g., sets 1 and 2 [tier 1] are equal but may be more or less difficult than sets 3 and 4 [tier 2]).

The time-lagged nature of the PTD makes a control condition unnecessary; the time-lagged introduction is sufficient for ruling out threats due to maturation or history.

When six sets/chains are used with two independent variables, the PTD has seven sequentially implemented experimental conditions, which is identical to the multiple probe (conditions) variation. However, instead of conducting probes including 3–4 behavior sets (corresponding to tiers of intervention), you include 6–8 behavior sets. The PTD uses two rules to order the experimental conditions: (a) rapid iterative implementation of the instructional procedures (independent variables) across sessions, and (b) time-lagged application of the instructional procedures across pairs of sets/chains. Figure 11.10 shows a hypothetical PTD study in the left panel; the right panel highlights the time-lagged AATD components. Figure 11.11 shows how that study also represents two multiple probe designs. We *do not* recommend presenting data similar to that shown in Figure 11.11 since it makes visual analysis of differences between conditions more difficult, although some authors have published using this type of display (Leaf et al., 2012). Presenting data in this way does allow for clear representation of data regarding the *demonstration* question (i.e., were there changes in baseline and intervention conditions); these are generally not the primary focus when PTDs are used. Analyzing data from a study using a PTD requires using visual analysis rules for an AATD for comparing the two interventions and visual analysis rules for MP designs for analyzing each intervention in comparison to baseline. Table 11.8 lists published studies using PTDs.

Procedural Guidelines

When using a PTD, adhere to the following guidelines:

1. Identify and define at least six nonreversible behavior sets/chains that are independent and functionally similar.
2. Select a sensitive, reliable, valid, and feasible data collection system and pilot the system and your behavior definitions.
3. Determine a priori frequency of reliability and fidelity data collection (e.g., 33% of sessions), and conduct data collection for the duration of the study.
4. Randomly assign one behavior set/chain to each instructional strategy and to a tier.
5. Determine what rules you will use for ordering conditions—it is critical to have an equal number of sessions that are roughly evenly spaced—thus, authors typically randomly select one condition and then automatically conduct the other condition for the next session.
6. Conduct an initial probe condition by collecting data on all behavior sets/chains for a minimum of 3 sessions or until data are stable for all behavior sets/chains.
7. Implement the comparison condition in the first tier by applying each strategy to its respective behavior set/chain in alternating sessions. If multiple sessions are

conducted each day, use counterbalancing to detect effects of time of day.

8. After both sets reach criterion, implement a second probe condition, collecting data on all behavior sets/chains for a minimum of 3 sessions or until all data are stable.
9. Implement the comparison condition in the second tier.
10. Repeat steps 8 and 9 for remaining tiers.
11. Replicate with similar participants.

Internal Validity

Studies have adequate internal validity when all likely threats are controlled for, and experimental control is demonstrated when adequate internal validity is present and when there are differences in the rate of learning (e.g., slope, sessions to criterion) between the two interventions and when behaviors change *when and only when* interventions are implemented (e.g., no behavior change in pre-intervention probe conditions).

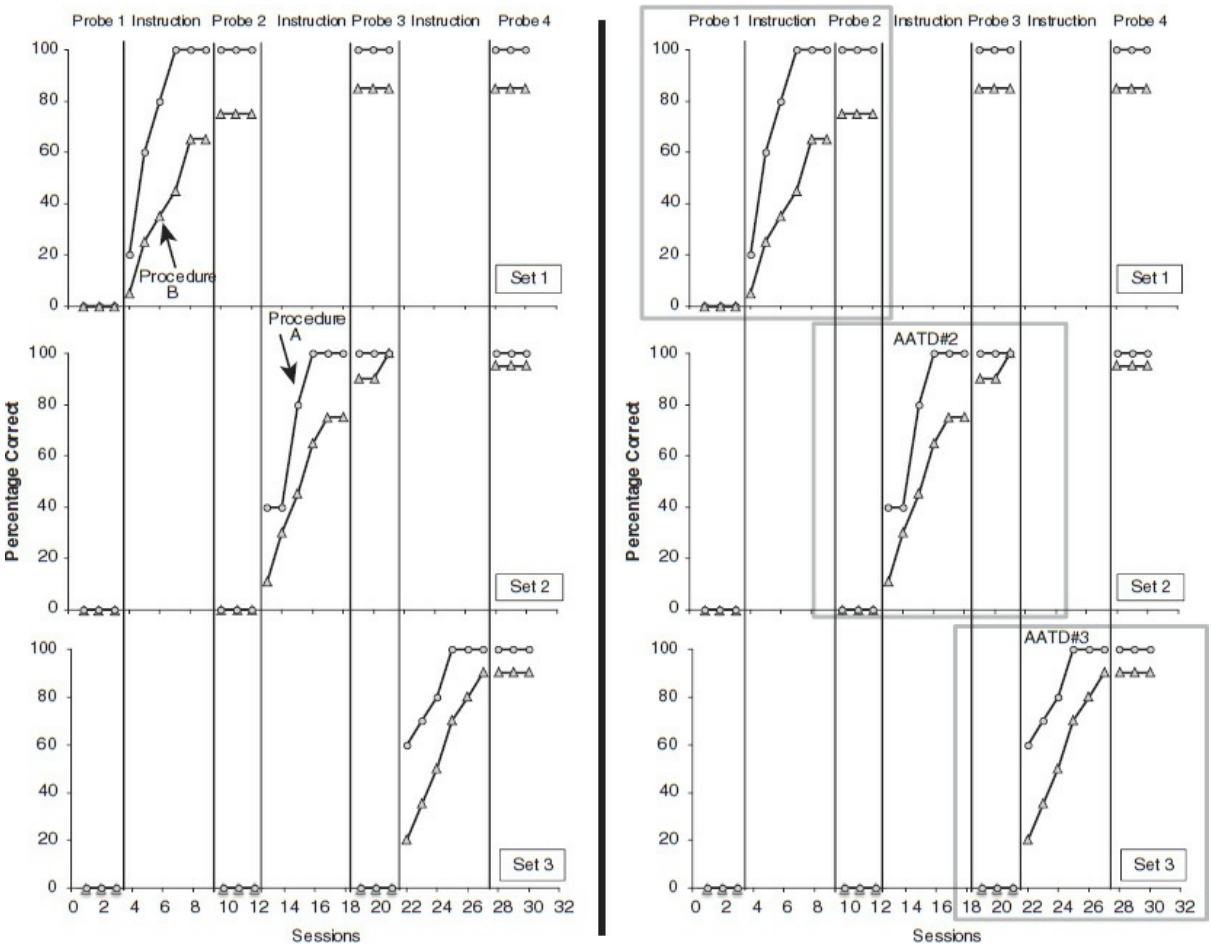


Figure 11.10 Typical presentation for parallel treatments design (left) and the same design with AATD comparisons highlighted (right).

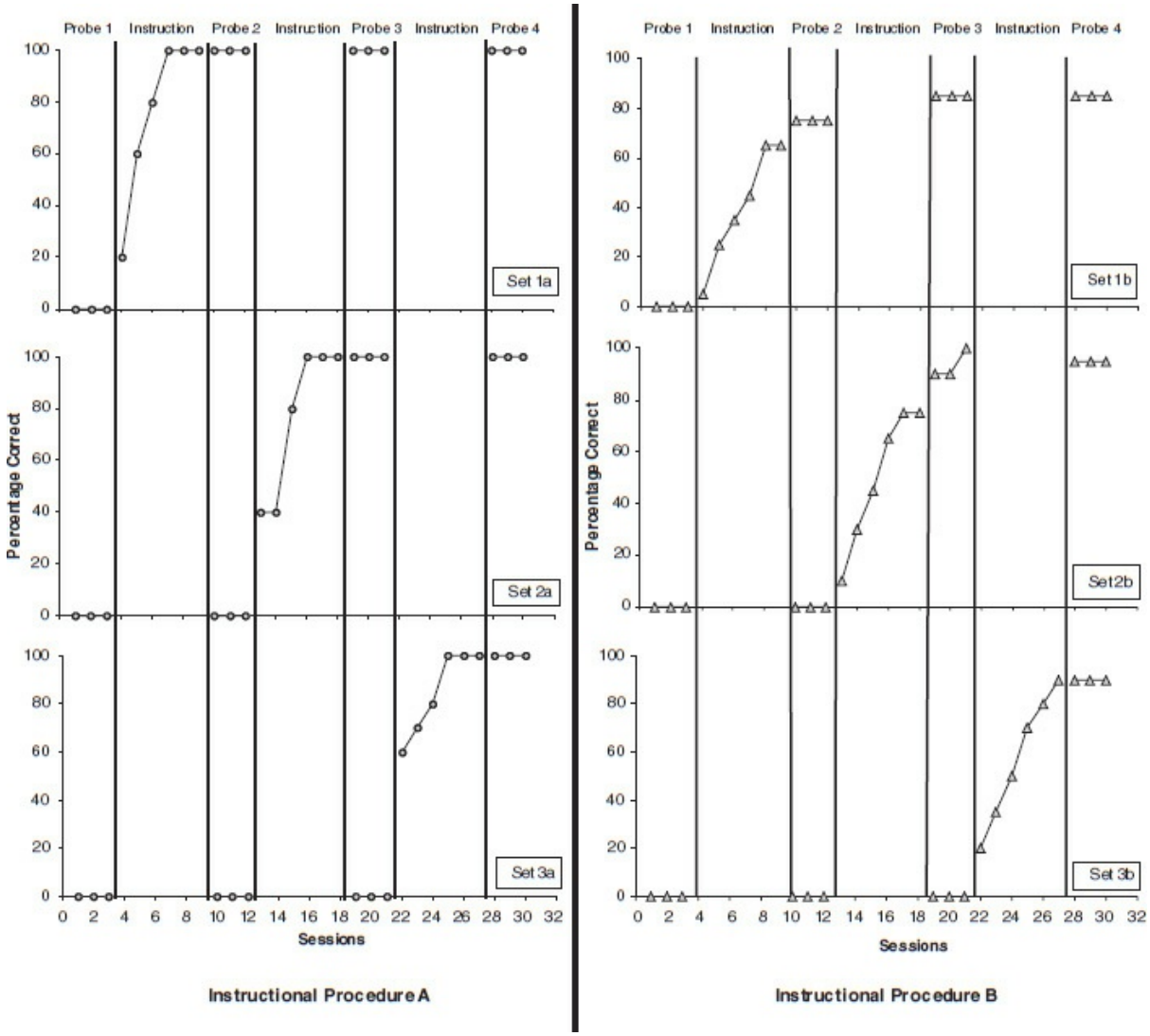


Figure 11.11 Parallel treatments design shown as two separate multiple probe designs. Note this is the same design as the one depicted in Figure 11.10 but this presentation makes it difficult to compare efficiency and effectiveness.

Table 11.8 *Studies Using Parallel Treatments Design (PTD)*

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
Leaf, J. B., Oppenheim-Leaf, M. L., Call, N. A., Sheldon, J. B., & Sherman, J. A. (2012). Comparing the teaching interaction procedure to social stories for people with autism. *Journal of Applied Behavior Analysis, 45*, 281–298.	Number: 6 Sex: M Age Range: 5–13 Disability/ diagnosis: ASD	Setting: Clinic, homes Arrangement: 1:1	Two alternating conditions (teaching interaction procedure and social stories)	Percent correct
Murzynski, N. T., & Bourrett, J. C. (2007). Combining video modeling and least-to-most prompting for establishing response chains. *Behavioral Interventions, 22*, 147–152.	Number: 2 Sex: M Age Range: 8–9 Disability/diagnosis: Autism	Setting: Participant's residential homes Arrangement: 1:1	Two alternating conditions (video modeling plus least-to-most prompting and least-to-most alone)	Number of steps completed independently
Rohena, E. I., Jitendra, A. K., & Browder, D. M. (2002). Comparison of the effects of Spanish and English constant time delay instruction on sight word reading by Hispanic learners with mental retardation. *Journal of Special Education, 36*, 169–184.	Number: 4 Sex: 2 M, 2 F Age Range: 12–15 Disability/diagnosis: Moderate ID	Setting: Self-contained classroom Arrangement: 1:1	Two alternating conditions (time delay presented in English or Spanish)	Percent correct
Schlosser, R. W., Belfiore, P. J., Nigam, R., Blischak, D., & Hetzroni, O. (1995). The effects of speech output technology in the learning of graphic symbols. *Journal of Applied Behavior Analysis, 28*, 537–549.	Number: 3 Sex: 2 M, 1 F Age Range: 24–25 Disability/ diagnosis: Severe to profound ID	Setting: Community-based workshop Arrangement: 1:1	Two alternating conditions (voice output communication device available and unavailable)	Percent correct
West, E. A., & Billingsley, F. (2005). Improving the system of least prompts: A comparison of procedural variations. *Education and Training in Developmental Disabilities, 40*, 131–144.	Number: 4 Sex: 3 M, 1 F Age Range: 5–6 Disability/diagnosis: Autism	Setting: Children's early childhood inclusive classroom. Arrangement: 1:1	Two alternating conditions (traditional and revised least to most procedures)	Percent correct
Wolery, M., Ault, M. J., Gast, D. L., Doyle, P. M., & Griffen, A. K. (1990). Comparison of constant time delay and the system of least prompts in teaching chained tasks. *Education and Training in Developmental Disabilities, 25*, 243–257.	Number: 4 Sex: 2 M, 2 F Age Range: 10–14 Disability/diagnosis: Moderate ID, Ds, AD, microcephaly	Setting: Self-contained classroom Arrangement: 1:1	Two alternating conditions (constant time delay and system of least prompts)	Percent correct

Note: ASD=autism spectrum disorder, ID=intellectual disability, Ds=Down syndrome, AD=articulation disorder

Some threats to internal validity (instrumentation, infidelity, attrition) are assessed in

the same way described for other designs; typical procedures for detecting and controlling for these threats should be used. History and maturation are controlled if behavior changes when and only when the behaviors are introduced to an intervention condition (the time-lagged component of the design). Testing may be particularly likely in this design because there are a large number of stimuli (e.g., 6–8 sets for 3–4 tier designs) and many testing occasions (e.g., 4 probe conditions). Some authors have used the multiple probe (days) variation to design PTD studies, but although this might allow you to conduct shorter, more frequent probes, it does not decrease the number of assessments required (e.g., total number of probe trials). You can minimize testing threats by designing non-aversive probe conditions and dividing probe sessions into multiple measurement opportunities.

Multitreatment interference is possible with the PTD. As with the AATD, it can be minimized by increasing the amount of time between sessions during instructional comparison conditions. Although alternating sessions by day is possible, the length of PTD studies usually calls for having multiple sessions per day. Baseline and maintenance probe conditions are controls for this threat. PTDs do not have the separation of treatments issue, because each strategy is applied to separate sets/chains. The PTD is not compromised by the reversibility issue; nonreversible behaviors are selected and independent variables are applied to separate (but equally difficult) sets/chains. As with the AATD, ensuring equal difficulty of behavior sets/chains is an extremely important issue.

Advantages

The primary advantage of the PTD is ability to compare two interventions for teaching non- reversible behaviors while controlling for history and maturation threats by time-lagging the comparisons. Another advantage is that repeated probe conditions of previously taught sets/chains allow study of the relative maintenance of behaviors assigned to each strategy. Two strategies may be effective and equally efficient, but may result in greater maintenance. Finally, unlike the AATD, there is no need for a control set, which may be seen as beneficial from a practical standpoint; there is no "untaught" behavior set, which aligns more closely with practice.

Limitations

Three major limitations exist with the PTD. First, identifying six equally difficult behavior sets/chains is challenging. Although procedures exist for determining whether sets/chains are equally difficult (Romer et al., 1988), it is often time consuming to find a sufficient number of behaviors. Second, a great deal of time is spent conducting probe sessions. Third, the design requires a great deal of time to complete; thus, the PTD should only be used when adequate time and availability of participants exists. For

example, the AATD traditionally has three conditions (initial probe, instructional comparison, and final probe), and these are similar to Probe 1, the first instructional comparison, and Probe 2 of the PTD. The PTD provides more intra-participant replications than the AATD when studying effectiveness, efficiency, and maintenance; but if time does not allow, the AATD should be used.

Conclusions

The PTD is a methodologically rigorous design that uses two combined condition-ordering strategies (time lagged implementation and rapid iterative alternation) to compare two intervention strategies *and* control for history and maturation by time-lagging comparisons across multiple tiers. It is a powerful demonstration of effectiveness as well as comparison, but it requires a considerable amount of time. Although experimental control is demonstrated for single participants when PTDs are used, we recommend that multiple participants be recruited to improve external validity.

Summary

In this chapter four experimental designs (multitreatment design, ATD, AATD, and PTD) were discussed; in Table 11.9, the four designs are compared on various dimensions. All of the designs can be used to compare interventions. The multitreatment and ATD require reversible behaviors, and the AATD and PTD require non-reversible behaviors. With the multitreatment design and ATD, only one behavior is required and it can be an acceleration or deceleration target. With the AATD, three behavior sets/chains are required, and with the PTD at least six behavior sets/chains are required. Generalization of the interventions can be evaluated in all designs, but often is difficult with the ATD. Only the PTD has built-in assessments of maintenance, although follow-up assessments can be done with all four designs. Multitreatment interference is likely with all of these designs, and methods for detecting and controlling it when detected is specific to each design.

Table 11.9 *Similarities and Differences of the Multitreatment Design, ATD, AATD, and PTD*

Dimension	Multitreatment	ATD	AATD	PTD
Types of comparison questions for which design is useful	Compare IVs Component analyses Parametric analyses	Compare IVs Component analyses Parametric analyses Assess factors maintaining behavior	Compare instructional strategies Efficiency questions Parametric analyses Procedural fidelity	Compare instructional strategies Efficiency questions Parametric analyses Maintenance questions
Reversible or Nonreversible behavior?	Reversible	Reversible	Nonreversible	Nonreversible
Minimum number of behaviors required	One	One	3 behavior sets/ chains	6 behavior sets/ chains
Type of behavior	Acceleration or deceleration	Acceleration or deceleration	Acceleration only	Acceleration only
Generalization measures feasible?	Yes	Yes	Yes	Yes
Maintenance part of design?	No	No	No	Yes
Minimum number of participants	1	1	1	1
Is reversibility issue solved?	No	No	Yes	Yes
Is separation of treatments issue solved?	No	No	Yes	Yes
Special considerations	None	Participant must discriminate when each condition is in effect	Behaviors must be of equal difficulty	Behaviors must be of equal difficulty

Note: ATD=alternating treatments design; AATD=adapted alternating treatments design; PTD=parallel treatments design

Applied Example 11–4: PTD (Probe Days)

Jones, C. D., & Schwartz, I. S. (2004). Siblings, peers, and adults: Differential effects of models for children with autism. *Topics in Early Childhood Special Education*, *24*, 187–198.

Jones and Schwartz (2004) compared three types of models (peers, siblings, and adults) on the acquisition of language skills by three preschoolers with autism. Models were siblings who attended the same school, peers without disabilities from the child's class, and an adult from the child's class. Two participants (Erin & Jerry) were taught three sets of behaviors (actions, professions, and opposites); and one participant (Jennifer) was taught three sets of two types of skills (actions and professions). Each set contained three behaviors (e.g., three different actions). For each participant, one set of each type of language skills was assigned a different model (peer, sibling, adult).

A PTD was used involving intermittent single probe days after three consecutive days of stable baseline performance, rather than probe conditions. The order of conditions was: (a) pre-baseline in which stimuli were identified and selected, (b) baseline, (c) instructional comparison, and (d) maintenance probes. During baseline sessions, the investigator conducted individual assessments of all behavior sets. For each trial, the researcher showed a picture and asked a question ("What is this person doing?"; "Who is this person?"; "If this is [e.g., *open*], then this is _____.") The models were taught to answer the questions correctly prior to the study.

During the comparison, 15-minute sessions were comprised of three 5-minute segments. In each segment, a different set of the same type of language behaviors (e.g., actions) were taught, and the corresponding model was present. In each segment, three trials were delivered on each of the three behaviors in the set. First, the investigator showed the model a picture, delivered the task direction, and provided a response interval. After the model responded correctly, the investigator showed the participant the picture, delivered the task direction, and provided a response interval. Correct responses were praised and errors and no responses were ignored. Then, investigator, a different model, and participant completed another 5-minute segment with another set of behaviors. When this segment was completed, the process was repeated with the third model. The daily order of models was the same for each tier, but was randomly determined across tiers. IOA data were collected for 21% of the sessions with at least one check per condition for each child. Mean IOA was 97% (range = 95%–100%). Maintenance data were collected intermittently after criterion level performance was demonstrated.

In Figure 11.12, Erin's data are shown for each model type across three tiers. In initial baseline sessions, Erin did not answer correctly for any trial. When the

comparison condition was implemented for actions, her number of correct responses immediately increased from 0 to 3 correct responses on peer modeled set, 7 on adult modeled set, and 8 on sibling modeled set. Correct responding reached 100% on sibling and adult sets and 66% on the peer set. After criterion level responding on action behaviors, probe data were collected across sets. Untrained sets remained at 0 correct responses. The comparison condition was implemented for the next set. Performance replicated that of the action behaviors sets; the number of correct responses increased on behaviors modeled by her peer, sibling, and adult. When criterion level responding occurred, probe sessions were conducted on the final behavior set; performance remained at zero. The comparison was implemented for this set and behavior change was similar to previous comparisons. Correct responding reached 100% on peer and adult sets and 66% on sibling set.

Maintenance probes were conducted on each behavior set after the respective comparison conditions. Performance maintained across all sets, regardless of the model type. Similar findings were reported with other participants. The researchers concluded all three models were effective in increasing the number of correct responses, and child models were as effective as adult models.

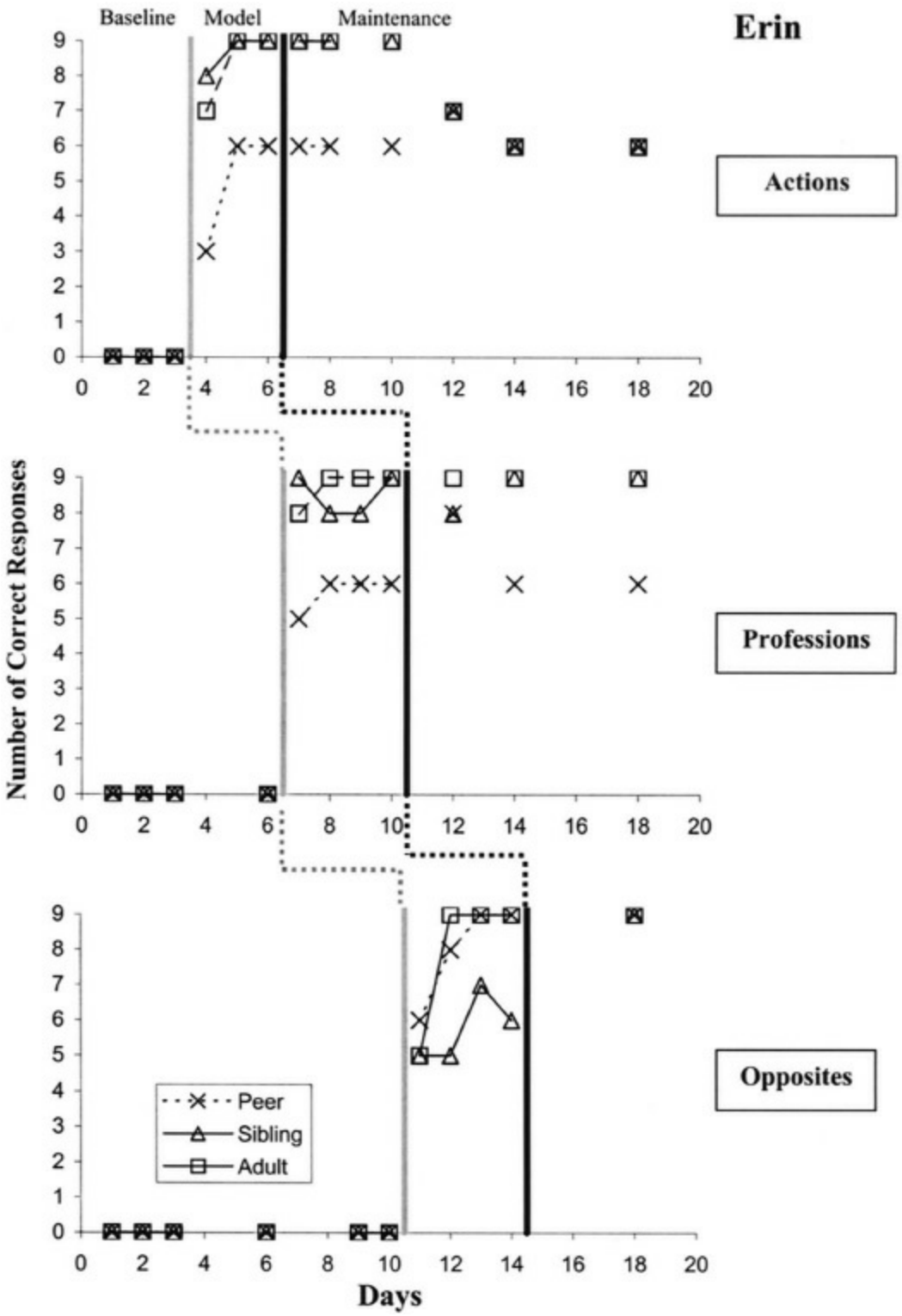


Figure 11.12 Parallel treatments design.

Source: Jones, C. D., & Schwartz, I. S. (2004). Siblings, peers, and adults: Differential effects of models for children with autism. *Topics in Early Childhood Special Education*, *24*, 187–198.

Appendix 11.1

Visual Analysis for Multitreatment Designs

Adequate design	**Examples**: B-C-B-C, A-B-C-B-C, A-B-C-D-C-D **Non-Examples**: A-B-A-C, A-B-C-B
Visual analysis considerations specific to design	• **Consistency across counterbalanced designs**. When multiple participants are included, the ability to draw definitive conclusions about the superior intervention is enhanced when the order is counterbalanced (e.g., B-C-B-C for one participant, and C-B-C-B for the other participant).
Common and potentially problematic data patterns	• **Delayed change across conditions**. A delayed change is less problematic if (a) you continue conditions until data are stable, (b) a delay was predicted a priori, (c) the delay occurs in both intervention conditions, and (d) the latency and magnitude of the delay are consistent. • **Small magnitude changes**. Small magnitude changes are not problematic if data patterns are consistent for similar conditions (e.g., behavior changes were small for both B_1-C_1 and B_2-C_2) and if between-condition level change exceeds within-condition variability (e.g., no overlap is present). Small magnitude changes are potentially problematic if agreement data are discrepant (e.g., data from a second observer might suggest no change occurred; assessed via visual analysis of plotted data from both observers). • **Highly variable data in one or more condition**. Variable data are less problematic if changes in level are above and beyond variability (e.g., no overlap), or if changes in variability predictably change across conditions (e.g., high variability in Intervention A followed by low variability during Intervention B). Variability is problematic if there is a high percentage of overlapping data points or variability otherwise precludes making a decision regarding behavior change.
Convincing Functional Relation (*B-C-B-C design)*	• Behavior patterns in B_1 and B_2 are similar • Behavior patterns in C_1 and C_2 are similar • Changes from B_1-C_1 and B_2-C_2 are similarly therapeutic • Changes from C_1-B_2 are contra-therapeutic • All changes are abrupt and concurrent with condition changes • Overlap is minimal • Variability and trends in any condition do not preclude ability to identify condition changes

Visual Analysis for Alternating Treatments Design

Adequate design	**Examples**: At least two alternating conditions with five data points each (with or without baseline and best-alone conditions) and a planned procedure for sequencing conditions (preferably based on randomization). **Non-Examples**: Alternating conditions with fewer than five data points each.
Visual analysis considerations specific to design	• **Differentiation**. In general, describe data patterns in ATDs by describing the degree of differentiation between two data paths rather than individual changes between data points. If three or more data paths are present, describe pairwise comparisons (e.g., A vs. B, B vs. C, A vs. C). Differentiation is defined as a consistent difference in level between adjacent data points from different conditions (e.g., data points 1, 3, 5, and 7 in Condition A are higher than corresponding points 2, 4, 6, and 8 in Condition B, respectively).
Common and potentially problematic data patterns	• **Small magnitude differentiation**. Small differences between conditions are not problematic if the differences are consistent over time. Small magnitude changes are potentially problematic if agreement data are discrepant (e.g., data from a second observer might suggest no change occurred; assessed via visual analysis of plotted data from both observers). • **Highly variable data in one or more condition**. Variable data are less problematic if differentiation is still present (e.g., no overlap). Variability is problematic if there is a high percentage of overlapping data points or variability otherwise precludes making a decision regarding differentiation. • **Trends during the comparison condition**. Unlike in other designs, trends are not particularly problematic in ATDs, even if they occur during baseline conditions, so long as differentiation is still present. For example, even if challenging behavior was decreasing over time in all three conditions (two treatments and a continuing baseline), if data were consistently lower in level in one treatment, a functional relation can be established.
Convincing Functional Relation	• Data paths with at least 5 points each do not overlap. • Variability and trends in any condition do not preclude ability to identify differentiation.

Visual Analysis for Adapted Alternating Treatments Design

Adequate design	**Examples**: At least two alternating conditions and a control condition with five data points each, following a baseline (probe) condition. **Non-examples**: Conditions with fewer than five data points or not including a control condition.
Visual analysis considerations specific to design	• **Differentiation**. Like the ATD, data patterns in AATDs are often described in terms of differentiation. Because non-reversible behaviors are measured, this differentiation is often in terms of the steepness of slopes (e.g., acquisition rate, time to acquire the behaviors). • **Control set comparisons**. Because AATDs are generally conducted with two treatments previously shown (in demonstration studies) to be effective, both interventions are likely to result in behavior change. Thus, the use of a control set is critical to rule out history and maturation threats; data for this set should be collected throughout the course of the study and plotted alongside data from the treatment sets. If behaviors assigned to the control condition improve over time, these threats are likely and experimental control is weakened.
Potentially problematic data patterns	• **Lack of differentiation between treatments**. When two treatments are equally effective (e.g., have the same slope), experimental control is weakened if no control set is present or if behavior change occurs for stimuli assigned to the control set.
Convincing Functional Relation	• One or both interventions result in behavior change • Control data (measured throughout the intervention) do not change over time

References

Addison, L. R., Piazza, C. C., Patel, M. R., Bachmeyer, M. H., Rivas, K. M., Milnes, S. M., & Oddo, J. (2012). Comparison of sensory integrative and behavioral therapies as treatment for pediatric feeding disorders. *Journal of Applied Behavior Analysis*, *45*, 455–471.

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

Barton, E. E., Stiff, L., & Ledford, J. R. (2017). The effects of contingent reinforcement on peer imitation in a small group play context. *Journal of Early Intervention.*

Billingsley, F. F., White, O. R., & Munson, A. R. (1980). Procedural reliability: A rationale and an example. *Behavioral Assessment*, *2*, 229–241.

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

Birnbrauer, J. S., Peterson, C. R., & Solnick, J. V. (1974). Design and interpretation of studies of single subjects. *American Journal of Mental Deficiency*, *79*, 191–203.

Chazin, K. T., Ledford, J. R., Barton, E. E., & Osborne, K. O. (2017). The effects of antecedent exercise on engagement during large group activities for young children. *Remedial and Special Education.* doi: 10.1177/0741932517716899

Cihak, D., Alberto, P. A., Taber-Doughty, T., & Gama, R. I. (2006). A comparison of static picture prompting and video prompting simulation strategies using group instructional procedures. *Focus on Autism and Other Developmental Disabilities*, *21*, 89–99.

Cox, A. L., Gast, D. L., Luscre, D., & Ayres, K. M. (2009). The effects of weighted vests on appropriate in-seat behaviors of elementary-aged students with autism and severe to profound intellectual disabilities. *Focus on Autism and Other Developmental Disabilities*, *24*, 17–26.

Douglas, K. H., Ayres, K. M., & Langone, J. (2015). Comparing self-management strategies delivered via an iPhone to promote grocery shopping and literacy. *Education and Training in Autism and Developmental Disabilities*, *50*, 446–465.

Dunlap, G. (1984). The influence of task variation and maintenance tasks on the learning and affect of autistic children. *Journal of Experimental Child Psychology*, *37*, 41–64.

Dunlap, G., Strain, P. S., Fox, L., Carta, J. J., Conroy, M., Smith, B. J., et al. (2006). Prevention and intervention with young children's challenging behavior: Perspective regarding current knowledge. *Behavioral Disorders*, *32*, 29–45.

Edgar, E. B., & Billingsley, F. F. (1974). Believability when N=1. *The Psychological Record*, *24*, 147–160.

Fiske, K. E. (2008). Treatment integrity of school-based behavior analytic interventions: A review of research. *Behavior Analysis and Practice*, *1*, 2,19–25.

Freeman, K. A., & Dexter-Mazza, E. T. (2004). Using self-monitoring with an adolescent with disruptive classroom behavior. *Behavior Modification*, *28*, 402–419.

Gast, D. L., & Wolery, M. (1988). Parallel treatments design: A nested single subject design for comparing instructional procedures. *Education and Treatment of Children*, *11*, 270–285.

Groskreutz, N. C., Groskreutz, M. P., & Higbee, T. S. (2011). The effects of varied levels of treatment integrity on appropriate toy manipulation in children with autism. *Research in Autism Spectrum Disorders*, *5*, 1358–1369.

Hains, A. H., & Baer, D. M. (1989). Interaction effects in multi-element designs: Inevitable, desirable, and ignorable. *Journal of Applied Behavior Analysis*, *22*, 57–69.

Hammond, J. L., & Hall, S. S. (2011). Functional analysis and treatment of aggressive behavior following resection of a craniopharngioma. *Developmental Medicine and Child Neurology*, *53*, 369–374.

Hanley, G. P., Piazza, C. C., Fisher, W. W., & Maglieri, K. A. (2005). On the effectiveness of and preference for punishment and extinction components of function-based interventions. *Journal of Applied Behavior Analysis*, *38*, 51–65.

Haydon, T., Conroy, M. A., Scott, T. M., Sindelar, P. T., Barber, B. R., & Orlando, A. (2010). A comparison of three types of opportunities to respond on student academic and social behavior. *Journal of Emotional and Behavioral Disorders*, *18*, 27–40.

Herzinger, C. V., & Campbell, J. M. (2007). Comparing functional assessment methodologies: A quantitative synthesis. *Journal of Autism and Developmental Disorders*, *37*, 1430–1445.

Holcombe, A., Wolery, M., & Gast, D. L. (1994). Comparative single subject research: Description of designs and discussion of problems. *Topics in Early Childhood Special Education*, *14*, 119–145.

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.

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

Ingersoll, B. (2011). The differential effect of three naturalistic language interventions on language use in children with autism. *Journal of Positive Behavior Interventions*, *13*, 109–118.

Iwata, B. A., Dorsey, M. F., Slifer, K. J., Bauman, K. E., & Richman, G. S. (1994). Toward a functional analysis of self-injury. *Journal of Applied Behavior Analysis*, *27*, 197–209 (reprinted from *Analysis and Intervention in Developmental Disabilities*, *2*, 3–20).

Johnston, J. M. (1988). Strategic and tactical limits of comparison studies. *The Behavior Analyst*, *11*, 1–9.

Jones, C. D., & Schwartz, I. S. (2004). Siblings, peers, and adults: Differential effects of models for children with autism. *Topics in Early Childhood Special Education*, *24*, 187–198.

Kodak, T., Northup, J., & Kelley, M. E. (2007). An evaluation of the types of attention that maintain problem behavior. *Journal of Applied Behavior Analysis*, *40*, 167–171.

Leaf, J. B., Oppenheim-Leaf, M. L., Call, N. A., Sheldon, J. B., & Sherman, J. A. (2012). Comparing the teaching interaction procedure to social stories for people with autism. *Journal of Applied Behavior Analysis*, *45*, 281–298.

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.

Logan, K. R., Jacobs, H. A., Gast, D. L., Murray, A. S., Daino, K., & Skala, C. (1998). The impact of typical peers on the perceived happiness of students with profound multiple disabilities. *Journal of the Association for Persons With Severe Handicaps*, *23*, 309–318.

Lynch, A., Theodore, L. A., Bray, M. W., & Kehle, T. J. (2009). A comparison of group-oriented contingencies and randomized reinforcers to improve homework completion and accuracy for students with disabilities. *School Psychology Review*, *38*, 307–324.

McComas, J. J., Thompson, A., & Johnson, L. (2003). The effects of presession attention on problem behavior maintained by different reinforcers. *Journal of Applied Behavior Analysis*, *36*, 297–307.

McGonigle, J. J., Rojahn, J., Dixon, J., & Strain, P. S. (1987). Multiple treatment interference in the alternating treatments design as a function of the intercomponent interval length. *Journal of Applied Behavior Analysis*, *20*, 171–178.

Mechling, L. C., & Ayres, K. M. (2012). A comparative study: Completion of fine motor office related tasks by high school students with autism using visual models on large and small screen sizes. *Journal of Autism and Developmental Disorders*, *42*, 2364–2373.

Milo, J., Mace, F. C., & Nevin, J. A. (2010). The effects of constant versus varied reinforcers on preference and resistance to change. *Journal of the Experimental Analysis of Behavior*, *93*, 385–394.

Mueller, M. M., Sterling-Turner, H. E., & Moore, J. W. (2005). Towards developing a classroom-based functional analysis condition to assess escape-to-attention as a variable maintaining problem behavior. *School Psychology Review*, *34*, 425–431.

Murzynski, N. T., & Bourrett, J. C. (2007). Combining video modeling and least-to-most prompting for establishing response chains. *Behavioral Interventions*, *22*, 147–152.

Neef, N. A., Cihon, T., Kettering, T., Guld, A., Axe, J. B., Itoi, M., & DeBar, R. (2007). A comparison of study session formats on attendance and quiz performance in a college course. *Journal of Behavioral Education*, *16*, 235–249.

Neef, N. A., & Peterson, S. M. (2007). Functional behavior assessment. In J. O. Cooper, T. E Heron, & W. L. Heward (Eds.), *Applied behavior analysis* (2nd ed., pp. 500–524). Upper Saddle River, NJ: Pearson.

Reichow, B., Barton, E. E., Sewell, J. N., Good, L., & Wolery, M. (2010). The effects of weighted vests on the engagement of children with developmental delays and

autism. *Focus on Autism and Other Developmental Disabilities*, *25*, 3–11.

Reichow, B., & Wolery, M. (2009). Comparison of everyday and every-fourth-day probe sessions with the simultaneous prompting procedure. *Topics in Early Childhood Special Education*, *29*, 79–89.

Reinhartsen, D. R., Garfinkle, A. N., & Wolery, M. (2002). Engagement with toys in two-year-old children with autism: Teacher selection and child choice. *Journal of the Association for Persons With Severe Handicaps*, *27*, 175–187.

Rohena, E. I., Jitendra, A. K., & Browder, D. M. (2002). Comparison of the effects of Spanish and English constant time delay instruction on sight word reading by Hispanic learners with mental retardation. *Journal of Special Education*, *36*, 169–184.

Romer, L. T., Billingsley, F. F., & White, O. R. (1988). The behavior equivalence problem in within-subject treatment comparisons. *Research in Developmental Disabilities*, *9*, 305–315.

Sanetti, L. M. H., Luiselli, J. K., & Handler, M. W. (2007). Effects of verbal and graphic performance feedback on behavior support plan implementation in a public elementary school. *Behavior Modification*, *31*, 454–465.

Savaiano, M. E., Compton, D. L., Hatton, D. D., & Lloyd, B. P. (2016). Vocabulary word instruction for students who read braille. *Exceptional Children*, *82*, 337–353. Schlosser, R. W., Belfiore, P. J., Nigam, R., Blischak, D., & Hetzroni, O. (1995). The effects of speech output technology in the learning of graphic symbols. *Journal of Applied Behavior Analysis*, *28*, 537–549.

Simonsen, B., MacSuga, A., Fallon, L. M., & Sugai, G. (2013). The effects of self-monitoring on teachers' use of specific praise. *Journal of Positive Behavior Interventions*, *15*, 5–15.

Sindelar, P. T., Rosenberg, M. S., & Wilson, R. J. (1985). An adapted alternating treatments design for instructional research. *Education and Treatment of Children*, *8*, 67–76.

Singleton, D. K., Schuster, J. W., Morse, T. E., & Collins, B. C. (1999). A comparison of antecedent prompt and test and simultaneous prompting procedures in teaching grocery words to adolescents with mental retardation. *Education and Training in Developmental Disabilities*, *34*, 182–199.

State, T. M., & Kern, L. (2012). A comparison of video feedback and in vivo self-monitoring on the social interactions of an adolescent with Asperger syndrome. *Journal of Behavioral Education*, *21*, 18–33.

Torelli, J. N., Lloyd, B. P., Diekman, C. A., & Wehby, J. H. (2017). Teaching stimulus control via class-wide multiple schedules of reinforcement in public elementary school classrooms. *Journal of Positive Behavior Interventions*, *19*, 14–25.

Tostanoski, A., Lang, R., Raulston, T., Carnett, A., & Davis, T. (2014). Voices from the past: Comparing the rapid prompting method and facilitated communication. *Developmental Neurorehabilitation*, *17*, 219–223.

Travers, J. C., & Fefer, S. A. (2017). Effects of shared active surface technology on the communication and speech of two preschool children with disabilities. *Focus on*

Autism and Other Developmental Disabilities, *32*, 44–54.

Ulman, J. D., & Sulzer-Azaroff, B. (1975). Multi-element baseline design in educational research. In E. Ramp & G. Semb (Eds.), *Behavior analysis: Areas of research and application* (pp. 377–391). Englewood Cliffs, NJ: Prentice-Hall.

VanDerHeyden, A. M., Snyder, P., Smith, A., Sevin, B., & Longwell, J. (2005). Effects of complete learning trials on child engagement. *Topics in Early Childhood Special Education*, *25*, 81–94.

Viel-Ruma, K., Houchins, D., & Fredrick, L. (2007). Error self-correction and spelling: Improving the spelling accuracy of secondary students with disabilities in written expression. *Journal of Behavioral Education*, *16*, 291–301.

Vollmer, T., Sloman, K., & Pipkin, C. (2008). Practical implications of data reliability and treatment integrity monitoring. *Behavior Analysis in Practice*, *1*, 2, 4–11.

Werts, M. G., Wolery, M., Holcombe, A., & Frederick, C. (1993). Effects of instructive feedback related and unrelated to the target behaviors. *Exceptionality*, *4*, 81–95.

Werts, M. G., Wolery, M., Holcombe, A., & Gast, D. L. (1995). Instructive feedback: Review of parameters and effects. *Journal of Behavioral Education*, *5*, 55–75.

West, E. A., & Billingsley, F. (2005). Improving the system of least prompts: A comparison of procedural variations. *Education and Training in Developmental Disabilities*, *40*, 131–144.

Wolery, M., Ault, M. J., & Doyle, P. M. (1992). *Teaching students with moderate and severe disabilities: Use of response prompting strategies*. White Plains, NY: Longman.

Wolery, M., Ault, M. J., Gast, D. L., Doyle, P. M., & Griffen, A. K. (1990). Comparison of constant time delay and the system of least prompts in teaching chained tasks. *Education and Training in Developmental Disabilities*, *25*, 243–257.

Zimmerman, K. N., Ledford, J. R., & Severini, K. E. (2017). Brief report: The effects of a weighted blanket on engagement during circle time for a student with ASD. *Manuscript Under Review*.

12
Combination and Other Designs

Jennifer R. Ledford and David L. Gast

Important Terms

changing criterion design, simultaneous treatments design, repeated acquisition design, brief experimental design

In this chapter we elaborate on variations of the basic and widely-used research designs discussed in previous chapters, including changing criterion designs (Hartmann & Hall, 1976), simultaneous treatments designs (Barlow & Hayes, 1979), repeated acquisition designs (Kennedy, 2005), and brief experimental designs (e.g., Cooper, Wacker, Sasso, Reimers, & Donn, 1990). In addition, we provide examples of how researchers have combined single case designs (SCDs) to strengthen their evaluation of experimental control and to resolve some of the ambiguities that may arise during the course of an experiment. Each of the four stand-alone designs presented in this chapter are less-widely used, due to both the considerable challenges that exist in controlling threats to

internal validity in the context of these designs and to their relatively constrained utility (e.g., each design is appropriate for a few specific applications). Combination designs, while sometimes difficult to implement, may assist researchers in controlling threats to internal validity and answering multiple or complex questions.

Changing Criterion Designs

Sidman (1960) described a research design that Hall (1971) named the **changing criterion design**. This design may be appropriate for practitioners and applied researchers who wish to evaluate instructional or therapy programs that require gradual, stepwise changes in behavior. This design can be used to increase or decrease behaviors already in a participants' repertoire. Hartmann and Hall (1976) describe the changing criterion design as follows:

> The design requires initial baseline observations on a single target behavior. This baseline phase is followed by implementation of a treatment program in each of a series of treatment phases. Each treatment phase is associated with a stepwise change in criterion rate for the target behavior. Thus, each phase of the design provides a baseline for the following phase. When the rate of the target behavior changes with each stepwise change in the criterion, therapeutic change is replicated and experimental control is demonstrated.
>
> (p. 527)

Though the changing criterion design has not been widely cited in the applied research literature (e.g., only used in just over 100 articles; Klein, Houlihan, Vincent, & Panahon, 2017), Hartmann and Hall (1976) have suggested it may be useful to monitor a wide range of programs (e.g., systematically increasing correct homework completion, decreasing number of cigarettes smoked per day). Researchers interested in assessing intervention programs that employ differential reinforcement procedures may find the changing criterion design helpful. For example, if a child completed 20–25% of math problems assigned during an independent work period during baseline conditions, a changing criterion design could be used such that each subsequent condition required an increasing percentage of completion (e.g., criterion 1=30%, criterion 2=50%, criterion 3=80%, criterion 4=65%, criterion 5=100%) to receive reinforcement.

To demonstrate experimental control using the changing criterion design you must show that each time the criterion level is changed (increased or decreased), there is concomitant change in the dependent variable. This change should be immediate and should follow a stable level and trend in the data at the preceding criterion level; this close alignment with criterion levels is required to rule out maturation effects. In addition, these effects can be detected by including a withdrawal criterion at some point during the study. If the data move in a contra- therapeutic direction during the withdrawal criterion, maturation is unlikely. Although this reversal is imperative for demonstrating maturation is unlikely to be a threat, it has been done relatively rarely (e.g., less than 40% of published changing criterion studies; Klein et al., 2017). It is imperative to demonstrate stability before changing the criterion level, for each phase serves as a baseline measure for the subsequent phase within the intervention condition. Replication of effect is demonstrated if each stepwise change in criterion results in a behavior change to the new criterion level. Figure 12.1 illustrates, with hypothetical data, the use of a changing criterion design to evaluate the effectiveness of an intervention

designed to decrease number of talk-outs. Note the criterion changes from 8 to 6 to 5 to 3 and then back up to 5 (a withdrawal). Because data closely align with the withdrawal criterion, we can be confident maturation is not a likely threat (e.g., talk-outs were not gradually improving, independent of the intervention.

The changing criterion design has been employed to monitor desensitization (Koegel, Openden, & Koegel 2004; Ricciardi, Luiselli, & Camre, 2006), prompt fading (Luiselli, 2000; Flood & Wilder, 2004), fluency building (Nes, 2005), exercise (DeLuca & Holborn, 1992), and self- monitoring (Ganz & Sigafoos, 2005) programs, among others (see Klein et al., 2017, for a review). It can be used to evaluate behavior acceleration (e.g., rate of problem completion) and deceleration (e.g., smoking, talk-outs) behaviors; monitoring of acceleration behaviors is somewhat more common (Klein et al., 2017).

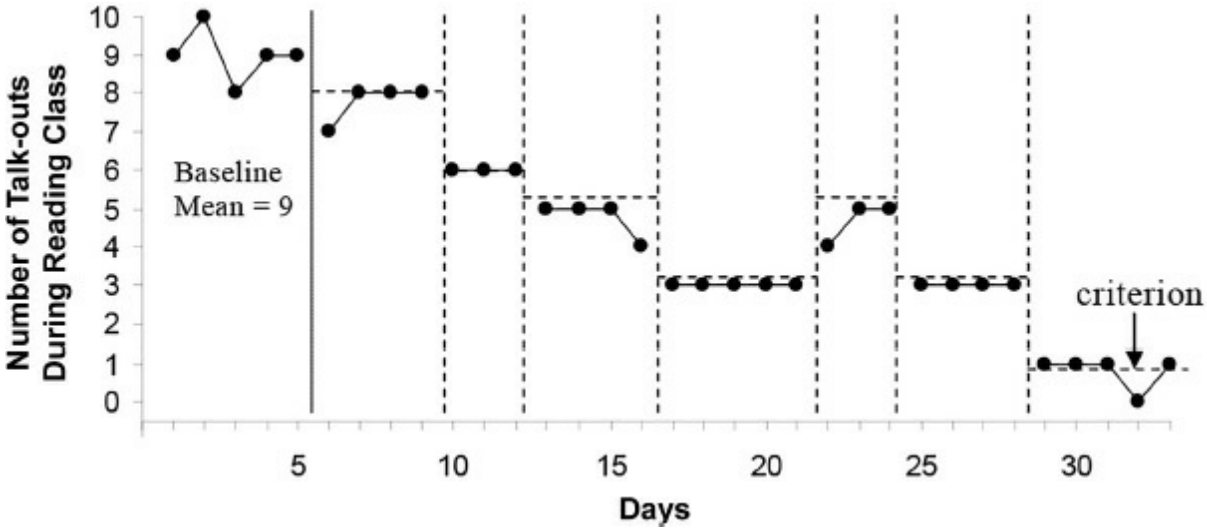


Figure 12.1 Prototype changing criterion design (talk-outs, decrease).

Procedural Guidelines

Changing criterion designs are useful for answering a small but important group of research questions, including the use of differential reinforcement strategies. When using a changing criterion design, adhere to the following guidelines:

1. Identify and define a reversible target behavior that is likely to closely align with changing criteria.
2. Select a sensitive, reliable, valid, and feasible data collection system and pilot the system and your behavior definitions.
3. Determine a priori frequency of reliability and fidelity data collection (e.g., 33% of sessions), and conduct data collection for the duration of the study.
4. Determine a priori your criterion changes, varying the magnitude and planning at least reversal (see below).
5. Collect continuous baseline data (A) on target behaviors for a minimum of 3 consecutive days or until data are stable.

6. Implement intervention condition with the first criterion in place.
7. Change the criterion level only after stable criterion-level responding has been attained in the preceding phase.
8. Repeat for at least four changes in criterion level (i.e., replications).
9. Allow unconstrained responding during all sessions (e.g., if you plan to systematically increase the number of minutes of exercise from 5 minutes to 20 minutes, do not prevent the participant from exercising for the full 20 minutes during each session).
10. Replicate with similar participants.

Internal Validity

Studies have adequate internal validity when all likely threats are controlled for, and experimental control is demonstrated when adequate internal validity is present and when the level of the dependent variable closely corresponds to criterion levels.

The most critical threat to internal validity is maturation, given the slow, step-wise changes in changing criterion designs. To control for this threat, introduce at least one contra-therapeutic criterion change. If behaviors worsen to correspond to the new criterion levels, maturation is not a likely threat. There are no design-specific concerns when detecting and controlling for instrumentation, data instability, and fidelity threats; typical procedures for detecting and controlling for these threats should be used (see Chapter 1 and Table 10.1).

To minimize the risk of bias, you should pinpoint criterion levels, or a strategy for determining criterion changes before initiating a study. If data consistently follow each of these criterion changes, there is a high probability that the intervention is responsible for changes in the target behavior.

It is important *not* to constrain responding by preventing your participant from reaching the target goal during initial sessions. For instance, if you wanted to evaluate the effects of contingent reinforcement on increasing number of minutes of exercise, and your first criterion is 7 minutes, you should not prevent the participant from working out for more than 7 minutes. During each session, participants should have equal opportunities to engage in the behavior, even though reinforcement may be contingent on a specific level of performance. An example of constrained responding would be providing a student with a worksheet with increasing numbers of problems across criterion levels such that the student could not respond to more problems than the set criterion.

Variations

Two variations of changing criterion designs are noteworthy. The first is a changing criterion design with behavior measurement across response classes; the second is useful

when several mutually exclusive behaviors are measured, with different contingencies applying to each. In the typical changing criterion design, behavior is shaped within a response class (e.g., the amount of behavior is changed; the topography is not). A variation of the design may be used such that progressively difficult behaviors or similar behaviors under different environmental conditions are required. For example, Koegel et al. (2004) required changes behavior topography to decrease problem behavior related to noise sensitivity. Changes in criterion for one participant included: Walking by closed bathroom door without toilet flushing, standing 75 feet from open bathroom door while toilet is flushed; standing inside closed stall while toilet is flushed. A similar modified changing criterion design was used by Birkan, Krantz, and McClannahan (2011) to teach children with autism to cooperate with injections.

The second variation of the changing criterion design is the distributed criterion design (McDougall, 2006). This design is appropriate when varying amounts of time should be allotted for engaging in multiple, mutually exclusive tasks. McDougall provides an applied example related to research productivity and work towards three writing tasks. This design might also be appropriate for shaping appropriate social behavior (e.g., reinforcing a certain amount of time spent on responding to peers, listening to peers, and engaging in solitary behavior) or independent after school behaviors (e.g., completing chores, engaging in physical activity, doing homework, and playing video games) that may vary over time. This design has not been widely used but may be advantageous when several mutually exclusive behaviors are of interest.

Advantages

The changing criterion design is appropriate to evaluate programs designed to shape behaviors that are in a person's repertoire but do not occur at an acceptable rate. Unlike MB and MP designs across behaviors, the changing criterion design has the advantage of requiring only one target behavior. In contrast to the A-B-A-B design, no withdrawal condition is required, though one return to a preceding criterion level is recommended, which can strengthen the demonstration of experimental control. Most importantly, at least from an educator's or therapist's perspective, the changing criterion design, through its small-step increments in criterion level, permits a student or client to change behavior slowly, perhaps decreasing the likelihood of failure due to a sudden large change in response effort.

Limitations

There are, perhaps, two reasons that the changing criterion design has been used infrequently: first, it is limited to a relatively small range of target behaviors and instructional procedures. The changing criterion design is not an appropriate paradigm to assess acquisition of new behaviors. Consequently, the changing criterion design is

limited to programs that manipulate consequences for the purpose of increasing or decreasing the frequency of behaviors already established in an individual's repertoire.

The second limitation is that it can be difficult to determine appropriate criterion levels. Whenever you are required to specify a criterion level of acceptable performance, there is always some degree of subjectivity or "professional guesswork" involved. The investigator who uses the changing criterion design has the tedious responsibility of making criterion changes that are large enough to be "detectable", small enough to be "achievable", and not so small that the behavior will far exceed the criterion level. In other words, a demonstration of experimental control depends upon an "a priori" prediction or strategy for setting a progression of criterion levels, as well as acceptable response ranges at each criterion level, predictions that may or may not prove appropriate. One strategy that takes some of the guesswork out of deciding individual criterion levels is to decide upon a percentage to change the criterion in each subsequent phase (e.g., 20–50% increase over the previous phase). For example, if during the baseline condition the mean frequency of daily talk-outs for a class of 12 students was 27, you could use a "criterion size change rule" of 15% of the preceding phase to determine the acceptable number of talk-outs for the next phase. The first criterion level would be 4 fewer talk-outs by the group (27 x 0.15 = 4.05 rounded to the nearest whole number), that is, 23 talk-outs for the group to access the reinforcer; the second criterion level would be 15% lower than the preceding criterion level (23 x 0.15 = 3.45 rounded to the nearest whole number), or 3 fewer talk-outs by the group, 20; the third criterion level would be 15% lower than the preceding criterion level (20 x 0.15 = 3) or 3 fewer talk-outs by the group, 17.

Applied Example 12–1: Changing Criterion Design

Johnston, R. J., & McLaughlin, T. F. (1982). The effects of free time on assignment completion and accuracy in arithmetic: A case study. *Education & Treatment of Children*, *5*, 33–40.

In this study, a changing criterion design was used to assess the effect of contingent free time on completion of daily math assignments during a daily math segment in a self- contained 2nd-grade classroom. The participant, a 7-year-old girl, had low daily worksheet completion, despite the fact that she consistently scored above grade level on achievement tests and averaged 100% correct on attempted problems. Daily assignments ranged from 6–43 items, including computational and "thought" problems. The two dependent variables were percentage of problems completed and percentage correct per assignment. To provide interobserver reliability checks a parent-aide re-graded at least one daily assignment during each phase of the study. For both dependent measures, point-by-point (i.e., problem-by-problem) reliability checks yielded 100% agreement.

During an initial 10-day baseline condition the second grade teacher presented the child with her daily math assignment, worked one of the problems as a model, and asked the student to complete as many problems as possible within the 35 minute session. At the end of this baseline period, an average daily baseline completion rate was computed at 35%. A changing criterion procedure was then introduced in which investigators "successively changed the criterion for reinforcement, usually in graduated steps, from baseline level until the desired terminal behavior was achieved" (p. 35). Most new criterion levels required a 5% increment in percentage completion above the preceding level (i.e., Phase 1 = 35%, Phase 2 = 40%, Phase 3 = 45%) for three consecutive days. During the intervention condition the teacher continued to present the child with the daily assignment and worked one example as a model; however, in addition she informed the child of the minimum number of problems that had to be completed accurately in order for criterion to be met and reminded her, that upon meeting the criterion, she was eligible to enjoy free time for the remainder of the 35-minute time period. If she did not meet criterion within the allowable 35 minutes, she was required to remain at her desk until criterion was attained.

A changing criterion design was used to assess intervention effectiveness, shown in Figure 12.2, which included 16 criterion changes during the intervention condition. A brief reversal in criterion level was instituted to strengthen the demonstration of experimental control. The investigation concluded with a three-session follow-up condition that was identical to baseline conditions (e.g., no free play contingency). The child met or exceeded all identified criterion levels in

interventions conditions. During the final intervention phase, she demonstrated 100% assignment completion. Furthermore, her accuracy remained stable throughout the study, even though greater numbers of more difficult problems were selected.

This study provides demonstration of an assessment of a simple and convincing strategy to shape a child's percentage of task completion by requiring small increments in performance to ensure continued successful responding. Interestingly, after the 60th session the investigators withdrew the free time contingency (i.e., reinstated baseline condition) and found that 100% task completion and accuracy were maintained when measured by three follow-up probes conducted 5, 15, and 25 days after terminating the intervention condition. Furthermore, the researchers reported that the free time contingency required very little teacher time or expense. As a final note, you might suspect that the choice of 5% level increments may have been smaller than actually required since the child abruptly reached and often exceeded each new criterion level. Perhaps the outcome objective could have been attained more rapidly had the step-wise requirements been greater. This demonstrates one difficulty with using changing criterion designs—the challenge of identifying criterion levels that will permit the demonstration of experimental control without impeding optimal learning rates.

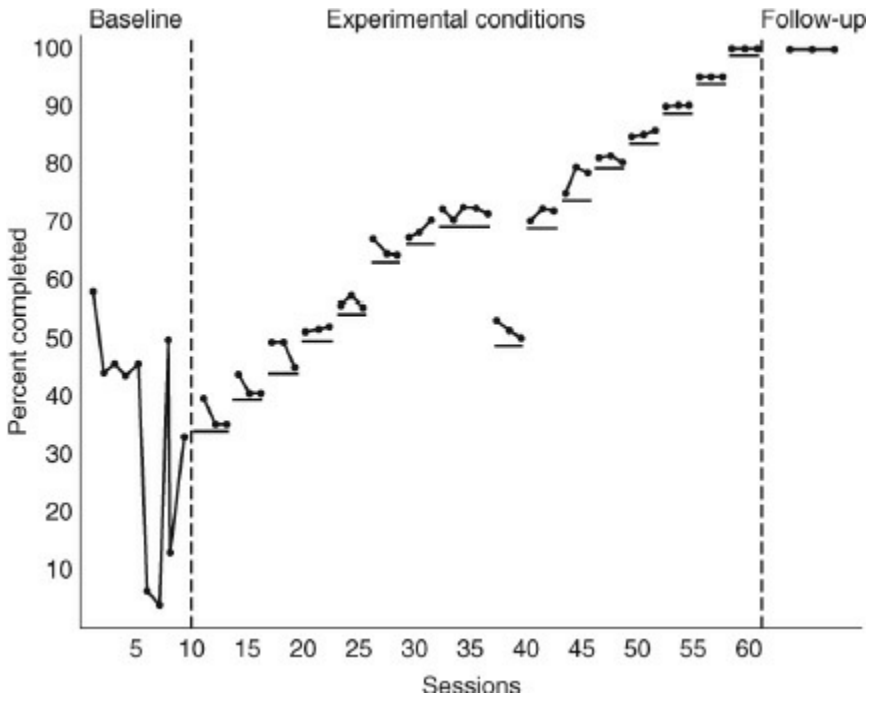


Figure 12.2 Changing criterion design with one reversal.

Source: Johnston, R. J., & McLaughlin, T. F. (1982). The effects of free time on assignment completion and accuracy in arithmetic: A case study. *Education and Treatment of Children*, *5*, 35–40.

Conclusions

The changing criterion design is one of several experimental paradigms available to educators and clinicians to evaluate the effectiveness of intervention programs. Table 12.1 summarizes several studies that have used the changing criterion design in clinical and educational settings. Though it has not been cited as frequently in the applied research literature as other common designs, it does offer a practical way to monitor performance when stepwise criterion changes are both desirable and practical. It can be used to monitor programs designed to increase or decrease the rate of responding. Those who decide to employ the changing criterion design must closely follow guidelines previously discussed. You are cautioned to use the changing criterion design only if the target response is in the individual's repertoire and the objective of the intervention is to increase or decrease the frequency responding. Under these conditions you will find the changing criterion design appropriate and useful in evaluating program effectiveness.

Table 12.1 *Studies Using Changing Criterion Designs*

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
DeLuca, R. V., & Holborn, S. W. (1992). Effects of a variable-ratio reinforcement schedule with changing criteria on exercise in obese & nonobese boys. *Journal of Applied Behavior Analysis, 25*, 671–679.	Number: 6 Sex: M Age: 11 Disability/ Diagnosis: Obesity (3)	Setting: Public school clinic Arrangement: Individual	Tangible reinforcement based on reaching criterion levels of responding	Revolutions per minute, duration of session
Easterbrooks, S. R., & Stoner, M. (2006). Using a visual tool to increase adjectives in the written language of students who are deaf or hard of hearing. *Communication Disorders Quarterly, 27*, 95–109.	Number: 3 Sex: 2 M, 1 F Age: 17–18 Disability/ Diagnosis: Severe to profound hearing loss	Setting: Special education classroom Arrangement: Individual	Printed visual aid for writing	Number of adjectives included in a written product
Flood, W. A., & Wilder, D. A. (2004). The use of differential reinforcement and fading to increase time away from a caregiver in a child with separation anxiety disorder. *Education and Treatment of Children, 27*, 1–8.	Number: 1 Sex: M Age: 11 Disability/ Diagnosis: ADD & anxiety disorder	Setting: Clinic Arrangement: Individual	DRO procedure (student was reinforced for absence of emotional behavior)	Latency to emotional behavior after departure of mother
Ganz, J. B., & Sigafoos, J. (2005). Self-monitoring: Are young adults with MR and autism able to utilize cognitive strategies independently? *Education and Training in Developmental Disabilities, 40*, 24–33.	Number: 2 Sex: M Age: 19–20 Disability/ diagnosis: Autism (1), ID (2)	Setting: Special education classroom Arrangement: Individual	Self-monitoring training (visual system to track successes & reinforcement for an identified number of tokens)	Participant 1: Number of independent tasks completed, Participant 2: Number of independent requests for help
Grey, I., Healy, O., Leader, G., & Hayes, D. (2009). Using a Timer to increase appropriate waiting behavior in a child with developmental disabilities. *Research in Developmental Disabilities, 30*, 359–366.	Number: 1 Sex: F Age: 11 Disability/ diagnosis: Cerebral palsy, moderate ID	Setting: Special education classroom Arrangement: Individual	Discrimination training	Number of seconds of appropriate waiting behavior
Luiselli, J. K. (2000). Cueing, demand fading, and positive reinforcement to establish self-feeding and oral consumption in a child with food refusal. *Behavior Modification, 24*, 348–358.	Number: 1 Sex: M Age: 4 Disability/ diagnosis: Lung disease & g-tube dependence	Setting: Home Arrangement: Individual (instruction provided by parents)	Positive reinforcement in the form of tangibles for self-feeding	Number of self-fed bites per session

Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
Nes, S. L. (2005). Using paired reading to enhance the fluency skills of less-skilled readers. *Reading Improvement, 40,* 179–192.	Number: 4 Sex: 3 M, 1 F Age: 9–12 Disability/ diagnosis: None	Setting: School library Arrangement: Individual	Oral reading by a skilled reader prior to reading of the passage by target participant	Number of words read per minute
O'Connor, A. S., Prieto, J., Hoffmann, B., DeQuinzio, J. A., & Taylor, B. A. (2011). A stimulus control procedure to decrease motor and vocal stereotypy. *Behavioral Interventions, 26,* 231–242.	Number: 1 Sex: M Age: 11 Disability/ Diagnosis: Autism	Setting: Clinic room and special education classroom Arrangement: Individual	Discrimination training	Percent of intervals with stereotypy
Rapp, J. T., Cook, J. L., McHugh, C., & Mann, K. R. (2017). Decreasing stereotypy using NCR and DRO with functionally matched stimulation: Effects on targeted and non-targeted stereotypy. *Behavior Modification, 41,* 45–83. (Study 4)	Number: 3 Sex: M Age: 3–7 Disability/ Diagnosis: Autism spectrum disorders	Setting: Clinic room Arrangement: Individual	Differential reinforcement of other behaviors plus response interruption	Latency to stereotypy (targeted and non-targeted), and problem behavior (for Xander only)
Ricciardi, J. N., Luiselli, J. K., & Camare, M. (2006). Shaping approach responses as intervention for specific phobia in a child with autism. *Journal of Applied Behavior Analysis, 39,* 445–448.	Number: 1 Sex: M Age: 8 Disability/ Diagnosis: Autism, specific phobia	Setting: Inpatient clinic Arrangement: Individual	Differential reinforcement of approach responses	Distance from avoided stimuli
Schumacher, B. I., & Rapp, J. T. (2011). Increasing compliance with haircuts for a child with autism. *Behavioral Interventions, 26,* 67–75.	Number: 1 Sex: M Age: 5 Disability/ Diagnosis: Autism	Setting: Home Arrangement: Individual	Differential reinforcement for sitting	Number of seconds sitting in chair
Warnes, E., & Allen, K. D. (2005). Biofeedback treatment of paradoxical respiratory distress in an adolescent girl. *Journal of Applied Behavior Analysis, 38,* 529–532.	Number: 1 Sex: F Age: 16 Disability/ Diagnosis: Paradoxical vocal fold motion	Setting: Outpatient clinic Arrangement: Individual	Biofeedback instruction	Electromyography, anecdotal pain and adaptive functioning reports

Note: M=male, F=female, ADD=attention deficit disorder, ID=intellectual disability, DRO=differential reinforcement of other behaviors

Simultaneous Treatments Designs

Simultaneous treatments (ST) designs have the single purpose of describing choice behavior when two concurrently available conditions exist. Researchers can use this design when two or more options are simultaneously available and when a participant's choice between the options is of interest. For example, if a researcher is interested in whether a child with a disability prefers to play in a center with or without peers present, the researcher could design a study such that the two conditions were simultaneously available. These conditions would be identical except for the presence of peers. During each session, the researcher could give the child with a disability the choice of playing in either area (e.g., "You can play *here* or *here*"). The researcher could then measure the number of times each area was chosen, or the percentage of time during which the child remained in each area. A consistent difference in the number of times chosen or amount of time spent in each condition, across sessions, is important for determining that a functional relation exists.

ST designs were differentiated from alternating treatments designs (ATD) and multi-element designs (M-ED) by Barlow and Hayes in 1979, and have been termed "concurrent operants" paradigms, particularly in the basic literature (e.g., *Journal of the Experimental Analysis of Behavior*), where they have been widely used since the 1960s (e.g., Duncan & Silberberg, 1982; Hackenberg & Joker, 1994; Richardson & Clark, 1976). Occasionally (particularly in "old" studies), investigators will use the term "simultaneous treatments design" to refer to alternating treatments designs; this is not the current preferred terminology.

ST designs should be used when choice behavior is the dependent variable of interest. In recent years, ST designs have been used to evaluate participant preferences for: embedded versus massed trial prompting (Heal & Hanley, 2011); making choices (Schmidt, Hanley, & Layer, 2009); punishment and extinction-components of interventions (Hanley, Piazza, Fisher, & Maglieri, 2005); and video versus in-vivo modeling (Geiger, LeBlanc, Dillon, & Bates, 2010). In addition, Tullis, Cannella-Malone, and Fleming (2012) used an ST design in combination with an A-B-A-B design to determine whether preferences for stimuli changed over time within preference assessment sessions.

ST designs may be appropriate to determine which is a client-preferred intervention when more than one is appropriate, practical, and effective (e.g., functional communication training and noncontingent reinforcement; Hanley, Piazza, Fisher, Contrucci, & Maglieri, 1997). For example, in one study, AATDs were used to compare the acquisition of pre-academic behaviors for 12 young children who received massed trial instruction and instruction embedded into game play (Ledford, Chazin, Harbin, & Ward, 2017). Concurrently, researchers used ST designs to evaluate children's preference for each instructional variation; these data were plotted on cumulative graphs.

Approximately half of the children preferred massed trials while the other half preferred embedded trials; most children preferred the format that resulted in most efficient learning for them.

In another study using an ST design, Slocum and Tiger (2011) used backward and forward chaining to teach non-functional chained tasks to students with ADHD, speech delays, or learning disabilities. All four participants learned the tasks, with relatively little variation in the number of trials to mastery for forward and backward chaining procedures (see top panel in Figure 12.3). Experimenters also used an ST design to evaluate preference for each procedure, and displayed results in a cumulative graph. Three of four participants had similar numbers of selections for each procedure; the fourth participant showed a consistent preference for forward chaining (see bottom panel in Figure 12.3).

Procedural Guidelines

When using an ST design, adhere to the following guidelines:

1. Identify two or more interventions or contexts that will be concurrently available to participants.
2. Determine a priori frequency of reliability and fidelity data collection (e.g., 33% of sessions), and conduct data collection for the duration of the study.
3. Identify a procedure by which participants can choose one of the available interventions and behaviorally define a choice (e.g., movement to a specific area, pointing to a picture, naming one of the options when asked "Which one do you want to do?").
4. Ensure participants are able to discriminate between conditions (e.g., that they can make an "informed" choice between interventions).
5. Repeatedly measure the degree to which participants choose one intervention more than the others (e.g., cumulative count of number of times chosen, percentage of times chosen per session, percentage of time spent in each of two areas).
6. Replicate with similar participants.

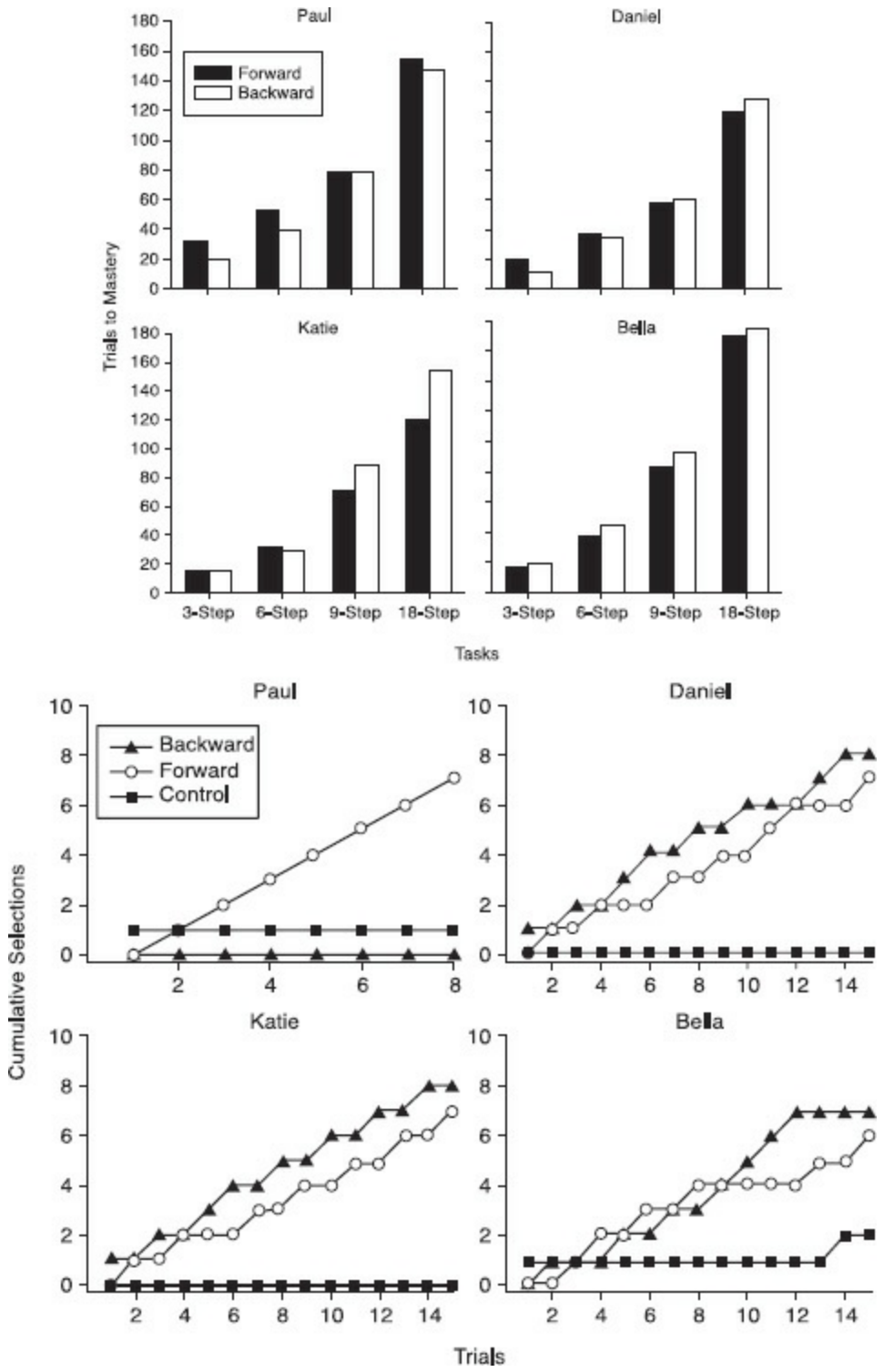


Figure 12.3 Data related to acquisition and choice behaviors for participants taught using backward and forward chaining in the context of a simultaneous treatments design (bottom).

Source: Slocum, S. K., & Tiger, J. H. (2011). An assessment of the efficiency of and choice for forward and backward chaining. *Journal of Applied Behavior Analysis*, *44*, 793–805.

Advantages

The ST design is uniquely appropriate for assessing choice behavior of participants. When the DV of interest is choice or preference among several concurrently available options, the ST design should be used. This design may be helpful, alongside another SCD, when researchers are interested in both effectiveness of multiple interventions (e.g., evaluated using an ATD or AATD) and the preference of participants regarding which intervention should be used. This design is a useful way to assess the social validity of different interventions, even when participants might not be able to verbally respond to questions regarding their preferences about instruction.

Limitations

The ST design is appropriate for assessing the choice behavior of participants; no other dependent variables should be assessed with this design. When using this design, it may be difficult for investigators to verify that participants were making "informed" choices rather than non- discriminately choosing one option, particularly for participants with limited communication skills.

Conclusions

Although ST designs are appropriate for a restricted number of research questions, they are helpful and appropriate when measuring participant choice and preference behaviors. They are best used in conjunction with other SCDs, especially those used to compare effectiveness and efficiency of two or more interventions (e.g., ATD and AATD).

Repeated Acquisition Designs

Repeated acquisition (RA) designs, like ST designs, are rarely reported in SCD literature, although a few recent examples of use do exist (Bouck, Flanagan, Joshi, Sheikh, & Schleppenach, 2011; Spencer et al., 2013; Sullivan, Konrad, Joseph, & Luu, 2013). However, RA designs are much more broadly applicable than ST designs, and are one of few SCDs appropriate for comparing interventions for teaching non-reversible behaviors (e.g., for comparing academic interventions). RA designs are used when the behaviors of interest will be quickly acquired by the participant (e.g., during one or only a few sessions), and when two interventions are being compared (e.g., errorless prompting versus non-errorless prompting). Recently, RA designs have also been used to evaluate single interventions (e.g., demonstration questions; Butler, Brown, & Woods, 2014).

When using this design, many behaviors should be identified (e.g., 100 sight words). For this design, it is preferable that all target stimuli are of equal difficulty; in this case, you can use Method 1 for assigning stimuli:

1. Randomly assign each stimulus to one intervention.
2. Randomly assign each stimulus to a set.
3. Randomly assign each set of stimuli to an intervention order (separately for each intervention).

If all stimuli are not of equivalent difficulty, you can use Method 2 for assigning stimuli:

1. Divide the stimuli into sets of equal difficulty.
2. Randomly assign half the stimuli from each set to each of two interventions.
3. Keeping matched sets together, randomly assign each set of stimuli to an intervention order.

RA designs are similar to ATDs; authors sometimes identify use of an ATD when they use rapidly alternating instructional conditions to teach different sets of non-reversible behaviors during each session (Bickford & Falco, 2012; Malanga & Sweeney, 2008). However, as mentioned in Chapter 11, true ATDs measure a *single reversible* behavior of interest in two or more conditions and AATDs measure *multiple non-reversible behaviors* repeatedly over time. In RA designs, the non-reversible behaviors of interest change every session or every few sessions. For example, although you might be interested in word-reading throughout the study, the actual words taught will vary across sessions in an RA design. An ATD is not appropriate for measuring these behaviors. Often, when authors use the variation they term as an ATD—using rapidly alternating conditions and frequently-changing non-reversible behaviors, they omit the

pretest session that is generally used in RA designs—there is an assumption by these investigators that the baseline performance would be at floor levels (e.g., 0% correct). Using RA designs with a single pre-test measure is preferable to this design variation without any pre-test (baseline) measurement.

One example of when it would be appropriate to use an RA design is if you want to compare errorless and non-errorless prompting procedures (EP and NEP, respectively) to teach word-reading to a student who is likely to reach acquisition quickly. After dividing the words into sets, you would conduct a pre-test for only the first 5 words assigned to EP (EP Set 1) and the first 5 words assigned to NEP (NEP Set 1). This single probe session is represented as the first data point for each set in Figure 12.4. Then you would immediately teach both sets, rapidly alternating between conditions (e.g., one EP session and one NEP session each day). Immediately after teaching the first set, you would do a single pre-test session for the second sets of words assigned to each condition (EP 2nd set & NEP 2nd set), then you would teach those sets. This would continue through all sets of words. A functional relation is demonstrated (and one intervention is deemed superior) if the time to acquisition is *consistently* lower for one intervention than for another. In Figure 12.4, for example, EP resulted in faster acquisition for each of the 10 hypothetical comparisons. In general, each set is taught for a pre-specified number of sessions (e.g., one session in each condition; four sessions in each condition) and acquisition is measured on the final day of instruction. Alternatively, and from a clinical standpoint, preferably, you could teach until the participant reaches a pre-set criterion on at least one set of behaviors, as shown in Figure 12.4. Because of the relative dearth of data in this design (similar to the ATD), we recommend completing a minimum of five (rather than three) comparisons (potential demonstrations of effect).

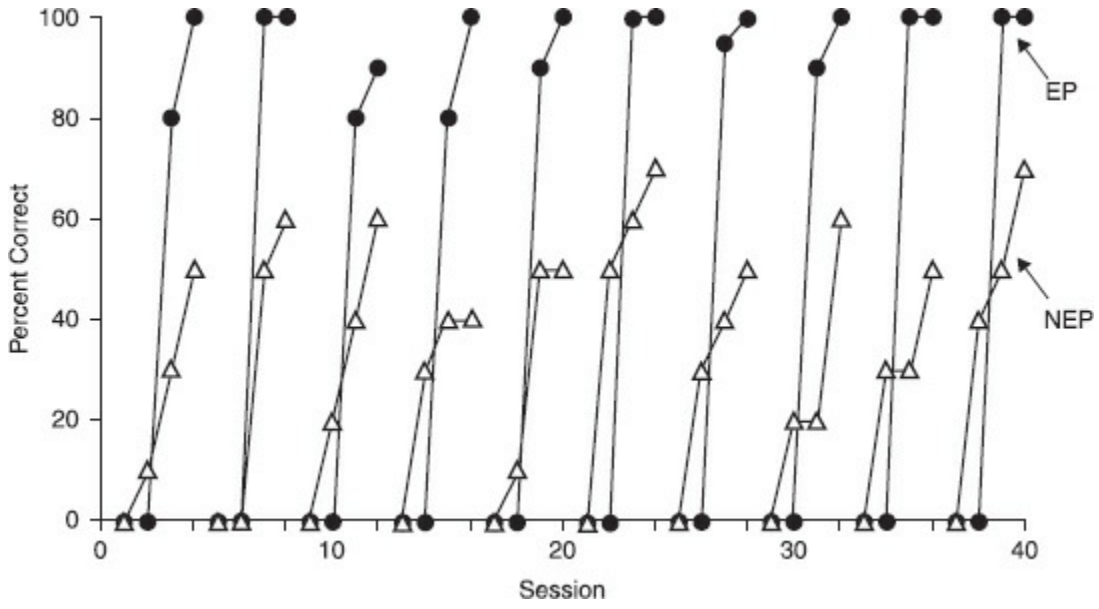


Figure 12.4 Hypothetical data using repeated acquisition design to compare errorless prompting (EP) to non-errorless prompting (NEP).

The data in Figure 12.4 are presented in keeping with the typical preferences for RA design. One variation on this presentation, when acquisition is measured during a single session, is to show data for each comparison with connected data points; these data look the same as those presented using ATDs (see Figure 12.5).

In a recent variation using an RA design, Spencer and colleagues (2012) used the RA design to answer a question that included a kind of component analysis. They evaluated whether an intervention designed to increase vocabulary and comprehension was more effective than repeated reading alone (see Figure 12.6). You should note that Spencer and colleagues only used the comparison for the first three sets of behaviors; this is not advised from a research standpoint, given the minimal baseline measurements used in RA designs.

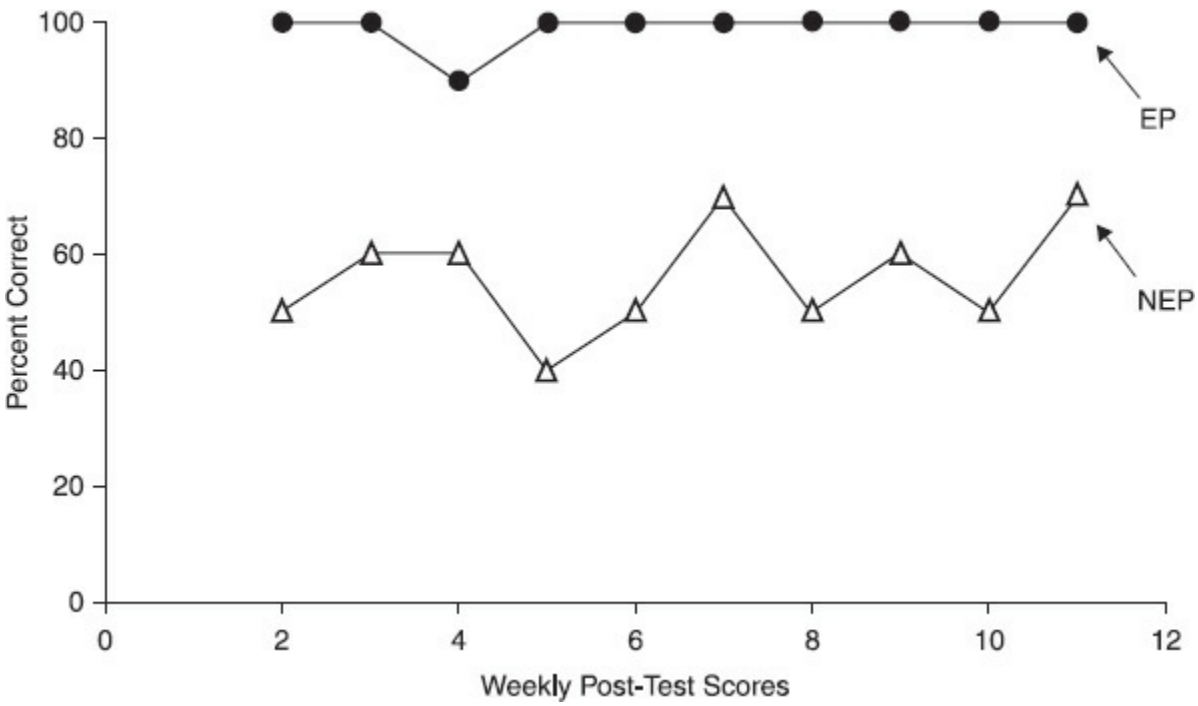
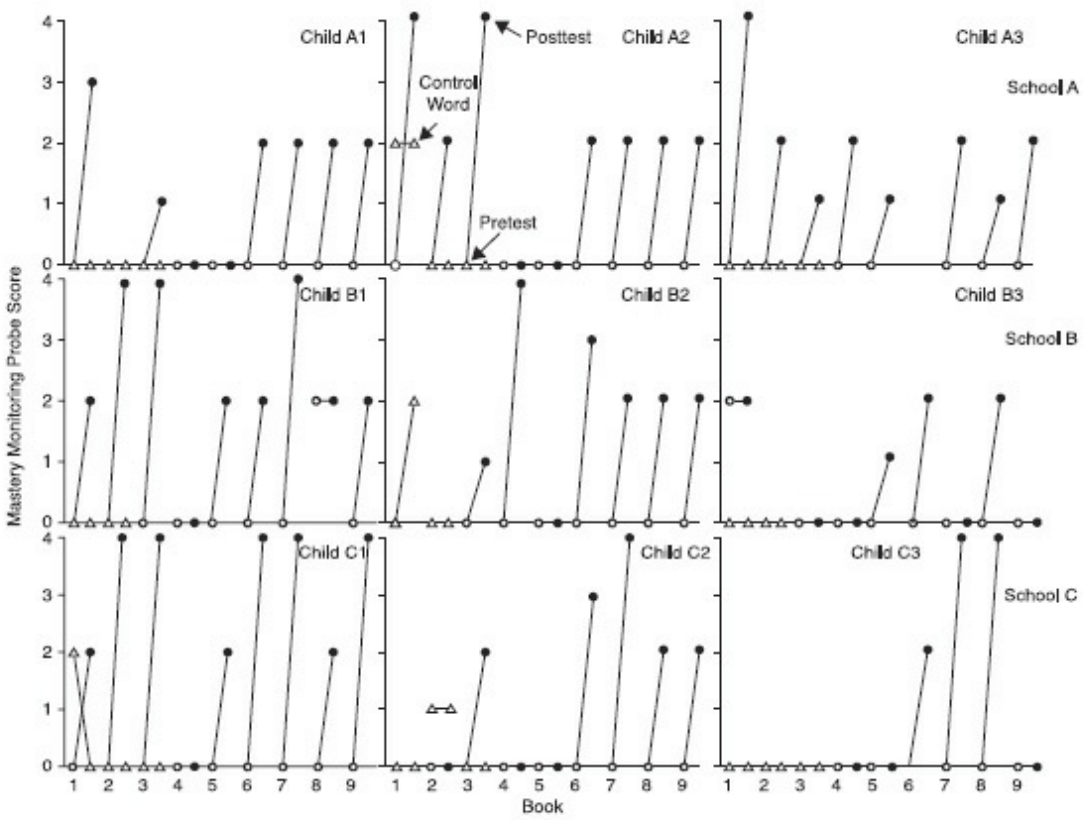


Figure 12.5 Variation of graphing data using repeated acquisition design; these data correspond to those shown in Figure 12.5. No pre-instruction data are plotted.

Figure 1

Vocabulary mastery monitoring probe scores at pretest and posttest



Note: Maximum score at pretest and posttest was 4 for each book and included the points for the two vocabulary targets for each book. Untaught words were assessed at pretest and posttest for three books. For untaught words, maximum score at pretest and posttest was 2 and included the points for one untaught word for each book. The three children at School B did not complete the items for untaught words for the third book. Child A3 was unavailable for the posttest for the sixth book.

Figure 12.6 Variation of repeated acquisition design with control sets for the first three comparisons.

Source: Spencer, E. J., Goldstein, H., Sherman, A., Noe, S., Tabbah, R., Ziolkowski, R., & Schneider, N. (2012). Effects of an automated vocabulary and comprehension intervention: An early efficacy study. *Journal of Early Intervention*, *34*, 195–221.

Guidelines

When using an RA design, adhere to the following guidelines:

1. Identify many (e.g., 20 or more) non-reversible behaviors that are unknown, but will be quickly acquired by participants (this will likely require a screening condition).
2. Select a sensitive, reliable, valid, and feasible data collection system and pilot the system and your behavior definitions.
3. Determine a priori frequency of reliability and fidelity data collection (e.g., 33% of sessions), and conduct data collection for the duration of the study.

4. Assign behaviors to sets, using Method 1 or Method 2 (described above).
5. Conduct one probe session for the first behavior set.
6. Provide instruction for each of the first behavior sets, using a different intervention for each, quickly alternating between interventions, until acquisition has been demonstrated.
7. Repeat steps 5 and 6 for each remaining comparison.
8. Replicate with similar participants.

Advantages

The RA design is appropriate when comparisons between two interventions are of interest and when the dependent variables of interest are non-reversible behaviors that will be rapidly acquired by all participants. The RA design is potentially more advantageous for practitioners because it does not require repeated testing prior to introduction of an intervention like the AATD, multiple probe across behaviors, and parallel treatments designs. Unlike the AATD, the RA design includes multiple comparisons for each participant (intra-participant replication) and results in a quick comparison between interventions.

Limitations

There are methodological disadvantages rendering RA designs less desirable than other designs intended to evaluate non-reversible behaviors (multiple probe, AATD, PTD). No post-instruction probe conditions are built into this design; therefore, there are no built-in opportunities for assessing short-term maintenance. Studies using this design have rarely evaluated maintenance; thus, conclusions regarding the efficiency and effectiveness of interventions compared may be incomplete. In addition, baseline measurement using this design is usually measured during one pre-instruction session; evaluation of potential threats due to history and maturation are not possible; potential increasing trends also cannot be evaluated. Because of these methodological constraints, it is best to include many comparisons for each participant.

Conclusions

The repeated acquisition design provides a relatively fast comparison between instructional conditions for teaching non-reversible behaviors. Although there are considerable methodological disadvantages, limited pre-instruction testing and relative speed may make the use of this design compelling for practitioners.

Brief Experimental Designs

The **brief experimental (BE) design** is a group of SCDs that are variations of commonly used designs, specifically withdrawal (A-B-A-B; e.g., McComas et al., 1996) and alternating treatments (M-ED variation; McComas et al., 2009; Mong & Mong, 2012) designs. The BE design requires fewer sessions, which makes their use practical in applied settings; however, fewer replications reduce confidence in conclusions. Often, the BE design has been used when conducting functional analyses, and has often been followed by the analysis of an intervention using a second SCD that confirms the findings from the brief design used. The BE design may be preferable to the more usual M-ED designs when evaluating the results of a functional analysis by decreasing the probability of problem behavior associated with an extended ME-D design. Recently, Martens and Gertz (2009) discussed the use of brief experimental designs in the context of functional analysis of behavior, stating that the use of these procedures were "… gaining recognition among researchers and educators alike as a valuable tool for making treatment decisions about children who are unresponsive to classroom instruction" (p. 93). These designs may be particularly effective for determining which among several interventions is likely to be effective for increasing desirable academic behaviors (Martens, Eckert, Bradley, & Ardoin, 1999). When each condition in an FA is evaluated only once and a tentative relation is found (e.g., problem behavior is highest during "escape" condition when compared to all other conditions), you can then further evaluate only the condition with high levels of problem behavior without exposing the participant to additional conditions that may result in increased problem behavior. For example, in the case of high levels of problem behavior in an "escape" condition, you could conduct an intervention whereas escape extinction is evaluated in the context of an A-B-A-B withdrawal design following the brief analysis.

One published example of the use of a BE design (LeGray, Dufrene, Sterling-Turner, Olmi, & Bellone, 2010) shows its use to test four conditions in an FA (attention, tangible, escape, and free play) conducted in regular education preschool and kindergarten classrooms. Rather than using the traditional M-ED design, with three to five replications for each condition, LeGray et al. conducted each condition during only one 10-minute session. Then additional conditions were introduced by using a contingency reversal for the one condition, of the original four, that had the highest level of problem behavior (e.g., true reversal; if attention resulted in high levels of problem behavior, the contingency reversal resulted in the student being ignored when engaging in problem behavior and being given attention for any non-problem behaviors) Then, the condition with the highest levels of problem behavior was re-instated (top panel of Figure 12.7), and finally, another contingency-reversal condition was introduced. For each condition, a single datum point (single session) was used. Following confirmation the contingency reversal resulted in low levels of problem behavior, additional intervention data were

collected in the context of an ATD (see bottom panel of Figure 12.7), which confirmed the results of the functional analysis conducted in the context of the BE design.

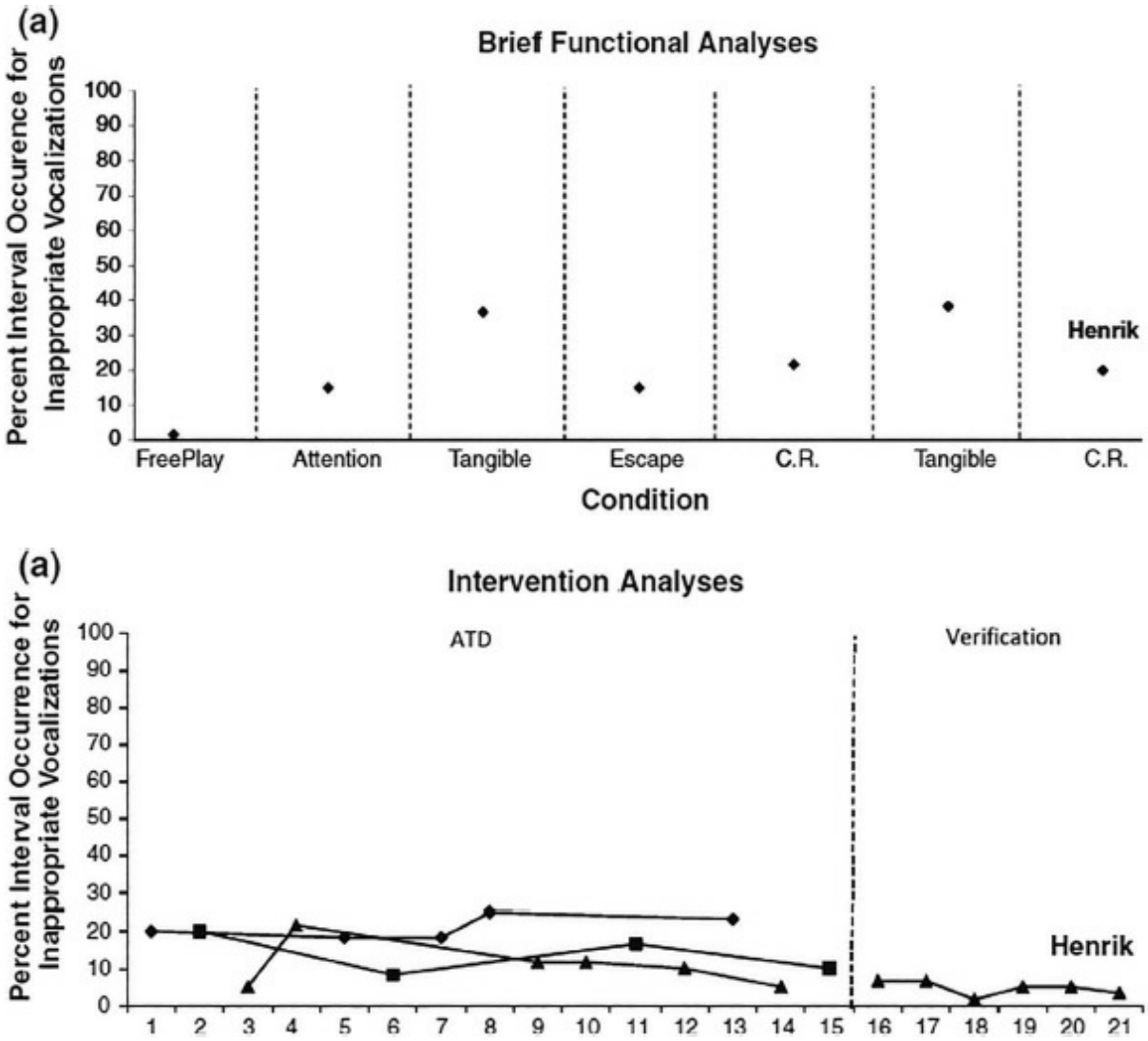


Figure 12.7 Brief experimental design followed by ATD evaluation. Source: LeGray, M. W., Dufrene, B. A., Sterling-Turner, H., Olmi, D. J., & Bellone, K. (2010). A comparison of function-based differential reinforcement interventions for children engaging in disruptive classroom behavior. *Journal of Behavioral Education*, *19*, 185–204.

Procedural Guidelines

When using a BE design, adhere to the following guidelines:

1. Identify and define a reversible target behavior.
2. Select a sensitive, reliable, valid, and feasible data collection system and pilot the system and your behavior definitions.
3. Determine a priori frequency of reliability and fidelity data collection (e.g., 33% of sessions), and conduct data collection for the duration of the study.
4. Introduce each condition during at least one session.

5. If a relation between levels of the dependent variable and one condition exists (e.g., behavior is lower or higher during one condition), confirm the relation by evaluating: (a) a contingency reversal, and/or (b) an intervention based on the relation.
6. Replicate with similar participants.

Advantages

Comparative data suggest functional analysis using the BE design is effective, resulting in the same conclusions drawn from extended analyses (Mong & Mong, 2012). Thus, indigenous implementers (e.g., practitioners) can spend less time evaluating the function of a behavior, resulting in faster implementation of effective interventions.

Limitations

Brief functional analyses, though shown to be accurate and effective in leading to implementation of effective interventions, do not have adequate replication. When used, further confirmation using a different design is needed to confirm a functional relation. Confirmation can be accomplished using a contingency reversal (A-B-A-B), where "A" is the condition in the brief analysis that resulted in higher levels of problem behavior, for example. You can also confirm findings by implementing an intervention (chosen using data from the brief analysis) in the context of another experimental design (e.g., a multiple baseline across behaviors).

Conclusions

Brief experimental designs have been used with increasing popularity (e.g., Cihak, Alberto, & Fredrick, 2007; Dufrene, Watson, & Kazmerski, 2008; LeGray et al., 2010; McComas et al., 2009; Mong & Mong, 2012; Petursdottir et al., 2009; Ward & Higbee, 2008). They may be helpful when an initial assessment is needed to determine the function of a behavior. However, additional confirmation (replication of effect) using a second SCD is needed.

Combination Designs

In this section we present a case for combining designs when the research question(s) or circumstances call for it. SCDs are combined when (a) planning a study, or (b) attempting to salvage experimental control during a study in progress. More specifically, applied researchers combine designs to:

1. Answer more than one research question in one investigation. For example, you might compare baseline and intervention conditions within an A-B-A-B design, wherein the B conditions both consist of ATDs (rapid alternation of two interventions).
2. Address the inherent limitations of a research design (e.g., multiple baseline design across participants) and to strengthen the demonstration of experimental control by adding another design (e.g., multiple probe design across behaviors).
3. Respond to covariation if behaviors, participant,s or conditions were not independent, as believed prior to the start of a multiple baseline or multiple probe design study, by changing to an A-B-A-B design with concurrent monitoring across untreated behaviors, participants or conditions.

Table 12.2 identifies and briefly summarizes several studies that have combined SCDs for one or more of these reasons. A common practice by researchers interested in the functional analysis of challenging behavior has been to combine an alternating treatments design (ATD—MED variation) with an A-B-A design (Baker, Hanley, & Mathews, 2006; Roantree & Kennedy, 2006) or A-B-A-B withdrawal design (Dwyer-Moore & Dixon, 2007; Hanley, Piazza, Fisher & Maglieri, 2005). Researchers who recognize the limitations of the nonconcurrent multiple baseline design have strengthened their evaluation of experimental control by combining it with an A-B-A-B withdrawal or "reversal" design (Freeman, 2006; Tiger, Hanley, & Hernandez, 2006), multiple baseline design (Schindler & Horner, 2005), or changing criterion design (Najdowski, Wallace, Doney, & Ghezzi, 2003). Others, who recognize the importance of intra-participant replication, and its not being addressed in multiple baseline (or probe) designs across participants, have combined them with multiple probe designs across behaviors (Smith et al., 2016; Trent, Kaiser, & Wolery, 2005), across conditions (Charlop-Christy, Lee, & Freeman, 2000), A-B-A-B withdrawal designs (Koegel, Werner, Vismara, & Koegel, 2005), and changing criterion designs (Levin & Carr, 2001). Multiple baseline across participants have also been combined with A-B-A-B designs (Charlop-Christy and Haymes (1998), alternating treatments designs (ATD; Lloyd, Bateman, Landrum, & Hallahan, 1989), and adapted alternating treatments designs (AATD; Canella-Malone, Sigafoos, O'Reilly, Cruz, Edrisinha, & Lancioni, 2006; Cuvo & Klatt, 1992; Worsdell, Iwata, Dozier, Johnson, Neibert & Thomason, 2005). These studies, and those presented

in Table 12.2, illustrate the range of combination designs that have been used by applied researchers, but this sample is by no means exhaustive. The important thing when designing your study is to select a research design, or combination of designs, that evaluate threats to internal validity and answers the research question(s) posed. As Baer, Wolf, and Risley (1987) stated, "—a good design is one that answers the question convincingly, and as such needs to be constructed in response to the question and then tested through argument in that context (sometimes called 'thinking through') rather than imitated from a book" (p. 319). "Perhaps the more important point is that convincing designs should be more important than 'proper' designs" (p. 320).

Table 12.2 *Studies Using Combination Single Case Research Designs*

Design	Reference	Participants	Setting/ Arrangement	Independent Variable	Dependent Variable
MP across contexts + ABAB	Alberto, P. L., Heflin, J., & Andrews, D. (2002). Use of the timeout ribbon procedure during community-based instruction. *Behavior Modification, 26*, 297–311.	Number: 2 Sex: M Age: 10–11 Disability/ Diagnosis: intellectual disability	Setting: Community work sites & school gym Arrangement: 1:1	Timeout ribbon procedure with token reinforcement for exhibiting appropriate behaviors	Percentage of inappropriate behaviors that interfered with community participation
MP across participants + MP across contexts	Charlop-Christy, M., Le, L., & Freeman, K. A. (2000). A comparison of video modeling with in vivo modeling for teaching children with autism. *Journal of Autism and Developmental Disorders, 30*, 537–552.	Number: 5 Sex: 4 M, 1 F Age: 7–11 years Disability/ Diagnosis: Autism	Setting: Private therapy center Arrangement: 1:1	In-vivo or video modeling for social, adaptive, or cognitive skills	Number of correct responses
MP across contexts+ ABAB	Hughes, M. A., Alberto, P. A., & Fredrick, L. L. (2006). Self-operated auditory prompting systems as a function-based intervention in public community settings. *Journal of Positive Behavior Interventions, 8*, 230–243.	Number: 4 Sex: 2 M, 2 F Age: 16–18 years Disability/ Diagnosis: Moderate ID	Setting: Community work sites Arrangement: 1:1	Self-operated auditory prompting systems with prompts that provided reinforcement for attention or escape-maintained behavior	Escape- and attention-maintained target behaviors (different for each child)
MB across participants + ABAB	Koegel, R. L., Werner, G. A., Vismara, L. A., & Koegel, L. K. (2005). The effectiveness of contextually supported play date interactions between children with autism and typically developing peers. *Research & Practice for Persons with Severe Disabilities, 30*, 93–102.	Number: 2 Sex: 1 M, 1 F Age: 8–9 Disability/ Diagnosis: Autism	Setting: Community/ home play sites Arrangement: Small group	Contextual support provided by providing set-up and structure that encouraged participation by both the child with autism and typically developing peer	Reciprocal interaction between child with autism and typically developing peer; Child affect
MB across participants and changing criterion	Levin, L., & Carr, E. G. (2001). Food selectivity and problem behavior in children with developmental disabilities: Analysis and intervention. *Behavior Modification, 25*, 443–470.	Number: 3 Sex: 2 M, 1 F Age: 5–7 Disability/ diagnosis: Autism (3), Moderate to severe ID (3)	Setting: Self-contained classroom Arrangement: 1:1	Positive reinforcement for consumption in the form of preferred edibles and withholding access to preferred foods prior to meals	Number of bites of non-preferred foods consumed during each session

Guidelines and Considerations for Combining Designs

The decision to combine two SCDs should be made after recognition of the experimental analysis limitations of using one design alone and the advantages of combining two designs. We recommend that you approach your decision to combine designs as follows:

1. Write your rationale for combining designs. This will necessitate identifying the limitations of each individual design you may be considering to answer your research question(s).
2. Select the two simplest research designs that will (a) answer your research question(s), (b) control for threats to internal validity, and (c) be practical given the demands of your setting.
3. Identify the primary design for your study, the one that will "drive" the decision making process. This will typically be the one that you wanted to use in the first place, but in recognition of its limitations, you may decide to add a second design to address those limitations. For example, in recognition of the failure of a multiple baseline (or probe) design across participants to evaluate intra-participant replication, you decide to combine a multiple probe design across participants (N=3) and a multiple probe design across behaviors (N=2). When Participant 1 reaches criterion on the first behavior, you will introduce the independent variable to the second behavior for Participant 1, while concurrently introducing the independent variable to the first behavior for Participant 2. When Participant 2 reaches criterion on the first behavior, you will introduce the independent variable to the second behavior for Participant 2, while concurrently introducing the independent variable to the first behavior for Participant 3. In this example the primary design is the multiple probe design across participants, with the multiple probe design across two behaviors addressing the primary limitation of the multiple probe design across participants (i.e., direct intra-participant replication). Figure 12.8 graphically depicts this combination design.

Consider logistics when choosing designs to combine. If you can combine a multiple probe design across participants with a multiple probe design across behaviors there are practical advantages for doing so. Yes, you could combine multiple baseline designs that would generate continuous rather than intermittent measures prior to introduction of the independent variable, but it is likely to be impractical, particularly if the research is to be conducted in a classroom or clinic setting. Though the frequency of measures will be less, the combined multiple probe designs will still permit an evaluation of history and maturation threats, and testing threats to a lesser degree, and will be more practical to implement in an applied research setting. Follow the design recommendations and guidelines for each of the two designs you have chosen to combine.

When designing your study, whether it is with a single design, or combination of designs, your first priority is to choose a design that will answer your research question(s). This should be done using the simplest research design and data collection procedures that will evaluate potential threats to internal validity. When combining

designs the same principle of parsimony applies. You should “be your worst critic”; anticipate criticisms and adjust your measurement and research design decisions based on the shortcomings you identify. Keep your research design and measurement procedures simple, but not so simple as to fail to adequately address threats to internal validity that would undermine your findings.

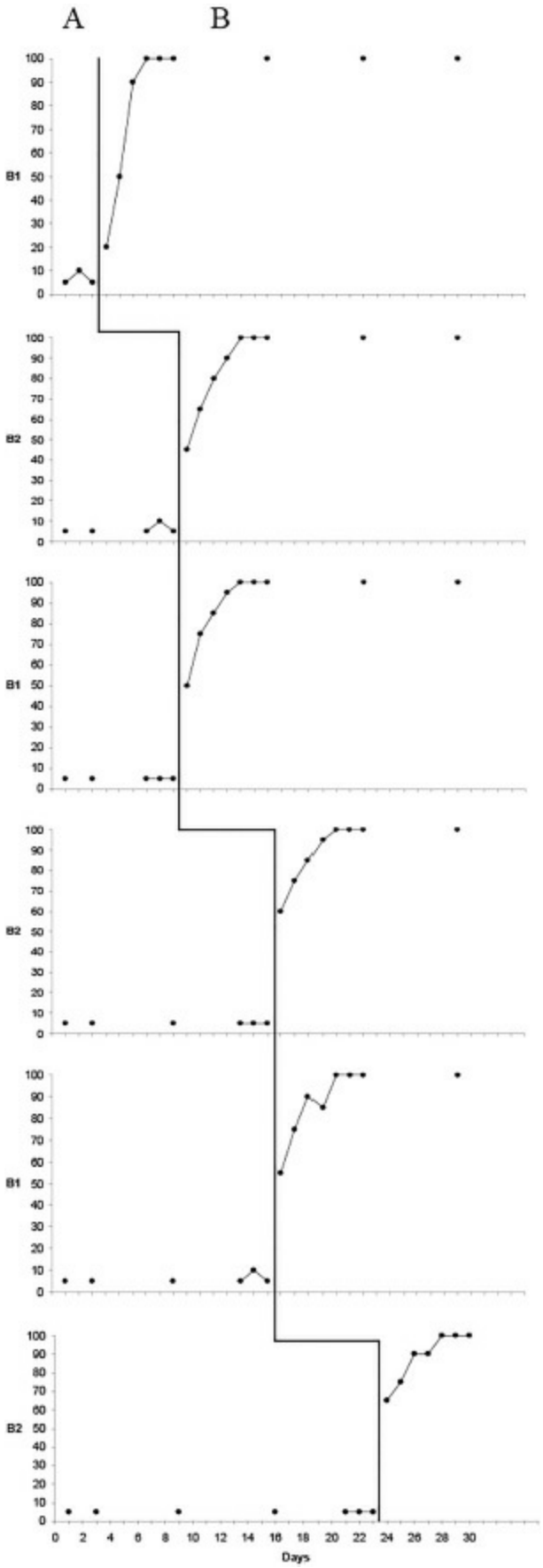


Figure 12.8 Prototype multiple probe design across participants and multiple probe across behaviors.

Applied Example 12–2: MB Design Across Behaviors and Participants

Trent, J. A., Kaiser, A. P., & Wolery, M. (2005). The use of responsive interaction strategies by siblings. *Topics in Early Childhood Special Education*, *25*, 107–118.

In this study, researchers assessed the effectiveness of teaching siblings of children with Down syndrome positive interaction strategies in their homes. Participants were two sibling dyads, each with an older, typically developing sister, and a younger sister with Down syndrome. Sibling dyads were 7 and 5 years, and 9 and 7 years old. Mirroring (motor imitation) and verbal responding were taught to siblings.

Primary data were the number of intervals in which siblings engaged in mirroring (imitation) and responding to verbal turns, using a partial interval recording system. Several secondary variables were also coded, including whether turns were verbal (any vocalization) and topic-related (intelligible to the observer). They also calculated the percentage of times the siblings were responsive to her sister with Down syndrome; this measure was dependent on the number of times the child with Down syndrome initiated.

Reliability data for both dependent (IOA) and independent (PF) variables were collected during 25% of all sessions across all conditions. Using the point-by-point (interval by interval) method, inter-observer agreement (IOA) ranged from 83–100% and procedural fidelity ranged from 89%-100%.

There were three experimental conditions (baseline, intervention, and follow-up). The effects of the intervention were evaluated in the context of a multiple baseline design across behaviors and participants. Sessions were conducted in the family home twice each week, each lasting 30–60 minutes. During baseline sessions, the observer asked siblings to play together for 10 minutes with no other family members in the room. Use of responsive intervention strategies by the typical sibling and verbal turns by the sibling with Down syndrome were recorded using the 10-second partial interval recording system. During intervention, each session was divided into three segments. The first segment consisted of teaching or reviewing interactive response strategies to the typically developing sibling. The second segment consisted of a 10-minute play session with both siblings. During the third segment, both siblings and the observer watched the videotape of the play session while the observer provided positive and corrective feedback. Follow-up sessions, which were conducted identical to baseline sessions, were conducted one month after completion of the intervention condition.

Figure 12.9 displays the number of 10-second intervals in which typical developing siblings used a responsive interactive strategy during baseline,

intervention, and follow-up conditions evaluated within the context of a multiple baseline across behaviors and participants design. Prior to instruction, there were few intervals during which they used responsive interaction strategies when interacting with their sister with Down syndrome. A visual analysis of the data shows that the use of responsive interactive strategies changed from a stable, low level, and zero or decelerating trend in baseline conditions, to a variable accelerating trend in a therapeutic direction upon introduction of training. These data were replicated across siblings. During the one-month follow-up session, siblings maintained use of responsive interaction strategies above baseline levels.

The researchers also measured responsiveness to verbal turns in a multiple baseline across participants design; though with only two tiers (data not presented here). For both participants, responsiveness did not increase during the mirroring condition but did increase during the verbal responding condition. Data on behaviors of the siblings with Down syndrome were more variable. Number of verbal turns remained variable across all conditions of the study. Although typical siblings took more verbal turns than their sibling with Down syndrome throughout the intervention, the ratio of turns taken by the typical developing sibling and the sibling with Down syndrome was more balanced at the end of intervention condition compared to baseline condition.

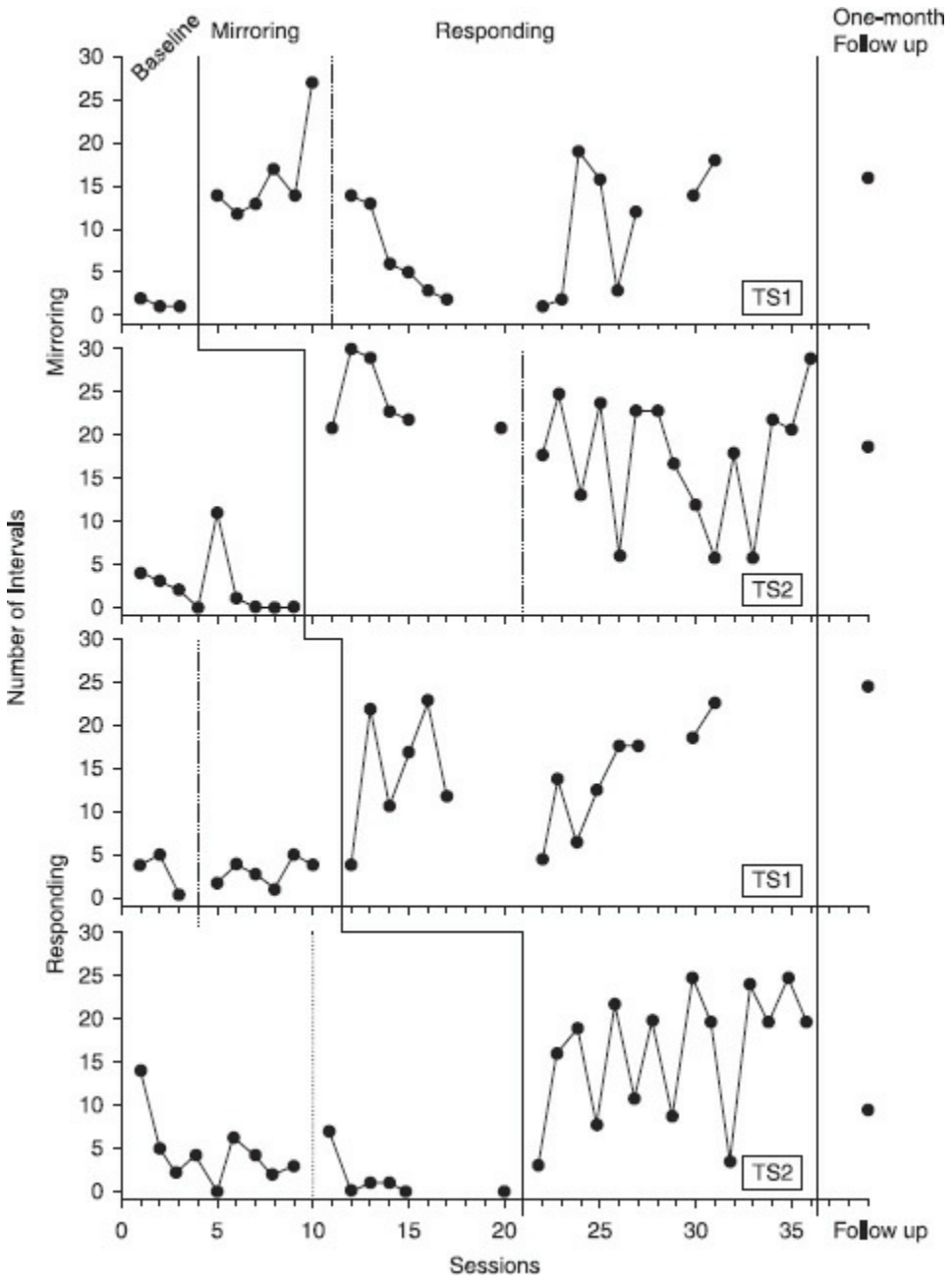


Figure 12.9 Multiple baseline across behaviors and participants design.

Source: Trent, J. A., Kaiser, A. P., & Wolery, M. (2005). The use of responsive interaction strategies by siblings. *Topics in Early Childhood Special Education*, *25*, 107–118.

Applied Example 12 – 3: ATD and A-B-A-B Withdrawal Design

Dwyer-Moore, K. J., & Dixon, M. R. (2007). Functional analysis and treatment of problem behavior of elderly adults in long-term care. *Journal of Applied Behavior Analysis*, *40*, 679–683.

Dwyer-Moore and Dixon (2007) studied the use of functional analysis and function-based treatments in three elderly participants in a long-term care facility. Participants in the study all were diagnosed with dementia, ranged in age from 70 to 90 years old, and referred by administrative and nursing staff for exhibiting problematic behaviors. Two female participants were referred due to disruptive vocalizations (obscenities, repetitive statements, "irrelevant utterances"), and one male participant was referred due to wandering from the facility that made it difficult to ensure his safety within the home. The functional analysis and results of treatment were evaluated within the context of an ATD (MED variation) combined within an A-B-A-B withdrawal design. The dependent variable was number of responses per minute for the target undesirable behavior. Interobserver agreement (IOA) was calculated by dividing the smaller frequency count by the larger frequency count and multiplying by 100. Agreement ranged from 90%–100% for functional analysis sessions (M=94%) and from 92–100% for treatment sessions (M=97%).

During the functional analysis, four experimental conditions (attention, demand, control, alone) were conducted for 10 minutes each with 5-minute breaks between sessions. During the attention condition, the experimenter sat across the room and interacted with the participant only by giving 5–10 seconds of social attention when the problem behavior was exhibited. During the demand condition, an occurrence of the problem behavior resulted in removal of the demand (gross motor or academic tasks) for 30 seconds. Leisure items were readily available in the control condition, and the experimenter provided 5–10 seconds of social attention during each 30-second interval. In the alone condition, the experimenter observed "unobtrusively" through a gap in the door; no leisure items or social attention were available. Results of the functional analysis (ATD, MED variation) portion of the study showed that the target behaviors were maintained by attention for two participants (Alice and Carmen), and escape from demands for the third participant (Derek).

Intervention sessions were conducted individually for 10 minutes each. For Alice, a DRA procedure was used that consisted of 3–5 seconds of social attention contingent upon an appropriate vocalization. Inappropriate vocalizations were ignored. Derek's intervention consisted of noncontingent access to attention (NCA)

on a FT-30 second schedule and access to his most preferred leisure items, identified via a preference assessment. No consequences were provided for wandering. If he wandered out of the room he was redirected back to the family room once he was observed engaged in appropriate behavior. Carmen's intervention entailed FCT with extinction during demand situations. Demands were presented continuously and the experimenter prompted her to hand a break card to her. Prompted or unprompted presentation of the break card resulted in a 30-second break from the activity. Inappropriate vocalizations resulted in continued demands, with a prompt to use the card after 5 seconds with no vocalizations.

Figure 12.10 shows the results from the functional analysis and treatment for each participant. For Alice, the ATD showed that inappropriate vocalizations were maintained by attention. Treatment, using DRA for appropriate vocalizations, resulted in a 40% decrease in inappropriate vocalizations, and a 400% increase in appropriate vocalizations. The effectiveness of the intervention was demonstrated within the context of the A-B-A-B withdrawal design. Derek's wandering behavior also was maintained by attention and his treatment package of noncontingent attention and access to preferred leisure activities resulted in an 85% decrease in his wandering behavior. The functional analysis showed that Carmen's disruptive vocalizations were maintained by escape from demands. Functional communication training and extinction resulted in an 82% decrease in inappropriate vocalizations from the baseline condition to the intervention condition.

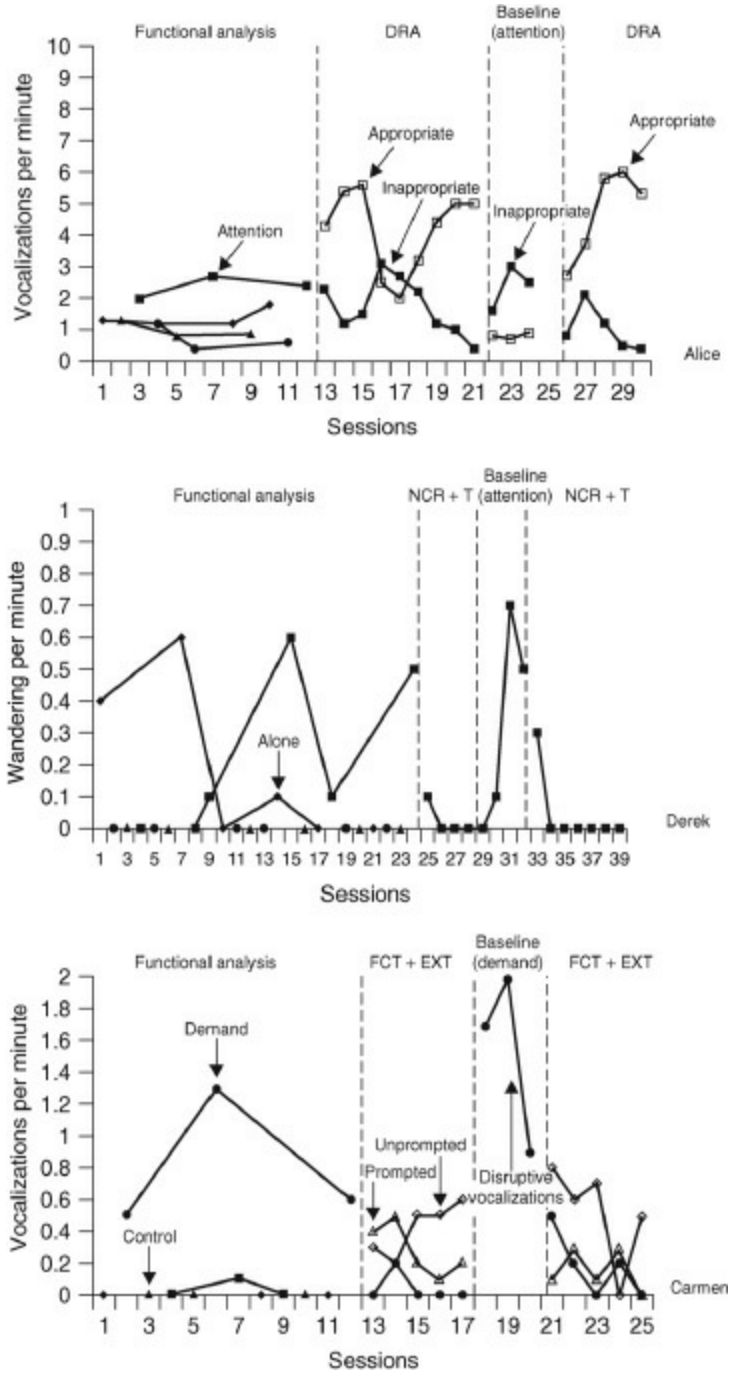


Figure 12.10 Combination A-B-A-B plus ATD.

Source: Dwyer-Moore, K. J., & Dixon, M. R. (2007). Functional analysis and treatment of problem behavior of elderly adults in long-term care. *Journal of Applied Behavior Analysis*, *40*, 679–683.

Summary

In this chapter we reviewed some SCDs that are used less often than the major designs covered in the preceding chapters. In addition, we provided an introduction to the rationale for combining designs in SCD research. Increasingly applied researchers are combining SCDs as a means of answering more than one research question in their investigations (e.g., ATD to assess the function of a behavior and an A-B-A-B to evaluate intervention effectiveness), and addressing the limitations of some designs, such as no intra-participant replication (e.g., multiple probe across participants combined with a multiple probe across behaviors). Because of the dynamic nature of SCDs, researchers have been able to add, or change, their experimental design midcourse for one or more study participants in an effort to salvage experimental control in response to behavior covariation. In our discussion of combining SCDs we have advocated that the simplest combination of designs be employed, provided that the research question can be answered and threats to internal validity controlled for. Prior to design selection we recommend that you assume the role of critic, systematically identifying all potential threats to internal validity and criticisms from reviewers that may arise. By doing so you will be better prepared to justify your design choice and explain its strengths and weaknesses in evaluating experimental control.

References

Alberto, P. L., Heflin, J., & Andrews, D. (2002). Use of the timeout ribbon procedure during community-based instruction. *Behavior Modification*, *26*, 297–311.

Baer, D. M., Wolf, M. M., & Risley, T. R. (1987). Some still-current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *20*, 313–327.

Baker, J. C., Hanley, G. P., & Mathews, R. M. (2006). Staff-administered functional analysis and treatment of aggression by an elder with dementia. *Journal of Applied Behavior Analysis*, *39*, 469–474.

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

Bickford, J. O., & Falco, R. A. (2012). Technology for early braille literacy: Comparison of traditional braille instruction and instruction with an electronic notetaker. *Journal of Visual Impairment & Blindness*, *106*, 679–693.

Birkan, B., Krantz, P. J., & McClannahan, L. E. (2011). Teaching children with autism spectrum disorders to cooperate with injections. *Research in Autism Spectrum Disorders*, *5*, 941–948.

Bouck, E. C., Flanagan, S., Joshi, G. S., Sheikh, W., & Schleppenbach, D. (2011). Speaking math—A voice input, speech output calculator for students with visual impairments. *Journal of Special Education Technology*, *26*, 1–14.

Butler, C., Brown, J. A., & Woods, J. J. (2014). Teaching at-risk toddlers new vocabulary using interactive digital storybooks. *Contemporary Issues in Communication Science and Disorders*, *41*, 155–168.

Canella-Malone, H., Sigafoos, J., O'Reilly, M., de la Cruz, B., Edrisinha, C., & Lancioni, G. E. (2006). Comparing video prompting to video modeling for teaching daily living skills to six adults with developmental disabilities. *Education and Training in Developmental Disabilities*, *41*, 344–356.

Charlop-Christy, M. H., & Haymes, L. K. (1998). Using objects of obsession as token reinforcers for children with autism. *Journal of Autism and Developmental Disorders*, *28*, 189–198.

Charlop-Christy, M. H., Le, L., & Freeman, K. A. (2000). A comparison of video modeling with in vivo modeling for teaching children with autism. *Journal of Autism and Developmental Disorders*, *30*, 537–552.

Cihak, D., Alberto, P. A., & Fredrick, L. D. (2007). Use of brief functional analysis and intervention evaluation in public settings. *Journal of Positive Behavior Interventions*, *9*, 80–93.

Cooper, L. J., Wacker, D. P., Sasso, G. M., Reimers, T. M., & Donn, L. K. (1990). Using parents as therapists to evaluation appropriate behavior of their children: Application to a tertiary diagnostic clinic. *Journal of Applied Behavior Analysis*, *23*, 285–296.

Cuvo, A. J., & Klatt, K. P. (1992). Effects of community-based, videotape, and flash card instruction of community- referenced sight words on students with mental retardation. *Journal of Applied Behavior Analysis*, *25*, 499–512.

DeLuca, R. V., & Holborn, S. W. (1992). Effects of variable-ratio reinforcement schedule with changing criterion on exercise in obese and non-obese boys. *Journal of Applied Behavior Analysis*, *25*, 671–679.

Dufrene, B. A., Watson, T. S., & Kazmerski, J. S. (2008). Functional analysis and treatment of nail biting. *Behavior Modification*, *32*, 913–927.

Duncan, H. J., & Silberberg, A. (1982). The effects of concurrent responding and reinforcement on behavioral output. *Journal of the Experimental Analysis of Behavior*, *38*, 125–132.

Dwyer-Moore, K. J., & Dixon, M. R. (2007). Functional analysis and treatment of problem behavior of elderly adults in long-term care. *Journal of Applied Behavior Analysis*, *40*, 679–683.

Easterbrooks, S. R., & Stoner, M. (2006). Using a visual tool to increase adjectives in the written language of students who are deaf or hard of hearing. *Communication Disorders Quarterly*, *27*, 95–109.

Flood, W. A., & Wilder, D. A. (2004). The use of differential reinforcement and fading to increase time away from a caregiver in a child with separation anxiety disorder. *Education and Treatment of Children*, *27*, 1–8.

Freeman, K. A. (2006). Treating bedtime resistance with the bedtime pass: A systematic replication and component analysis with 3-year olds. *Journal of Applied Behavior Analysis*, *39*, 423–428.

Ganz, J. B., & Sigafoos, J. (2005). Self-monitoring: Are young adults with MR and autism able to utilize cognitive strategies independently? *Education and Training in Developmental Disabilities*, *40*, 24–33.

Geiger, K. B., LeBlanc, L. A., Dillon, C. M., & Bates, S. L. (2010). An evaluation of preference for video and in-vivo modeling. *Journal of Applied Behavior Analysis*, *43*, 279–283.

Grey, I., Healy, O., Leader, G., & Hayes, D. (2009). Using a Time Timer to increase appropriate waiting behavior in a child with developmental disabilities. *Research in Developmental Disabilities*, *30*, 359–366.

Hackenberg, T. D., & Joker, V. R. (1994). Instructional versus schedule control of humans' choices in situations of diminishing returns. *Journal of the Experimental Analysis of Behavior*, *62*, 367–383.

Hall, R. V. (1971). *Managing behavior: Behavior modification, the measurement of behavior*. Lawrence, KS: H & H Enterprises.

Hanley, G. P., Piazza, C. C., Fisher, W. W., Contrucci, S. A., & Maglieri, K. A. (1997). Evaluation of client preference for function-based treatment packages. *Journal of Applied Behavior Analysis*, *30*, 459–473.

Hanley, G. P., Piazza, C. C., Fisher, W. W., & Maglieri, K. A. (2005). On the effectiveness of and preference for punishment and extinction components of function-based

interventions. *Journal of Applied Behavior Analysis*, *38*, 51–65.

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

Heal, N. A., & Hanley, G. P. (2011). Embedded prompting may function as embedded punishment: Detection of unexplained behavioral processes within a typical preschool teaching strategy. *Journal of Applied Behavior Analysis*, *44*, 127–131.

Hughes, M. A., Alberto, P. A., & Fredrick, L. L. (2006). Self-operated auditory prompting systems as a function- based intervention in public community settings. *Journal of Positive Behavior Interventions*, *8*, 230–243.

Johnston, R. J., & McLaughlin, T. F. (1982) The effects of free time on assignment completion and accuracy in arithmetic: A case study. *Education and Treatment of Children*, *5*, 35–40.

Kennedy, C. H. (2005). *Single-case designs for educational research*. Boston, MA: Pearson/Allyn and Bacon.

Klein, L. A., Houlihan, D., Vincent, J. L., & Panahon, C. J. (2017). Best practices in utilizing the changing criterion design. *Behavior Analysis in Practice*, *10*, 52–61.

Koegel, R. L., Openden, D., & Koegel, L. K. (2004). A systematic desensitization paradigm to treat hypersensitivity to auditory stimuli in children with autism in family contexts. *Research & Practice for Persons With Severe Disabilities*, *29*, 122–134.

Koegel, R. L., Werner, G. A., Vismara, L. A., & Koegel, L. K. (2005). The effectiveness of contextually supported play date interactions between children with autism and typically developing peers. *Research & Practice for Persons With Severe Disabilities*, *30*, 93–102.

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.

LeGray, M. W., Dufrene, B. A., Sterling-Turner, H., Olmi, D. J., & Bellone, K. (2010). A comparison of function- based differential reinforcement interventions for children engaging in disruptive classroom behavior. *Journal of Behavioral Education*, *19*, 185–204.

Levin, L., & Carr, E. G. (2001). Food selectivity and problem behavior in children with developmental disabilities: Analysis and intervention. *Behavior Modification*, *25*, 443–470.

Lloyd, J. W., Bateman, D. F., Landrum, T. J., & Hallahan, D. P. (1989). Self-recording of attention versus productivity. *Journal of Applied Behavior Analysis*, *22*, 315–323.

Luiselli, J. K. (2000). Cueing, demand fading, and positive reinforcement to establish self-feeding and oral consumption in a child with food refusal. *Behavior Modification*, *24*, 348–358.

Malanga, P. R., & Sweeney, W. J. (2008). Increasing active student responding in a university applied behavior analysis course: The effect of daily assessment and response cards on end of week quiz scores. *Journal of Behavioral Education*, *17*, 187–199.

Martens, B. K., Eckert, T. L., Bradley, T. A., & Ardoin, S. P. (1999). Identifying effective treatments from a brief experimental analysis: Using single-case design elements to aid decision making. *School Psychology Quarterly*, *14*, 163–181.

Martens, B. K., & Gertz, L. E. (2009). Brief experimental analysis: A decision tool for bridging the gap between research and practice. *Journal of Behavioral Education*, *18*, 92–99. McComas, J. J., Wacker, D. P., Cooper, L. J., Asmus, J. M., Richman, D., & Stoner, B. (1996). Brief experimental analysis of stimulus prompts for accurate responding on academic tasks in an outpatient clinic. *Journal of Applied Behavior Analysis*, *29*, 397–401.

McComas, J. J., Wagner, D., Chaffin, M. C., Holton, E., McDonnel, M., & Monn, E. (2009). Prescriptive analysis: Further individualization of hypothesis testing in brief experimental analysis of reading fluency. *Journal of Behavioral Education*, *18*, 56–70.

McDougall, D. (2006). The distributed criterion design. *Journal of Behavioral Education*, *15*, 237–247.

Mong, M. D., & Mong, K. W. (2012). The utility of brief experimental analysis and extended intervention analysis in selecting effective mathematics interventions. *Journal of Behavioral Education*, *21*, 99–118.

Najdowski, A. C., Wallace, M. D., Doney, J. K., & Ghezzi, P. M. (2003). Parental assessment and treatment of food selectivity in natural settings. Journal *of Applied Behavior Analysis*, *36*, 383–386.

Nes, S. L. (2005). Using paired reading to enhance the fluency skills of less-skilled readers. *Reading Improvement*, *40*, 179–192.

O'Connor, A. S., Prieto, J., Hoffmann, B., DeQuinzio, J. A., & Taylor, B. A. (2011). A stimulus control procedure to decrease motor and vocal stereotypy. *Behavioral Interventions*, *26*, 231–242.

Petursdottir, A., McMaster, K., McComas, J. J., Bradfield, T., Braganza, V., McDonald, J., Rodriguez, R., & Scharf, H. (2009). Brief experimental analysis of early reading interventions. *Journal of School Psychology*, *47*, 215–243.

Rapp, J. T., Cook, J. L., McHugh, C., & Mann, K. R. (2017). Decreasing stereotypy using NCR and DRO with functionally matched stimulation: Effects on targeted and non-Targeted Stereotypy. *Behavior Modification*, *41*, 45–83.

Ricciardi, J. N., Luiselli, J. K., & Camre, M. (2006). Shaping approach responses as intervention for specific phobia in a child with autism. *Journal of Applied Behavior Analysis*, *39*, 445–448.

Richardson, W. K., & Clark, D. B. (1976). A comparison of the key-peck and treadle-press in the pigeon: Differential reinforcement of low-rate schedule of reinforcement. *Journal of the Experimental Analysis of Behavior*, *26*, 237–256.

Roantree, C. F., & Kennedy, C. H. (2006). A paradoxical effect of presession attention on stereotypy: Antecedent attention as an establishing, not an abolishing, operation. *Journal of Applied Behavior Analysis*, *39*, 381–384.

Schindler, H. R., & Horner, R. H. (2005). Generalized reduction of problem behavior of young children with autism: Building trans-situational interventions. *American*

Journal on Mental Retardation, *110*, 36–47.

Schmidt, A. C., Hanley, G. P., & Layer, S. A. (2009). A further analysis of the value of choice: Controlling for illusory discriminative stimuli and evaluating effects of less preferred items. *Journal of Applied Behavior Analysis*, *42*, 711–716.

Schumacher, B. I., & Rapp, J. T. (2011). Increasing compliance with haircuts for a child with autism. *Behavioral Interventions*, *26*, 67–75.

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

Slocum, S. K., & Tiger, J. H. (2011). An assessment of the efficiency of and choice for forward and backward chaining. *Journal of Applied Behavior Analysis*, *44*, 793–805.

Smith, K. A., Ayres, K. A., Alexander, J., Ledford, J. R., Shepley, C., & Shepley, S. B. (2016). Initiation and generalization of self-instructional skills in adolescents with autism and intellectual disability. *Journal of Autism and Developmental Disabilities*, *46*, 1196–1209.

Spencer, E. J., Goldstein, H., Sherman, A., Noe, S., Tabbah, R., Ziolkowski, R., & Schneider, N. (2012). Effects of an automated vocabulary and comprehension intervention: An early efficacy study. *Journal of Early Intervention*, 45, 195–221.

Sullivan, M., Konrad, M., Joseph, L. M., & Luu, K. C. T. (2013). A comparison of two sight word reading fluency drill formats. *Preventing School Failure*, *57*, 102–110.

Tiger, J. H., Hanley, G. P., & Hernandez, E. (2006). An evaluation of the value of choice with preschool children. *Journal of Applied Behavior Analysis*, *39*, 1–16.

Trent, J. A., Kaiser, A. P., & Wolery, M. (2005). The use of responsive interaction strategies by siblings. *Topics in Early Childhood Special Education*, *25*, 107–118.

Tullis, C. A., Cannella-Malone, H. I., & Fleming, C. V. (2012). Multiple stimulus without replacement preference assessments: An examination of the relation between session number and effectiveness. *Journal of Developmental and Physical Disabilities*, *24*, 337–345.

Ward, R. D., & Higbee, T. S. (2008). Noncontingent reinforcement for tub-standing in a toddler. *Education and Treatment of Children*, *31*, 213–222.

Warnes, E., & Allen, K. D. (2005). Biofeedback treatment of paradoxical vocal fold motion and respiratory distress in an adolescent girl. *Journal of Applied Behavior Analysis*, *38*, 529–532.

Worsdell, A. S., Iwata, B. A., Dozier, C. L., Johnson, A. D., Neibert, P. L., & Thomason, J. L. (2005). Analysis of response repetition as an error-correction strategy during sight-word reading. *Journal of Applied Behavior Analysis*, *38*, 511–527.

13
Evaluating Quality and Rigor in Single Case Research

Jennifer R. Ledford, Justin D. Lane, and Robyn Tate

Important Terms

risk of bias, rigor, internally valid, quality, standards, quality indicators, rating frameworks, ecological validity, social validity, stimulus generalization, response generalization, maintenance

To this point, we have primarily discussed the *conduct* of single case design (SCD) research. Of course, both researchers and practitioners must also *analyze* the research studies of others in order to contextualize their own studies and determine the extent to which evidence exists for a given practice in given contexts (What works, for whom, and under what conditions?). In the next two chapters, we will suggest that *rigor* of evidence should be assessed separately from *outcomes*. Although outcomes might seem more

immediately relevant (especially for researcher- practitioners), we argue that outcomes assessment is ultimately meaningless without an accompanying meaningful analysis of the rigor of individual studies and the body of literature as a whole. Thus, in this chapter, we focus on assessing rigor of SCD studies. In the next chapter, we discuss more thoroughly the rationale and methods behind synthesis of SCD research, as well as analyzing outcomes across studies.

Rigor, Risk of Bias, Internal Validity, and Quality

The terms rigor, risk of bias, and internal validity have been used interchangeably to mean *to what extent are we confident that the results of a study are due to planned differences between conditions, and not to any other factors?* **Risk of bias** is a term used in group comparison research and refers to the likelihood that the outcomes of a study are *biased* due to some methodological decision made by the researchers, resulting in potential overestimation of effects (Higgins et al., 2011). For example, in both group comparison and SCD research, observer bias is possible, and is minimized by using blind observers (often referred to as blind assessors). **Rigor** is a more comprehensive term; it refers to the extent to which researchers planned and conducted the study in a manner that produces convincing outcomes. Rigor encompasses minimizing bias as well as other factors that decrease confidence in outcomes, such as having insufficient data to draw confident conclusions and choosing an inappropriate design to answer your research questions. A study with low risk of bias and high rigor is **internally valid**. Risk of bias, rigor, and internal validity are not rated dichotomously; rather they are generally considered to be low, adequate, or high. **Quality** is another term used in evaluating research; it generally refers to whether the study includes components that are considered to be important for generality or applicability; this term tends to be more domain-specific than the other terms. For example, it might be important that educational interventions are tested in typical school contexts, but this would not be relevant for experimental psychology. The ideal study has high quality, low risk of bias, and high rigor and internal validity. Below, we have divided characteristics of studies we determine to be *critical*—a study cannot be considered rigorous without these; and those we determine to be *important* for improving quality—via increasing applicability, importance, or generality—but less critical for ensuring adequate internal validity.

Critical Characteristics of SCD Studies

We identify five critical elements that are crucial for establishing adequate rigor. These include using an appropriate design to answer your research questions, having adequate opportunities to demonstrate effects, demonstrating adequate reliability of dependent and independent variables, and collecting a sufficient amount of data. These characteristics are generally considered important, but no consensus exists regarding whether each is critical (cf. Hitchcock et al., 2014; Wolery, 2013).

Design Appropriateness

Confident conclusions regarding functional relations can only be drawn when an appropriate design is used to answer your research questions. Basic study characteristics (e.g., behavior reversibility and research question type) should be used to narrow down design options (see Appendix 13.1). More nuanced decisions can be made among potentially appropriate design types by considering factors such as feasibility, likely threats to validity, and available participants, contexts, and behaviors. Examples of common inappropriate use of SCDs for answering research questions include: (a) measuring non-reversible behaviors in the context of a design intended to be used with reversible behaviors; (b) drawing conclusions relative to baseline conditions without adequate replications (e.g., in ATDs without a continuing baseline condition or in A-B-C-B-C designs); and (c) failure to include a control set when using AATDs, to control for history or maturation threats. Appendix 13.1 provides a structure for determining appropriate research designs based on the question type (demonstration or comparison) and the dependent variable type (reversible or non-reversible); it can be used both to determine what design you should use when conducting a study and to determine whether another researcher has used the most appropriate design.

Potential Demonstrations of Effect

If authors chose an appropriate design to answer their research questions, the next step is to determine whether they have included a sufficient number of demonstrations to adequately control for threats to internal validity allowing for an experimental demonstration of effect. Generally, three potential demonstrations at three different points in time is sufficient; common variations that are *insufficient* include multiple baseline or probe designs with two tiers and withdrawal or multitreatment designs with insufficient *adjacent conditions* (e.g., A-B-A, A-B-C-D-C, A-B-A-C). Note that multiple baseline designs with three or more tiers but fewer than three start points have an insufficient number of potential demonstrations and all non- concurrent multiple

baseline designs have insufficient rigor for drawing confident conclusions (see Chapter 10). Likewise, multiple baseline designs that have different interventions following a concurrent baseline are not appropriate (e.g., A-B for two tiers, and A-C in the third tier). If at least three potential demonstrations are not possible, at three different points in time, the study does not have adequate internal validity. We will note that these guidelines are relatively new in the field; thus, many studies conducted prior to the 2000s may not meet these and other rigor guidelines.

Reliability

In Chapter 5, we discussed the potential hazards of human observers, including error, bias, and drift. Here, we would like to reiterate that studies should include secondary observers, who are blind to condition if possible (Chazin et al., 2017; Tate et al., 2014). Secondary observers should be independent (e.g., not influenced by responding of the primary observer) and should collect data alongside the primary observer for approximately 1/3 of all sessions in all conditions for all participants (Ledford, Barton, Severini, & Zimmerman, 2017). We should note that others have recommended 20–25% of sessions (e.g., What Works Clearinghouse, 2013). Determining minimum frequency may vary based on code complexity and ongoing agreement (Kazdin, 2011; e.g., when low agreement occurs in a particular condition, additional agreement data should be collected). In addition, interobserver agreement should be calculated using the most precise method given the recording system (e.g., point-by-point if possible). Means and ranges across participants and conditions should be reported, and reasons for any low values should be described. Preferably, authors should indicate that they visually analyzed secondary data to assess for potential bias and drift, and should report procedures for retraining and discrepancy discussions. Many tools (see below) report that an 80% agreement value is acceptable; as outlined in Chapter 5, this is a somewhat arbitrary criterion. When determining acceptability, you should assess the complexity of the codes and contexts, the relative subjectivity of the dependent variable, and (most importantly) whether low agreement has the potential to alter data patterns (and thus, your decisions regarding functional relations; Barlow & Hersen, 1984). If reliability data are not collected and reported at a sufficiently high level for all dependent variables and participants, to permit confidence of changes between conditions, the study does not have adequate internal validity.

Reliability of the Independent Variable

We also consider reliability of independent variable implementation a critical factor; without confirmation that all conditions were conducted as planned, we cannot be confident that programmed changes between conditions occurred (Ledford & Gast, 2014; Ledford & Wolery, 2013). This, of course, precludes confidence that differences between

conditions resulted in changes in participant behavior. As discussed in Chapter 6, researchers should provide evidence that *all* conditions were implemented as intended, not just treatment conditions (Ledford & Gast, 2014; Ledford & Wolery, 2013). There are no consistent criterion levels defined as adequate, but implementation should exceed 90%, unless your research questions are related to fidelity levels (e.g., if you intend to answer the question of whether certain conditions lead to high-fidelity use of interventions). Low procedural fidelity can be mitigated by re-training (see Chapter 6). Reasons for low fidelity should be described and implications of low fidelity should be explicitly stated (e.g., if positive outcomes occurred despite intermittent low fidelity, this might serve as evidence that the intervention is powerful even if practitioners cannot complete it to 100% fidelity all of the time). If fidelity data across comparison conditions (e.g., baseline and intervention or Intervention 1 and Intervention 2) are not collected, reported, and at a sufficiently high level, the study does not have adequate internal validity.

Data Sufficiency

The final characteristic we will define as critical is the presence of a sufficient amount of data for determining whether a functional relation is present. Different minimum criteria have been specified (e.g., three versus five; CEC, 2014; WWC, 2013). However, rather than identifying a specific number, we suggest that analysts answer the question: Does the number of data points in one or more conditions prohibit or seriously inhibit the ability to identify (a) whether behavior change occurred, and (b) whether these changes were due to changes between conditions *and only* changes between conditions? If the answer to this question is "yes," the study does not have adequate internal validity. Generally, you need few data points if data are at floor or ceiling levels and stable; you need more data points if data are variable.

Characteristics that Increase Quality

While some characteristics are necessary for drawing confident conclusions regarding the relation between independent and dependent variables, others are desirable but not critical. These characteristics increase applicability, importance, or generality of study findings, but do not directly influence your ability to draw confident conclusions about results. Four such factors are ecological validity, social validity, generalization assessment, and maintenance assessment. Other factors related to *written descriptions* of your study are detailed below (see *Adequate Reporting*).

Ecological Validity

Ecological validity refers to the extent to which a study is relevant to typical contexts. Of course, the referent for and importance of *typical context* may be different depending on your interests (Bronfenbrenner, 1979). For example, a practicing speech pathologist might find relevant a series of studies conducted in clinics related to the use of responsive interaction practices, conducted by graduate students who are licensed special educators or speech pathologists. However, this group of studies may have low ecological validity for a preschool teacher who is interested in improving responsive interactions in her classroom of 18 children with and without disabilities. Thus, ecological validity depends on your research question. Two components of ecological validity you should assess are the extent to which (a) *contexts* in a study are similar to the typical environment and (b) *implementers* are similar to the ones who typically implement the intervention. In addition, some researchers have suggested assessment of typical contexts be divided into *physical* (e.g., inclusive classroom where a child typically attended school), *activity* (e.g., an activity in which the child would regularly participate, not a contrived situation), and *social* contexts (e.g., with others typically present in the environment; Clarke & Dunlap, 2008).

A study can have high internal validity and have little ecological validity; a study can also have high ecological validity without internal validity. In the latter case, despite high ecological validity, the study is not useful because we cannot draw confident conclusions—or as Bronfenbrenner (1979) said: "I question the seemingly automatic granting of scientific legitimacy to a research effort merely because it is conducted in a real-life setting" (p. 28–29). Nonetheless, studies that are both ecologically and internally valid may contribute more applicable and generalizable knowledge to the field, reducing the research-to-practice gap (Ledford, Hall, Conder, & Lane, 2016). When assessing groups of studies for the purpose of synthesizing findings, you should quantify ecological validity by analyzing the extent to which the group of studies use typical contexts and indigenous implementers, and explicitly address to what extent this might impact

generality of findings.

Social Validity

Social validity is the relative social importance of a study's procedures, goals, and outcomes (Wolf, 1978); it is closely related to ecological validity (Foster & Mash, 1999). Social validity has largely targeted indirect stakeholders (e.g., parents, practitioners, employers of participants), especially in the area of special education and behavioral sciences; these assessments of approval of the goals, procedures, or outcomes have tended to be questionnaires related to intervention acceptability. This is probably the *easiest* way to measure social validity, although some have questioned the usefulness of these assessments, given they are nearly always positive in nature (cf. Ledford et al., 2016; Machalicek et al., 2008). Other ways to measure social validity, discussed in Chapter 6, are less subject to bias. They include: (a) normative comparisons, (b) blind raters, and (c) provision of intervention choice (Hanley, 2010). In addition to these assessments, some have suggested that maintenance of behavior change in the absence of intervention serves as meaningful evidence of social validity (Kennedy, 2003). When assessing groups of studies for the purpose of synthesizing findings, you should assess to what extent studies measured social validity, and to what extent those measurements included questionnaires that are more subject to bias, or less subjective measures.

Generalization Assessment

In the areas of special education and behavioral sciences, measuring and improving generalized behavior change has received considerable attention (cf. Council for Exceptional Children, 2015; Wolery & Gast, 1984) and, in at least some areas of research, generalization is consistently assessed (cf. Ledford, Lane, Elam, & Wolery, 2012). Two kinds of generalization can be assessed: stimulus generalization and response generalization. **Stimulus generalization** refers to the use of a taught behavior in the presence of different contexts (e.g., varied materials, different instructor, different setting). For example, a participant who is taught to read sight words on notecards might generalize this skill to reading the words in books (generalization across materials), or a participant who is taught to self-monitor attention in math class might generalize this skill to reading class (generalization across settings). **Response generalization** refers to the use of a similar and related behavior that occurs without direct teaching. For example, a participant who is taught to verbally respond to peers might generalize this behavior and begin to *initiate* interactions, or a student who is taught to use a number line to add numerals might begin to use the number line to appropriately subtract numbers as well. When assessing the extent to which studies assess generalization, two questions are appropriate: (1) What types of generalization would be desirable outside of study contexts? and (2) How well was generalization measured? In some areas, stimulus

generalization may be of considerable importance (this type of generalization has been measured most often in SCD research); however, in other areas, response generalization may critical. Figure 13.1 shows a hierarchy of generalization measurement in SCD research; generally, post-test only generalization assessments are the least desirable and continuous measurement in the context of an SCD is the most desirable.

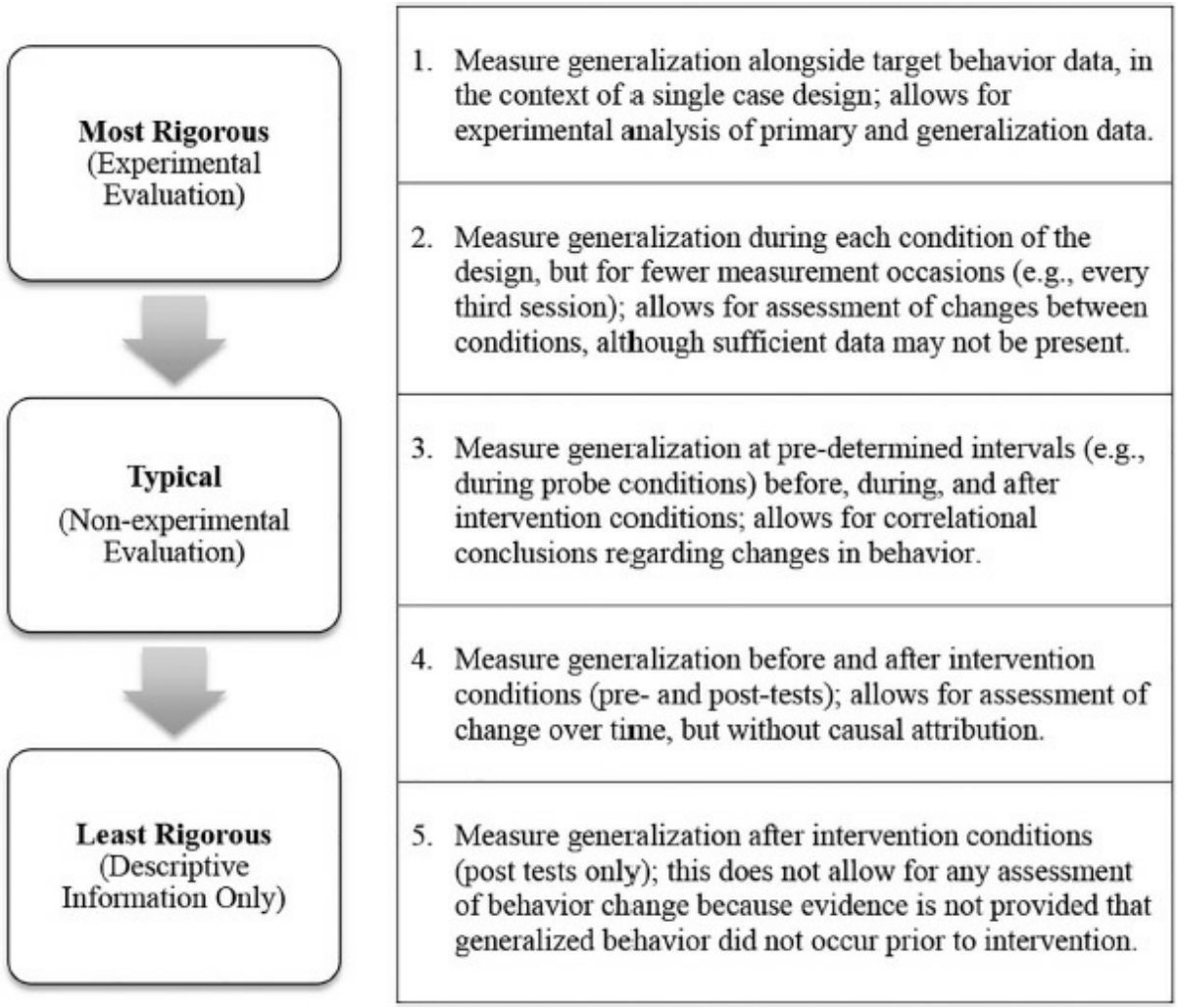


Figure 13.1 Descriptions of common temporal procedures for measuring generalization.

Maintenance Assessment

Maintenance refers to the continued effects of an intervention (e.g., increased accuracy, decreased problem behavior) in the absence of intervention. SCD research is generally not well-suited to experimentally evaluate maintenance of behavior change, although it is possible. For example, a multiple baseline across participants design, in which the first experimental evaluation is between baseline and intervention conditions, can also time lag a maintenance condition, allowing you to evaluate maintenance in relation to the intervention condition. However, this is rare and generally not the primary question of interest. The withdrawal of intervention in A-B-A-B designs serves as a type of

maintenance measure, but for behaviors that we do not *expect* to maintain. When assessing the extent to which studies adequately assess maintenance, you may want to ask three relevant questions. (1) Do I expect this behavior to maintain? Many behaviors (particularly those that are reversible) may not be expected to maintain in the absence of intervention. (2) If I expect maintenance, is it measured? (3) Is experimental control demonstrated for maintained outcomes? The answer to this final question is rarely "yes"; this is not a critical flaw, but you should recognize that conclusions drawn from non-experimental comparisons should be considered less compelling.

Randomization and Rigor

Some researchers have suggested that randomization should be liberally used in SCD research (Kratochwill & Levin, 2010). We do not believe in using randomization for the sake of increasing the similarities between SCD experimental work and group design experimental work. However, there are instances in which using randomization improves rigor by reducing bias.

The use of randomization in SCD is not new; randomization is reported in studies using SCD in the literature as many as 35 years ago (e.g., Wolery & Billingsley, 1982; Wolery, Holcombe, Werts, & Cipolloni, 1993). Randomization of intervention condition implementation has a long history of use in both rapid iteration designs (e.g., ATDs; discussed by Barlow & Hayes, 1979 and Holcombe, Wolery, & Gast, 1994) and multiple probe or baseline designs (Wolery & Billingsley, 1982). As described in Chapters 10 and 11, it is commonly used to order condition implementation in studies using ATDs and AATDs (cf. Haydon et al., 2010; Ingersoll, 2011; Lynch, Theodore, Bray, & Kehle, 2009) and less-commonly used to order intervention implementation in multiple probe and baseline designs (cf. Ledford, Gast, Luscre, & Ayres, 2008; Wolery, Holcombe, Werts, & Cipolloni, 1993).

More recently, additional uses of randomization have been suggested: (a) randomized start times for intervention conditions ("randomized phase start-point designs"; Kratochwill & Levin, 2010, p. 131) and (b) randomization of intervention condition implementation ("randomized phase order designs," p. 131). Randomized phase start-point designs would be used, for example, if a researcher randomly determined the session during which a participant would move between baseline (A) and intervention (B) conditions during an A-B-A-B design without consideration of data patterns (participant responding). Difficulties would arise when these random start points are selected without consideration for data patterns (e.g., when a therapeutic trend exists during baseline condition, which would usually result in postponement of condition changes). Randomized phase order designs include randomly determining which condition was implemented at a given time for a participant (e.g., beginning with A or B conditions in an A-B-A-B design). Difficulties exist with this method when it is logically important for a baseline condition to be completed prior to an intervention condition.

In group design comparison research, randomization serves the purpose of minimizing threats to internal validity by randomly assigning participants to interventions. When large numbers of participants exist, this randomization serves the purpose of "evening out" groups; thus, any pre-intervention differences (e.g., socioeconomic status, IQ) that exist among participants are theoretically evenly distributed between groups. However, randomly assigning a very small number of participants (or items or days) does not serve the same purpose—randomization only works as an equalizer when large numbers are present. This has been acknowledged in group research; statisticians generally cite 50 as

the smallest number of participants who can be randomly assigned to one of two intervention groups (cf. Kang, Ragan, & Park, 2008; Singh, 2006). Thus, although we agree that randomization is a reasonable addition when designing SCD research, we do not believe that it is always warranted or that it "rescues" the credibility of SCD research (Kratochwill & Levin, 2010).

As discussed in Chapter 11, counterbalancing serves as a control similar to matching participants and then randomizing group assignment in group design studies (i.e., matched pairs design). For example, if one student participates in a study with an A-B-C-B-C design, the second participant should participate in the interventions in a counterbalanced order: an A-C-B-C-B design. In a group design, matching participants allows researchers to confidently assume that pre-intervention differences aren't responsible for behavior change; likewise, counterbalancing conditions allows researchers to confidently assume that intervention order is not responsible for behavior change. The same is true of counterbalancing behaviors in an SCD comparison design; when randomly assigning six behaviors to two different instructional procedures within the context of an AATD, it is likely that the two behavior sets will not be of equal difficulty (the number of randomized elements is not large enough to "even out" the difficulty level of the behaviors). However, suppose you assign a set of behaviors to Intervention B for the first participant and to Intervention C for the second participant. If Intervention B consistently results in better outcomes, we can be confident that results were due to Intervention B being the superior intervention, rather than the alternate explanation that the behaviors taught with one intervention were easier than those taught with the other intervention. After choosing the counterbalancing method (e.g., two participants will participate in a study in the context of a B-C-B-C design and two participants will participate with a C-B-C-B design), the participants (or stimuli) could be randomly assigned with the counterbalancing rules in mind (e.g., rather than purposively selecting two participants to participate in the B-C-B-C variation, randomly select them).

Despite misgivings about some uses of randomization, we find the following uses of randomization acceptable, although control for threats to internal validity is still dependent on appropriate use of an SCD:

1. *Randomized start times for interventions*: Randomly determining the start date for an intervention in the context of any SCD is reasonable, given a small window that does not begin until baseline data are stable (e.g., within 0–2 days after a pre-determined minimum number of sessions with stable baseline data).
2. *Randomized condition implementation*: Randomly ordering conditions is typically done in rapid alternation designs, although restricted randomization rather than true randomization is often used. This is reasonable (and widely used) in ATD, AATD, PTD, and multielement designs.
3. *Randomized stimuli assignment*: When two interventions are compared in the contexts of AATD, PTD, or RA designs, it is critical to randomly assign stimuli or

sets of stimuli to interventions to avoid potential bias (e.g., assigning an "easier" set to a preferred intervention).

4. *Randomized assignment of tiers*: When using MB and MP designs, it is important to randomly assign order of implementation across tiers to minimize the likelihood of bias. This is especially critical when using MB and MP designs across participants.

Internal validity does not *depend* on randomization in SCD. However, the inclusion of randomization given the conditions above likely decreases risk of bias and thus increases the rigor.

Purposes of Evaluating Rigor

Researchers evaluate rigor for several interrelated reasons, including to determine (1) to what extent an individual study has been conducted in a manner to allow confidence in results, (2) to what extent a *group* of studies has been conducted in a manner to allow confidence in overarching conclusions regarding outcomes, and (3) areas of improvement for a body of research related to a specific independent or dependent variable (e.g., What improvements in rigor are needed in future research?) Regardless of whether you are assessing a single study or a group of related studies, it is critical that you assess rigor *before you assess outcomes*. This will allow you to determine whether changes in behavior are believable (i.e., whether you are confident that changes in behavior are due to experimental manipulations *and only* those manipulations).

Standards, Quality Indicators, and Rating Frameworks

A number of tools have been designed to assist researchers in assessing the rigor of SCD studies; the purpose, content, guidelines, and use of each is slightly different. For example, **standards** are generally designed to outline a basic set of benchmarks, which serve as minimum criteria for evaluation. Standards can be used to determine, for example, whether studies should be included in a review synthesis (discussed in the next chapter)—studies that do not meet standards are methodologically weak—their findings cannot be interpreted to draw conclusions in the same way as studies that meet standards. **Quality Indicators** and **Rating Frameworks** are groups of characteristics designed to determine to what extent a study includes factors determined to be critical to evaluating study quality and/or rigor. These tools are designed to answer questions regarding to what extent studies meet a wide variety of conditions, rather than only basic criteria critical for internal validity. A study may meet basic standards but still have low rigor, high risk of bias, or low generality or applicability (quality, external validity). Thus, when using a tool to assess rigor, ensure that you are familiar with guidelines regarding tool use and that you use a tool as intended or explicitly describe your deviations.

Tools for Characterizing Rigor

Below, we briefly describe some widely used tools for assessing rigor or quality of SCD research. Others are available, and we expect that additional ones will be developed in the near future. Syntheses of SCD research studies that have included use of each tool are shown in Table 13.1. Note that we do not consider this a comprehensive list of tools; information on a variety of tools is available in the literature (cf. Maggin, Briesch, Chafouleas, Ferguson, & Clark, 2014; Wendt & Miller, 2012). The critical factor when determining which tool to use is to choose one that (a) matches your purpose, and (b) includes valuation of critical factors. In addition to these published tools listed below, we have also included as Appendix 13.2 a *Rigor and Quality Checklist* at the end of this chapter. This checklist includes items we consider crucial to internal validity (rigor) and quality (generality, external validity).

Table 13.1 *Examples of Use of Tools for Evaluating Rigor*

Tool	Article Reference	Research Question	General Findi Regarding Rigor
Exceptional Children (Horner et al., 2005)	Knight, V., Sartini, E., & Spriggs, A. D. (2015). Evaluating visual activity schedules as evidence-based practice for individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *45*, 157–178.	What is the quality of the literature related to visual activity schedules? What is the magnitude of effects? Can VAS be considered an evidence-based practice?	Of 31 studies, received all possible poi (some indicators v divided into more than yes/no response, ea receiving a score of 0 o 15 received "acceptable ratings (inclusion c critical characterist 15 received "unacceptal ratings.
Rogers, L. A., & Graham, S. (2008). A meta-analysis of	Which writing practices have been shown to be effective for students in grades	Scores of 0–1 were assigned for each of 11 indicators; range was 4.0–	

single subject design writing intervention research. *Journal of Educational Psychology*, *100*, 879.	1–12?	11.0. Of 75 quality scores, only 2 were 11.0, indicating the study met all indicators.	
WWC Standards (Kratochwill et al., 2010)	Fallon, L. M., Collier-Meek, M. A., Maggin, D. M., Sanetti, L. M., & Johnson, A. H. (2015). *Exceptional Children*, *81*, 227–246.	Is performance feedback for educators an evidence-based practice?	More than hal the studies (n=81) met design standards; another qua (n=45) met standards v reservation Other studi (N=43) did meet standa Authors reported studies not meeting standards generally d not have sufficient number of potential demonstrat or data poi
	King, S. A., Lemons, C. J., & Davidson, K. A. (2016). Math Interventions for Students With Autism Spectrum Disorder: A Best-Evidence Synthesis. *Exceptional Children*, *82*, 443–462.	What are the features of high-quality empirical studies published in peer-reviewed journals in which the efficacy of math interventions was evaluated for students	Studies were included in review if th met *all* of t WWC crite Based on th criteria, 10 24 articles (29/57 case were exclu from analy

		with ASD?	
	Smith, K., Shepley, S., Alexander, J., & Ayres, K. (2015). The independent use of self-instructions for the acquisition of untrained tasks for individuals with an intellectual disability: A review of the literature. *Research in Developmental Disabilities*, *40*, 19–30.	To what extent have studies demonstrated improved generalized use of self-instruction materials to learn multi-step tasks for participants with intellectual disabilities?	One of 19 stud (5%) met standards, a 8 (45%) met with reservation Remaining studies (n= 55%) did no meet evider standards. Reasons for meeting standards include: DV reliability (N=7) and number of points (N=3
CEC Quality Indicators (CEC, 2014)	Kaldenberg, E. R., Watt, S. J., & Therrien, W. J. (2015). Reading instruction in science for students with learning disabilities: A meta-analysis. *Learning Disability Quarterly*, *38*, 160–173.	What interventions were most effective for increasing reading comprehension of science text for students with learning disabilities?	On average, studies met 20/24 indicators, a range of 1 22. All stud met criteria some doma context and setting, participants description practice, internal validity, da analysis. Fe studies had adequate internal reliability a many did n meet the criteria for implementa fidelity.
CEC Quality Indicators (CEC, 2014)	Losinski, M., Sanders, S. A., & Wiseman, N. M. (2016).	What are the relative effects of studies	One study me standards; t average

	Examining the use of deep touch pressure to improve the educational performance of students with disabilities: a meta-analysis. *Research and Practice for Persons with Severe Disabilities*, *41*, 3–18.	examining DTP with students with disabilities? Does the use of DTP with students with disabilities meet the CEC (2014) standards for an EBP?	percentage indicators r by study w 77% (range: 100%). Autl reported m common omissions v implementa fidelity, interventio agent, and social valid
RoBiNT Scale (Tate et al., 2014)	Sigmundsdottir, L., Longley, W., & Tate, R. L. (2016). Computerized cognitive training in acquired brain injury: A systematic review of outcomes using the International Classification of Functioning. *Neuropsychological Rehabilitation*, *26*, 673–741.	What are the outcomes of computerized cognitive training for adults with acquired brain injury?	The study included 1 single case design den as "experimen (e.g., A-B designs), ar 13 denoted "non-experiment The sole experiment study recei 12/30 possil points.
	Tate, R., Wakim, D., & Genders, M. (2014). A systematic review of the efficacy of community-based, leisure/social activity programmes for people with traumatic brain injury. *Brain Impairment*,	What are the outcomes of community-based leisure/social activity intervention programs for individuals with traumatic brain injury?	The study included 1 single case design, and received a score of 12/30.

15, 157–176.			
SCARF (Ledford et al., 2016)	Zimmerman, K. N., & Ledford, J. R. (2017). Evidence for the effectiveness of social narratives: Children without ASD. *Journal of Early Intervention, 39,* 199–217.	To what extent have social narratives been assessed for children who do not have autism?	Three designs scores of <1 (lowest quality/rigo five had sc between 1 a 2 (low); tw had scores between 2 a 3 (high), an zero had sc between 3 a 4 (highest quality/rigo Scores were lowest regarding measureme of generalizati and fidelity
Zimmerman, K. N., Ledford, J. R., Severini, K. E., Pustejovsky, J. E., Barton, E. E., & Lloyd, B. P. (2017). A Comparison of Methods to Evaluate Quality and Rigor when Synthesizing Single Case Research Designs. *Under Review.*	To what extent do frameworks, overlap metrics, and effect sizes suggest that antecedent sensory-based materials manipulations result in positive behavior change for young children?	Of 51 designs, 6 had scores of <1.0 (lowest quality/rigor), 22 had scores between 1 and 2 (low), 23 had scores between 2 and 3 (high), and 0 had scores between 3 and 4 (highest). Scores were lowest regarding measurement of generalization and fidelity.	
Risk of Bias Tool (Reichow et al., 2017)	Barton, E. E., Reichow, B., Schnitz, A., Smith, I. C., & Sherlock, D. (2015). A	What is the evidence of effectiveness for sensory-based	Studies had lo or unclear r of bias in sc areas, with or fewer

	systematic review of sensory-based treatments for children with disabilities. *Research in Developmental Disabilities*, *37*, 64–80.	interventions for children with disabilities?	studies showing hi risk (participant selection, blinding of assessment, reliability, sampling); 50% of stud showed hig risk (sequer generation) and more tl 50% of stud showed hig risk of bias both blindi of participa and person and proced fidelity.

Exceptional Children

In 2005, *Exceptional Children* (*EC*), a premier journal in special education, printed a special issue on research design. In the article regarding SCD research, Horner and colleagues described the use of SCD studies to establish evidence-based practices in special education, described experimental control in SCDs, and explicated that rigorously conducted SCD research was experimental in nature. They also described 21 quality indicators for SCD research, in seven areas: Descriptions of participants and settings, dependent variable measurement, independent variable measurement, baseline, experimental control, external validity, and social validity. As of 2017, this article remained the most highly cited paper in EC and numerous reviews of SCD research have used these quality indicators to assess rigor in groups of related studies (see Table 13.1). The primary drawback of these quality indicators is that they are dated (i.e., expectations for rigor have increased over time) and not all indicators are created equal. For example, two studies with similar numbers of addressed quality indicators might have very different adequacy in terms of rigor. This issue, which is not unique to this tool, is discussed by a number of authors in relation to evaluation of rigor (cf. Wendt & Miller, 2012; Maggin, Briesch, Chafouleas, Ferguson, & Clark, 2014). For example, if two studies each addressed 20 of the 21 indicators, but one had low ecological validity (e.g., used researcher rather than indigenous implementers) and one had low internal validity (e.g., did not include three potential demonstrations of effect), confidence in conclusions from

each is considerably different despite the identical quality score. Regardless, this is a highly used and well-regarded tool that includes critical components for assessing rigor and quality.

WWC Standards

The What Works Clearinghouse (WWC) "pilot" design standards were introduced in 2010. As described in Chapter 1, the purpose of the WWC is to review research evidence in the field of education to provide information to practitioners to inform evidence-based decisions. The standards were developed to "guide WWC study reviewers when making decisions about the internal-causal validity (i.e., internal validity) of a particular study …" (Hitchcock et al., 2014, p. 145). Hitchcock states that an intended benefit of the standards was to promote increased rigor in SCD research. The WWC standards are considerably less comprehensive than the EC quality indicators, as differences between the two types of tools (described above) indicate. The current WWC standards include indicators related to: (a) systematic implementation of the intervention; (b) adequate dependent variable reliability; (c) number of potential demonstrations; and (d) number of data points. One prominent difference is that the WWC Standards did not include an item regarding fidelity measurement (discussed at length in several published articles; Hitchcock et al., 2014; Wolery, 2013); we consider fidelity measurement to be critical. Nonetheless, a considerable improvement of the standards is the use of a gating system (Maggin et al., 2014), wherein studies are evaluated for rigor first; outcomes are only evaluated if rigor is sufficient. They also use two levels of meeting standards—a study *Meets Evidence Standards* if there are five or more data points in all conditions and *Meets Standards with Reservations* if all conditions have at least three data points, but at least one condition has fewer than five. A study *Does not Meet Evidence Standards* if there are fewer than three data points in any condition, fewer than three potential demonstrations of effect (five potential demonstrations in ATDs), reliability data are not collected or low agreement is present, or if the independent variable is not systematically manipulated. For studies that *Meet* or *Meet with Reservations*, outcomes can be assessed via visual analysis. The most recent WWC handbook (as of November, 2017) is available here:
https://ies.ed.gov/ncee/wwc/Docs/referenceresources/wwc_standards_handbook_v4.pdf.

CEC Quality Indicators

The Council for Exceptional Children (CEC) published their *Standards for Evidence-Based Practices in Special Education* in 2014. These guidelines, including quality indicators (QIs) applicable to group design, SCD, and both types of designs, include many of the same indicators in the Horner et al. (2005) article in *EC*. There are eight areas of assessment; the first four (context and setting, participants, intervention agent,

and description of practice) are related to adequate reporting of study characteristics. The next three areas (implementation fidelity, internal validity, and outcome measures/dependent variables) are related to both rigor and reporting (e.g., dependent variables should be measured reliably and described well). The final category, data analysis, requires the presentation of data via an SCD line graph. In order to be considered "methodologically sound," studies must meet *all quality indicators*.

Following quality appraisal, outcomes analysis is conducted. There are three potential outcome categories. *Positive effects* are present when a functional relation is established, with ¾ of participants showing positive and "meaningful" behavior change and no contra-therapeutic effects. Negative effects are present when behavior changes for ¾ of participants *in the unintended direction*. Neutral or mixed effects are present when neither positive nor negative effects can be established. We should note that outcomes analysis for this tool occurs *at the study level* rather than at the design level. This is incongruent with other tools (e.g., WWC, SCARF) and can lead to discrepant conclusions. To be considered an evidence-based practice based on SCD research, the CEC QIs stipulate there must be (a) at least five methodologically sound studies including at least 20 participants with positive effects *and* (b) include 0 methodologically sound studies with negative effects *and* (c) have more methodologically sound studies with positive effects than neutral or mixed effect (at least a 3:1 ratio). Other potential classifications of evidence include *potentially evidence-based practice*, *mixed evidence*, *insufficient evidence*, and *negative effects*. Because this tool stipulates that studies should include 100% of quality indicators in order to be designated as methodologically sound, it is likely that most practices would be evaluated as having "insufficient evidence" (i.e., an insufficient number of methodologically sound studies).

RoBiNT Scale

A group of researchers from Australia (Tate et al.) developed and updated the Risk of Bias in N-of-1 Trials (RoBiNT) Scale in 2013. The RoBiNT includes 15 items and uses terminology common to medicine (e.g., N-of-1). Unlike the previously mentioned tools, the items are not rated in a binary fashion, but rather using a three-point scale. It includes many of the same items as WWC standards (e.g., at least three demonstrations of effect, at least 5 data points per condition, IOA at 80% or higher) and Quality Indicators as described in *EC* and by CEC (e.g., descriptions of baseline and intervention conditions, operational definitions of target behaviors, graphed data for evaluation, treatment adherence [fidelity]). Additional items include replication to improve generality (not including within-design replication) and measurement of generalization in the context of the design. Some items more common in medical research (as compared to typical SCD research in education, psychology, or behavioral sciences) are the inclusion of randomization, blinding participants and implementers, and blinding data collectors. Most evaluated SCD studies received scores of 0 for all of these items (Tate et al., 2014). Items are divided into two sub-scales allowing the evaluator to calculate

separate scores for internal and external validity; the ability to separate rigor and generality is a considerable strength of this tool.

Single Case Analysis and Review Framework

The Single Case Analysis and Review Framework (SCARF; Ledford, Lane, Zimmerman, Chazin, & Ayres, 2016) was developed in 2016 and has only been used in a single published review (Zimmerman & Ledford, 2017). Like the WWC tool, but dissimilar to the CEC tool, it is designed to evaluate studies at the design level. In addition to avoiding generalizations across designs, evaluating at the design level also allows for the evaluation of some designs in a particular study rather than the study as a whole (e.g., if your research question is related to children with autism and a study includes two A-B-A-B designs, one for a child with autism and one for a child with developmental delays, you can choose to complete the SCARF for only one of the participants).

The SCARF is divided into three sections. The first assesses rigor (internal validity) and is more heavily weighted than the second section, which is designed to assess quality and breadth of measurement. All items are scored on a 0–4 point scale, and item scores are averaged to calculate a section score. The three components of rigor are reliability, data sufficiency, and fidelity. The seven components scored for quality and generality include descriptions of (1) participants, (2) conditions, and (3) dependent variables; (4) social and ecological validity; and measurement of (5) maintenance, (6) stimulus generalization, and (7) response generalization. The average score from each section is calculated; the total score for the study is

$$((\mathrm{rigor\ average}) \times 2 + (\mathrm{quality\ and\ breadth\ of\ measurement\ average})) \div 3$$

and is also on a 0–4 point scale. The final section, designed for outcome evaluation, is also scored on a scale of 0–4. Generally, a 4 is scored if there are at least 3 demonstrations and no weak effects or non-effects. Weak effects are behavior changes that occur between conditions, but that are not immediate and abrupt; this decreases confidence that the intervention and only the intervention is responsible for behavior change. A score of 3 indicates at least three demonstrations and no non-effects. A score of 2 indicates at least 3 demonstrations but at least 1 non-effect. A score of 1 is fewer than three demonstrations, with one non-effect. A score of 0 is given when there is more than one demonstration of no behavior change in a single design. Modified criteria are provided for MB and MP designs (related to concurrence and vertical analysis) and to ATDs (with a focus on differentiation rather than behavior change). The scoring sheet for the SCARF is available online (http://vkc.mc.vanderbilt.edu/ebip/scarf/).

One novel quality of the SCARF tool is that there is a convention for presenting the results of analysis of multiple studies, via a scatterplot (see Figure 13.2). Each data point represents a single design, with the total score for quality and rigor on one axis and the outcomes score on the other. Using these coordinates, studies that have high rigor and

positive outcomes will be in the upper right quadrant (labeled B in Figure 13.2) and studies that have high rigor and less optimal outcomes will be in the lower right quadrant (labeled D in Figure 13.2). Studies depicted to the left of the mid-line generally have quality that is insufficient for drawing conclusions (quadrants A and C in Figure 13.2); thus, outcomes from high- and low-rigor studies are plotted using the same conventions but are separated via their quality and rigor scores to allow for analysis. This allows readers to quickly determine what proportion of the studies are high quality (to the right of the vertical midline) and have positive outcomes (above the horizontal midline).

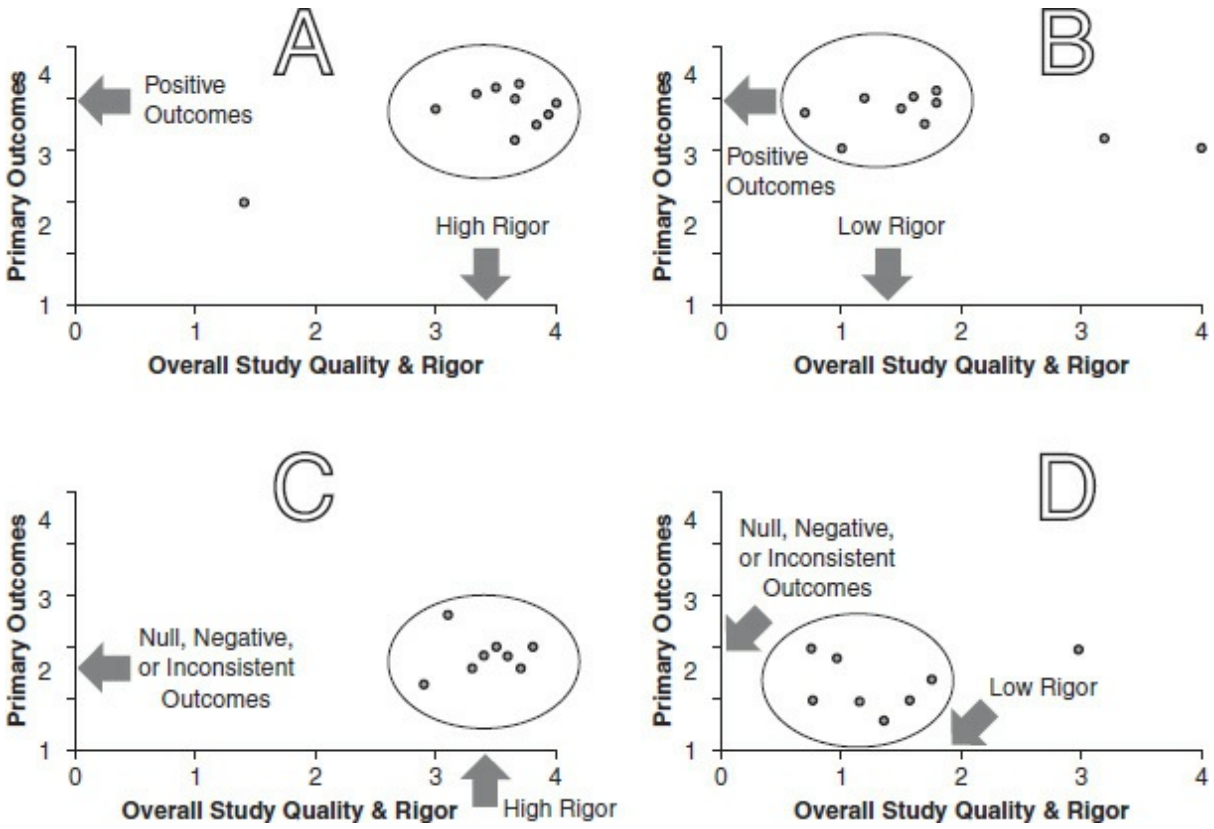


Figure 13.2 Sample data from the SCARF tool showing (A) a group of studies with high rigor and positive outcomes; (B) a group of studies with low rigor and positive outcomes; (C) a group of studies with high rigor and null, negative, or inconsistent outcomes; and (D) a group of studies with low rigor and null, negative, or inconsistent outcomes.

Risk of Bias Tool

Reichow, Barton, and Maggin (2017) have developed a specific tool designed to evaluate the risk of bias in SCD studies; this tool was modified from the risk of bias tool used in Cochrane Collaboration meta-analyses for group design studies (Higgins et al., 2011). The risk of bias tool assesses biases in seven areas, including sequence generation, participant selection, blinding of participants and personnel, procedural fidelity, blinding of outcome assessment, dependent variable reliability, and data sampling. This tool has

been used in only two published reviews (Barton, Pustejovsky, Maggin, & Reichow, 2017; Barton, Reichow, Schnitz, Smith, & Sherlock, 2015). The risk of bias tool's primary strength is that it is somewhat easily compared to the risk of bias tool designed for group design studies. For example, as shown in Figure 13.3, in a review of sensory-based interventions, Barton et al. found that both group and SCD studies had high risks of bias in the area of procedural fidelity. Thus, when a review includes both group comparisons and SCD studies, this tool might be particularly valuable to allow readers to draw conclusions regarding bias across similar categories.

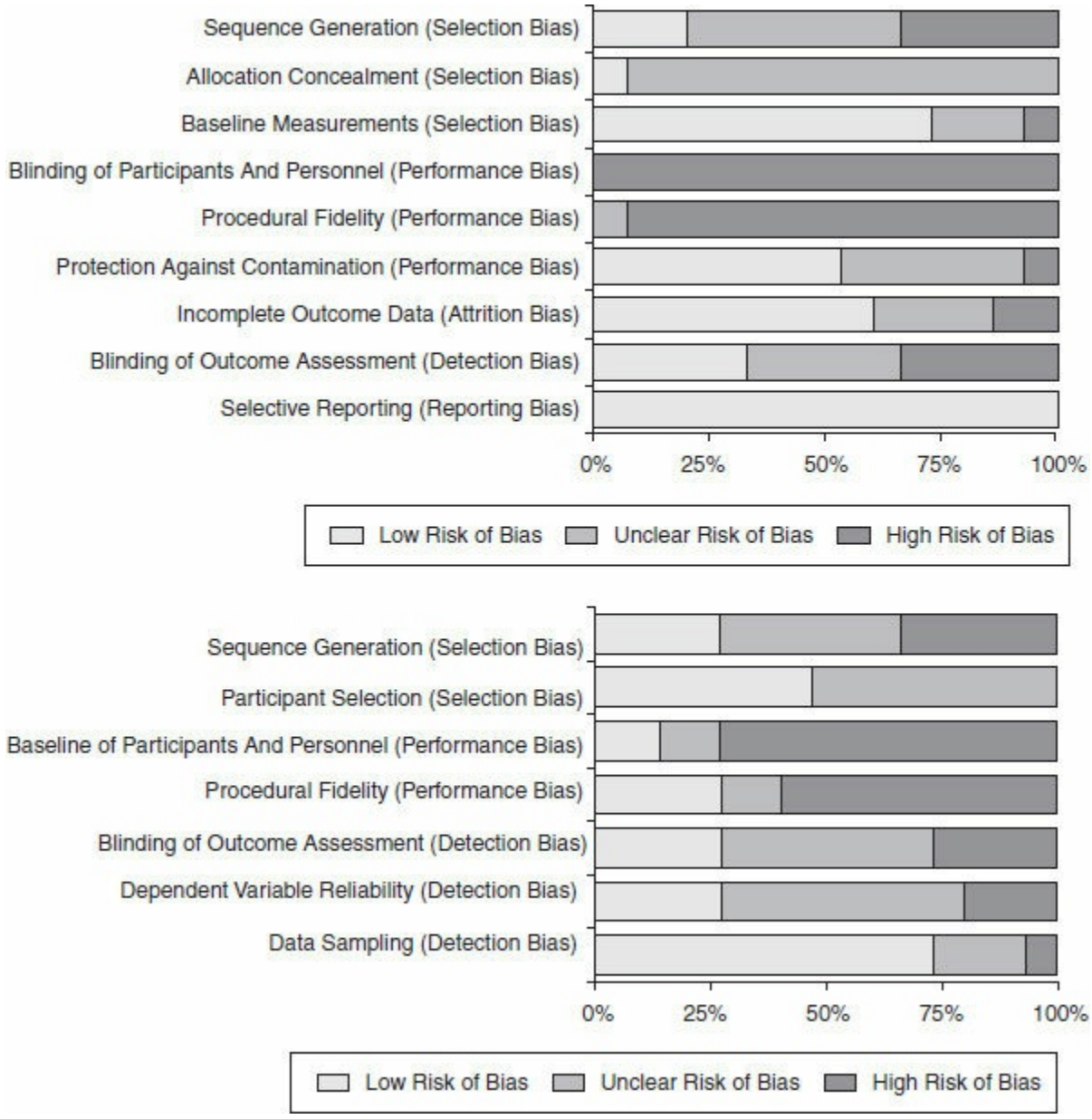


Figure 13.3 Results from risk of bias tool.

Source: Barton, E. E., Reichow, B., Schnitz, A., Smith, I. C., & Sherlock, D. (2015). A systematic review of sensory-based treatments for children with disabilities. *Research in Developmental Disabilities*, *37*, 64–80.

Adequate Reporting

The written report of an SCD study is the public record of the research. Thus, the report needs to provide an accurate account of the study and its findings and should be written with clarity, transparency, and completeness. This allows for replication of procedures and further assists with answering for whom, for what behaviors, and under what conditions an intervention is appropriate; inaccurate or incomplete descriptions of procedures impedes further establishment of evidence, or lack thereof, for an intervention. To facilitate adequate reporting practices in the literature, guidelines are available to assist authors in creating written reports of SCD studies. The **S**ingle-**C**ase **R**eporting guideline **I**n **Be**havioral interventions (SCRIBE; Tate et al., 2016a, 2016b) is a contemporary reporting guide developed specifically for SCDs and was a response to inadequate or uneven reporting practices observed in the literature. For example, Maggin et al. (2011) found that even basic demographic information was often not reported in 24 SCD studies evaluating token economies for challenging behaviors in students (details of age were absent in 42% of reports and of sex in 33%); similarly, in 253 SCD studies in the neuro-rehabilitation field, Tate et al. (2014) found no information reported on inter-rater agreement of the target behavior (46% of reports) or whether the assessor was independent of the therapist (86%); in an extensive survey of 409 SCD reports in education and psychology journals, Smith (2012) found no provision of baseline data in 22% of articles, and no visual or statistical analysis of the data in 52%. These examples highlight the need to improve the completeness of reporting SCD studies.

SCRIBE was developed using procedures recommended by Moher and colleagues (2010a) for the CONSORT (***CON****solidated* ***S****tandards* ***O****f* ***R****eporting* ***T****rials*) family of reporting guidelines. The first CONSORT reporting guide was published in 1996 by Begg and colleagues (the most recent revision appears in Moher et al. (2010b) with the aim of improving incomplete reporting of randomized controlled trials in the medical field. In the intervening 20 years, a large number of reporting guides for many different types of methodologies have been published using the CONSORT procedures as a model. Reporting guidelines are available for systematic reviews, observational and diagnostic studies, and qualitative research (see EQUATOR Network, www.equator-network.org, which archives all reporting guidelines using CONSORT procedures).

The incentive to create SCRIBE derived from a similar endeavor to develop a reporting guide for N-of-1 Trials in the medical literature (Shamseer et al., 2015; Vohra et al., 2015). SCRIBE is intended to be applicable to all fields of the behavioral sciences. Accordingly, a group of world experts was assembled, with representation from content experts in clinical and neuropsychology, educational psychology and special education, medicine, occupational therapy and speech pathology, as well as single-case methodologists and statisticians, journal editors and a medical librarian, and guideline developers. SCRIBE items were evaluated in two rounds of an online Delphi survey, and subsequently

finalized during a two-day consensus conference. The methodology and procedures used to develop SCRIBE are described in Tate et al. (2016a). That article was published in 10 journals simultaneously, representing a broad range of disciplines, to facilitate widespread dissemination of the work. A more detailed 'explanation and elaboration' article (Tate et al., 2016b) provides the rationale for including each of the items of SCRIBE and examples of adequate reporting from the literature.

The main product of SCRIBE is a 26-item checklist, which users can download from the SCRIBE website (www.sydney.edu.au/medicine/research/scribe), and is reproduced in Table 13.2. SCRIBE provides authors with information on *what* to report in sections, and sub-sections, commonly included in published literature (Introduction, Method, Results, Discussion).

Table 13.2 *The Single-Case Reporting guideline In BEhavioural interventions (SCRIBE) 2016 Checklist*

Item number	Topic	Item description
TITLE and ABSTRACT		
1	Title	Identify the research as a single-case experimental design in the title
2	Abstract	Summarize the research question, population, design, methods including intervention/s (independent variable/s) and target behavior/s and any other outcome/s (dependent variable/s), results, and conclusions
INTRODUCTION		
3	Scientific background	Describe the scientific background to identify issue/s under analysis, current scientific knowledge, and gaps in that knowledge base
4	Aims	State the purpose/aims of the study, research question/s, and, if applicable, hypotheses
METHODS		
	DESIGN	
5	Design	Identify the design (e.g., withdrawal/reversal, multiple-baseline, alternating-treatments, changing-criterion, some combination thereof, or adaptive design) and describe the phases and phase sequence (whether determined *a priori* or data-driven) and, if applicable, criteria for phase change
6	Procedural changes	Describe any procedural changes that occurred during the course of the

		investigation after the start of the study
7	Replication	Describe any planned replication
8	Randomization	State whether randomization was used, and if so, describe the randomization method and the elements of the study that were randomized
9	Blinding	State whether blinding/masking was used, and if so, describe who was blinded/masked
PARTICIPANT/S or UNIT/S		
10	Selection criteria	State the inclusion and exclusion criteria, if applicable, and the method of recruitment
11	Participant characteristics	For each participant, describe the demographic characteristics and clinical (or other) features relevant to the research question, such that anonymity is ensured
CONTEXT		
12	Setting	Describe characteristics of the setting and location where the study was conducted
APPROVALS		
13	Ethics	State whether ethics approval was obtained and indicate if and how informed consent and/or assent were obtained
MEASURES and MATERIALS		
14	Measures	Operationally define all target behaviors and outcome measures, describe reliability and validity, state how they were selected, and how and when they were measured
15	Equipment	Clearly describe any equipment and/or materials (e.g., technological aids, biofeedback, computer programs, intervention manuals or other material resources) used to measure target behavior/s and other outcome/s or deliver the interventions
INTERVENTIONS		
16	Intervention	Describe intervention and control condition in each phase, including how and when they were actually administered, with as much detail as

		possible to facilitate attempts at replication
	Procedural fidelity	Describe how procedural fidelity was evaluated in each phase
NALYSIS		
18	Analyses	Describe and justify all methods used to analyze data
RESULTS		
19	Sequence completed	For each participant, report the sequence actually completed, including the number of trials for each session for each case. For participant/s who did not complete, state when they stopped and the reasons
20	Outcomes and estimation	For each participant, report results, including raw data, for each target behavior and other outcome/s
21	Adverse events	State whether or not any adverse events occurred for any participant and the phase in which they occurred
DISCUSSION		
22	Interpretation	Summarize findings and interpret the results in the context of current evidence
23	Limitations	Discuss limitations, addressing sources of potential bias and imprecision
24	Applicability	Discuss applicability and implications of the study findings
DOCUMENTATION		
25	Protocol	If available, state where a study protocol can be accessed
26	Funding	Identify source/s of funding and other support; describe the role of funders

Source: Tate et al. (2016). The single case reporting guideline in behavioral interventions (SCRIBE) 2016: Explanation and Elaboration. *Archives of Scientific Psychology*, *4*, 10–31.

Readers should be aware of the different purposes of a reporting guide versus tools for assessing rigor. The former is a guide for authors writing a report, to instruct on *what* to report. Such guides are also helpful for journal editors and reviewers to determine whether a written report provides all the necessary information. In contrast, tools for assessing rigor inform the reader of *how well* a study was conducted (see previous sections). For example, item 14 of SCRIBE (concerned with measures) asks the author to describe "… how and when [the target behaviors] were measured." An author might report, for example, that in an A-B-A-B design study, the target behavior was measured

weekly during each two-week baseline (A) and each two-week intervention (B) conditions. This is an example of adequate reporting, because the reader knows exactly what was done (measures of the target behavior were collected twice in every baseline and intervention condition). However, such a study would likely not meet contemporary standards for adequate internal validity, as assessed via a tool for assessing rigor.

It is important to remember the distinction between reporting guides and tools for assessing rigor; when used appropriately, they are helpful resources to improve the conduct and reporting of SCD studies. Using reporting guidelines as a metric for determining if a report is technologically sound, for purposes of transparency and options for replication, will likely improve the quality of written reports (Turner et al., 2012). As the examples provided in the 'explanation and elaboration' article (Tate et al., 2016b) demonstrate, adequate reporting can be achieved for SCDs without the need for exhaustive detail. In cases where adequate reporting does require extensive description, or where reports are written for journals that impose restrictive word lengths on articles, online supplements are a useful way to include additional information.

Conclusions

The impetus of applied research is identifying efficacious, effective, and efficient interventions to establish well-developed technologies for indigenous implementers in typical environments. This process encompasses (1) conducting rigorous and high-quality SCD studies and (2) evaluating the collection of available studies to answer *what we already know* about a given topic and *what is next* for the corresponding field of study. Evaluating rigor of SCD studies is the critical first step in determining to what extent subsequent analysis of data should occur; inadequate rigor negates further analysis of data, indicating little to no confidence that the independent variable, and the independent variable alone, was responsible for changes in the dependent variable. Adherence to guidelines, especially transparency in reporting practices, when preparing reports on SCD studies will facilitate evaluations of rigor in future reviews. As fields strive to scale up interventions for indigenous implementers, reviews of past research and adherence to contemporary standards promote a unified approach to research and practice that aligns with the basic tenets of science and research.

Appendix 13.1

Choosing an Appropriate Research Design

Is your research question a **demonstration** question (e.g., comparing intervention to baseline) or a **comparison** question (e.g., comparing two interventions)?

_____

Is your primary dependent variable of interest **reversible** (i.e., likely to reverse to baseline levels if you withdraw the intervention) or **non-reversible** (i.e., once the participant learns it, they will perform it accurately even in the absence of intervention?)

_____

Reversible Behaviors

*Demonstration/Reversible: If you are interested in a **demonstration** question and you are measuring **reversible** behaviors, the following designs may be feasible for use:*		
Design	Choose if…	Avoid if…
A-B-A-B	In most situations	Participants, implementers, and stakeholders find intervention withdrawal unacceptable
Multiple baseline	An A-B-A-B design is not feasible and instability threats are more likely than testing threats	You identify multiple targets but they are likely to covary; if testing threats are likely; if participants object to a prolonged baseline
Multiple probe	An A-B-A-B design is not feasible and testing threats are more likely than instability threats	You identify multiple targets but they are likely to covary; if instability threats are likely; if participants object to a prolonged baseline
Changing criterion	Your independent variable can be administered according to step-wise criterion, and your dependent variable is likely to closely align with criterion level	Your dependent variable is an acquisition behavior (i.e., appropriate for performance rather than acquisition deficits); your independent variable cannot be implemented according to desired dependent variable levels

*Comparison/Reversible: If you are interested in a **comparison** question and you are measuring **reversible** behaviors, the following designs may be feasible for use:*		
Design	Choose if…	Avoid if…
ATD	It is feasible to rapidly alternate conditions; you require a short-duration comparison; you are interested in a demonstration + comparison question (use a continuing baseline condition)	It will be difficult for implementers to rapidly alternate between conditions or for participants to discriminate between conditions
Multitreatment	It is preferable to slowly	You have limited

	alternate conditions; you are interested in only a comparison question	time with which to conduct the comparison
M-ED	You are interested in comparing assessment conditions (e.g., functional analysis or structural analysis)	You are interested in comparing intervention conditions
Simultaneous treatments	You are interested in assessing participant preference or choice	You are interested in comparing other dependent variables

Non-Reversible Behaviors

*Demonstration/Non-Reversible: If you are interested in a **demonstration** question and you are measuring **non-reversible** behaviors, the following designs may be feasible for use:*		
Design	Choose if…	Avoid if…
Multiple baseline	Testing threats are more likely than instability threats	Testing threats are likely in baseline conditions
Multiple probe	Instability threats are more likely than testing threats; continuous data collection is not feasible	Data instability is likely in baseline conditions

*Comparison/Non-Reversible: If you are interested in a **comparison** question and you are measuring **non-reversible** behaviors, the following designs may be feasible for use:*		
Design	Choose if…	Avoid if…
AATD	You can identify at least three sets of behaviors that are of equal difficulty; you can collect at least three sessions of baseline data	Stakeholders object to moderate pre-intervention assessments
RA design	Behaviors will be acquired very rapidly; many behaviors can be identified; stakeholders or participants object to repeated assessments prior to teaching	Participants are not likely to quickly learn target behaviors; many target behaviors cannot be identified and equated
PTD	You are interested in a demonstration + comparison question; you have extended time for the comparison; you want to conduct the most rigorous comparative test	Stakeholders object to numerous repeated assessments, you have limited time with which to conduct the comparison

How do I know which MB/MP design is appropriate?

1. If identifying three or more behaviors (or behavior sets) is possible, the *MB/MP across behaviors* allows for intra-participant replication; when you use an MB/MP

design across behaviors for multiple participants, inter-participant replication is possible. Do not use this design if you cannot identify at least three different behaviors (or behavior sets) or if maturation or covariation are likely.

2. If multiple contexts are available in which the skill is needed *and not likely to generalize*, the *MB/MP across contexts* can be used. Identifying three contexts with similar baseline characteristics but limited covariation is difficult; thus, this design is limited in utility under most circumstances.
3. If one of the other variations is not feasible, you may choose to use the *MB/MP across participants* variation; however, it does not allow for inter-participant replication and is sensitive to inconsistent intervention effects.

Appendix 13.2

Quality and Rigor Checklist

Domain	#	Criteria
Rigor	1	Is the design appropriate for answering the research question?
	2	Are there at least three demonstrations of effect at three different points of time, between two adjacent conditions?
		SID: Four adjacent conditions are required to meet this requirement (e.g., B-C-B-C but not A-B-A-C-A-B-C).
		TLI: Concurrent baselines are required to meet this requirement.
		RIA: Five alternations are generally preferred in these designs (e.g., 10 total sessions when comparing two conditions).
	3	Do authors present sufficient evidence for reliability of dependent variables? Generally, this requirement is met if inter-observer agreement data are collected regularly and across conditions and are sufficiently high to increase confidence in results.
	4	Do authors present sufficient evidence for reliability of independent variable implementation? Generally, this requirement is met if data are collected regularly and across conditions, data are collected on independent and control variables, and adherence is sufficiently high to indicate conditions were implemented as planned.
	5	Is there a sufficient amount of data in all primary comparison conditions? A minimum of three is required, but more are needed when data are variable or trends are present; five is not always sufficient.
	6	*If applicable*, is randomization used to decrease bias? Applicability varies based on design, but generally includes randomization (with or without restrictions) in RIA designs and random assignment to tiers in TLI designs.
Quality /Generality	7	Is the study ecologically valid? Criteria for this item may vary depending on your research question, but may include the use of typical settings, indigenous implementers, and meaningful outcomes.

		vary depending on your research question, but evidence of social validity include feedback from direct consumers (participants), indirect consumers (e.g., parents/teachers of participants), or other stakeholders (e.g., practitioners) and measures less subject to bias are valued more highly.
	9	Does the study adequately assess response and/or stimulus generalization? Generally, assessment in the context of a single case design is preferable; pre/post assessments provide some information but do not allow for experimental evaluations.
	10	Does the study adequately assess maintenance of behaviors in the absence of interventions? Note that we might not expect maintenance of reversible behaviors in the absence of intervention.
Reporting	11	Does the study include all relevant information regarding participant characteristics, condition descriptions, dependent variable definitions, and recording procedures?

Note: SID=sequential introduction and withdrawal designs (e.g., withdrawal, multitreatment). TLI=time-lagged implementation designs (multiple baseline, multiple probe designs). RIA=rapid iterative alternation designs (ATD, AATD).

References

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

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

Barton, E. E., Pustejovksy, J. P., Maggin, D. M., & Reichow, B. R. (2017). A meta-analysis of technology aided instruction and intervention for students with ASD. *Remedial and Special Education.* doi: 10.1177/0741932517729508

Barton, E. E., Reichow, B., Schnitz, A., Smith, I. C., & Sherlock, D. (2015). A systematic review of sensory-based treatments for children with disabilities. *Research in Developmental Disabilities*, *37*, 64–80.

Begg, C., Cho, M., Eastwood, S., Horton, R., Moher, D., Olkin, I., … Simel, D. (1996). Improving the quality of reporting of randomized controlled trials. *JAMA: The Journal of the American Medical Association*, *276*, 637–639.

Bronfenbrenner, U. (1979). *The ecology of human development: Experiments by design and nature*. Cambridge, MA: Harvard University Press.

Chazin, K. T., Ledford, J. R., Barton, E. E., & Osborne, K. O. (2017). The effects of antecedent exercise on engagement during large group activities for young children. *Remedial and Special Education.* doi: 10.1177/0741932517716899

Clarke, S., & Dunlap, G. (2008). A descriptive analysis of intervention research published in the Journal of Positive Behavior Interventions: 1999–2005. *Journal of Positive Behavior Interventions*, *10*, 67–71. Council for Exceptional Children. (2014). *Standards for evidence-based practices in special education.* Author: Arlington, VA. Retrieved May 4, 2017, from www.cec.sped.org/~/media/Files/Standards/Evidence%20based%20Practices%20and%2

Council for Exceptional Children. (2015). Initial preparation standards. In *What every special educator must know: Professional ethics and standards*. Arlington, VA: Author.

Fallon, L. M., Collier-Meek, M. A., Maggin, D. M., Sanetti, L. M., & Johnson, A. H. (2015). *Exceptional Children*, *81*, 227–246.

Foster, S. L., & Mash, E. J. (1999). Assessing social validity in clinical treatment research: Issues and procedures. *Journal of Consulting and Clinical Psychology*, *67*, 308–319.

Hanley, G. P. (2010). Toward effective and preferred programming: A case for the objective measurement of social validity with recipients of behavior-change programs. *Behavior Analysis in Practice*, *3*, 13–21.

Haydon, T., Conroy, M., Scott, T. M., Sindelar, P. T., Barber, B. R., & Orlando, A. (2010). A comparison of three types of opportunities to respond on student academic and social behaviors. *Journal of Emotional and Behavioral Disorders*, *18*, 27–40.

Higgins, J. P., Altman, D. G., Gøtzsche, P. C., Jüni, P., Moher, D., Oxman, A. D.,. . Sterne, J. A. (2011). The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *British Medical Journal*, *343*, d5928.

Hitchcock, J. H., Horner, R. H., Kratochwill, T. R., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2014). The what works clearinghouse single-case design pilot standards: Who will guard the guards? *Remedial and Special Education*, *35*, 145–152.

Holcombe, A., Wolery, M., & Gast, D. L. (1994). Comparative single-subject research: Description of designs and discussion of problems. *Topics in Early Childhood Special Education*, *14*, 119–145.

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.

Ingersoll, B. (2011). The differential effect of three naturalistic language interventions on language use in autism. *Journal of Positive Behavior Interventions*, *13*, 109–118.

Kaldenberg, E. R., Watt, S. J., & Therrien, W. J. (2015). Reading instruction in science for students with learning disabilities: A meta-analysis. *Learning Disability Quarterly*, *38*, 160–173.

Kang, M., Ragan, B. G., & Park, J. (2008). Issues in outcomes research: An overview of randomization techniques for clinical trials. *Journal of Athletic Training*, *43*, 215–221.

Kazdin, A. E. (2011). *Single-case research designs. Methods for clinical and applied settings* (2nd ed.). New York, NY: Oxford University Press.

Kennedy, C. H. (2003). The maintenance of behavior change as an indicator of social validity. *Behavior Modification*, *26*, 594–604.

King, S. A., Lemons, C. J., & Davidson, K. A. (2016). Math interventions for students with autism spectrum disorder: A best-evidence synthesis. *Exceptional Children*, *82*, 443–462.

Knight, V., Sartini, E., & Spriggs, A. D. (2015). Evaluating visual activity schedules as evidence-based practice for individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *45*, 157–178.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*, 124–144.

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

Ledford, J. R., Barton, E. E., Severini, K. E., & Zimmerman, K. N. (2017). A primer on single case research designs: Contemporary use & analysis. *American Journal on Intellectual and Developmental Disabilities.*

Ledford, J. R., & Gast, D. L. (2014). Measuring procedural fidelity in behavioral research. *Neuropsychological Rehabilitation*, *24*, 332–348.

Ledford, J. R., Gast, D. L., Luscre, D., & Ayres, K. M. (2008). Observational and incidental learning by children with autism during small group instruction. *Journal of Autism*

and Developmental Disorders, *38*, 86–103.

Ledford, J. R., Hall, E., Conder, E., & Lane, J. D. (2016). Research for young children with autism spectrum disorders: Evidence of social and ecological validity. *Topics in Early Childhood Special Education*, *35*, 223–233.

Ledford, J. R., Lane, J. D., Elam, K. L., & Wolery, M. (2012). Using response prompting procedures during small group direct instruction: Outcomes and procedural variations. *American Journal of Intellectual and Developmental Disabilities*, *117*, 413–434.

Ledford, J. R., Lane, J. D., Zimmerman, K. N., Chazin, K. T., & Ayres, K. A. (2016, April). *Single Case Analysis and Review Framework (SCARF)*. Retrieved from http://vkc.mc.vanderbilt.edu/ebip/scarf/

Ledford, J. R., & Wolery, M. (2013). Procedural fidelity: An analysis of measurement and reporting practices. *Journal of Early Intervention*, *35*, 173–193.

Losinski, M., Sanders, S. A., & Wiseman, N. M. (2016). Examining the use of deep touch pressure to improve the educational performance of students with disabilities: A meta-analysis. *Research and Practice for Persons With Severe Disabilities*, *41*(1), 3–18.

Lynch, A., Theodore, L. A., Bray, M. A., & Kehle, T. J. (2009). A comparison of group-oriented contingencies and randomized reinforcers to improve homework completion and accuracy for students with disabilities. *School Psychology Review*, *38*, 307–324.

Machalicek, W., O'Reilly, M. F., Beretvas, N., Sigafoos, J., Lancioni, G., Sorrells, A., Lang, R., & Rispoli, M. (2008). A review of school-based instructional interventions for students with autism spectrum disorders. *Research in Autism Spectrum Disorders*, *2*, 395–416.

Maggin, D. M., Briesch, A. M., Chafouleas, S. M., Ferguson, T. D., & Clark, C. (2014). A comparison of rubrics for identifying empirically supported practices with single-case research. *Journal of Behavioral Education*, *23*, 287–311.

Maggin, D. M., Briesch, A. M., & Chafouleas, S. M. (2013). Application of the what works clearinghouse standards for evaluating single-subject research: Synthesis of the self-management literature base. *Remedial and Special Education*, *34*, 44–58.

Maggin, D. M., Chafouleas, S. M., Goddard, K. M., & Johnson, A. H. (2011). A systematic evaluation of token economies as a classroom management tool for students with challenging behavior. *Journal of School Psychology*, *49*, 529–554.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gotzsche, P. C., Devereaux, P. J., … Altman, D. G. (2010a). CONSORT 2010 explanation and elaboration: Updated guidelines for reporting parallel group randomised trials. *British Medical Journal*, *340*, c869. doi:10.1136/bmj.c869

Moher, D., Schulz, K. F., Simera, I., & Altman, D. G. (2010b). Guidance for developers of health research reporting guidelines. *PLoS Medicine*, *7*(2), e1000217. doi:10.1371/journal.pmed.1000217

Reichow, B., Barton, E. E., & Maggin, D. M. (2017). Development and applications of the single case design risk of bias tool. *Manuscript under review*.

Rogers, L. A., & Graham, S. (2008). A meta-analysis of single subject design writing intervention research. *Journal of Educational Psychology*, *100*, 879.

Shamseer, L., Sampson, M., Bukutu, C., Schmid, C. H., Nikles, J., Tate, R., … and the CENT group. (2015). CONSORT extension for reporting N-of-1 trials (CENT) 2015: Explanation and elaboration. *BMJ*, *350*, h1793. doi:10.1136/bmj/h1793

Sigmundsdottir, L., Longley, W., & Tate, R. L. (2016). Computerised cognitive training in acquired brain injury: A systematic review of outcomes using the International Classification of Functioning. *Neuropsychological Rehabilitation*, *26*, 673–741.

Singh, G. (2006). Randomization made easy for small size controlled clinical trials. *Journal of the International Association of Medical Science Educators*, *16*, 75–78.

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, *17*, 510–550.

Smith, K., Shepley, S., Alexander, J., & Ayres, K. (2015). The independent use of self-instructions for the acquisition of untrained tasks for individuals with an intellectual disability: A review of the literature. *Research in Developmental Disabilities*, *40*, 19–30.

Stokes, T. F., & Baer, D. M. (1977). An implicit technology of generalization. *Journal of Applied Behavior Analysis*, *10*, 349–367. Tate, R. L., Perdices, M., McDonald, S., Togher, L., & Rosenkoetter, U. (2014). The design, conduct and report of single-case research: Resources to improve the quality of the neurorehabilitation literature. *Neuropsychological Rehabilitation*, *24*, 315–331.

Tate, R. L., Perdices, M., Rosenkoetter, U., McDonald, S., Togher, L., Shadish, W., … Vohra, S., for the SCRIBE Group. (2016a). The ***S***ingle-***C***ase ***R***eporting guideline ***I***n ***BE***havioural Interventions (SCRIBE) 2016: Explanation and elaboration. *Archives of Scientific Psychology*, *4*, 10–31.

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W., Barlow, D. H., Horner, R., … Wilson, B. (2016b). The ***S***ingle-***C***ase ***R***eporting guideline ***I***n ***BE***havioural Interventions (SCRIBE) 2016 statement. *Archives of Scientific Psychology*, *4*, 1–9.

Tate, R., Wakim, D., & Genders, M. (2014). A systematic review of the efficacy of community-based, leisure/social activity programmes for people with traumatic brain injury. *Brain Impairment*, *15*, 157–176.

Turner, L., Shamseer, L., Altman, D. G., Weeks, L., Peters, J., Kober, T., … Moher, D. (2012). Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. *Cochrane Database of Systematic Reviews*, (11), Art. No.: MR000030. doi:10.1002/14651858.MR000030.pub2

Vohra, S., Shamseer, L., Sampson, M., Bukutu, C., Schmid, C. H., Tate, R., … for the CENT group. (2015). CONSORT extension for reporting N-of-1 trials (CENT) 2015 Statement. *BMJ*, *350*, h1738 doi:10.1136/bmj/h1738

Wendt, O., & Miller, B. (2012). Quality appraisal of single-subject experimental designs: An overview and comparison of different appraisal tools. *Education and Treatment of Children*, *35*, 235–268.

What Works Clearinghouse. (2013). *What works clearinghouse procedures and standards handbook (Version 3.0)*. Washington, DC: Institute for Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/DocumentSum.aspx?sid=19

Wolery, M. (2013). A commentary: Single-case design technical document of the what works clearinghouse. *Remedial and Special Education*, *34*, 39–43.

Wolery, M., & Billingsley, F. F. (1982). The application of Revusky's Rn test to slope and level changes. *Behavioral Assessment*, *4*, 93–103.

Wolery, M., & Gast, D. (1984). Effective and efficient procedures for the transfer of stimulus control. *Topics in Early Childhood Special Education*, *4*, 52–77.

Wolery, M., Holcombe, A., Werts, M. G., & Cipolloni, R. M. (1993). Effects of simultaneous prompting and instructive feedback. *Early Education and Development*, *4*, 20–31.

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

Zimmerman, K. N., & Ledford, J. R. (2017). Evidence for the effectiveness of social narratives: Children without ASD. *Journal of Early Intervention, 39,* 199–217.

Zimmerman, K. N., Ledford, J. R., Severini, K. E., Pustejovsky, J. E., Barton, E. E., & Lloyd, B. P. (2017). A Comparison of Methods to Evaluate Quality and Rigor when Synthesizing Single Case Research Designs. *Under Review.*

14
Synthesis and Meta-analysis of Single Case Research

Mariola Moeyaert, Kathleen N. Zimmerman, and Jennifer. R. Ledford

Important Terms

synthesizing, meta-analysis, effect size, gray literature, data extraction, autocorrelation, heteroscedastic overlap metrics

Conducting one single case design (SCD) study answers a very narrow question; thus, like all research, the value of each study must be considered in the context of the literature as a whole. This is why, as discussed in Chapter 3, it is important to contextualize a study you are conducting relative to similar studies previously conducted. It is also why, in an attempt to answer a somewhat broader question, researchers are interested in **synthesizing** outcomes across studies. By synthesizing, we mean reviewing results of similar studies and drawing broad conclusions about the state

of the evidence. One way to synthesize outcomes is via **meta-analysis**, which is "a technique for encoding and analyzing the statistics that summarize research findings as they are typically presented in research reports" (Lipsey & Wilson, 2001, p. 2). Conceptualization of meta-analysis is based on the "typical" between-groups research design. Since visual analysis is the primary analysis technique for SCD data, meta-analysis cannot be easily modified to "fit" SCD data. Currently, researchers are pursuing attempts to develop and validate procedures to meaningfully combine data from SCD studies. This chapter will review some of those attempts, but first, we address the purposes of outcomes synthesis in general, guidelines for performing narrative reviews, synthesizing studies based on visual analysis, and data extraction. As discussed in Chapter 13, we emphasize that synthesis of outcomes, including meta-analyses, are generally not helpful in the absence of systematic assessment of the rigor of studies.

Purposes of Summative Evaluation of Outcomes

Summative evaluations of outcomes, and the development of quantitative metrics, have generally been conducted with the purpose of combining outcomes across multiple studies; however, some have advocated for the reporting of quantitative effect size metrics for single studies. We think this is misguided for several reasons: (a) many of the metrics that are the most widely- used and touted as "effect sizes" have considerable weaknesses (see sections below) and may mislead readers, (b) SCD researchers report all data in graphical format, and thus statistics are not needed for interpretation of single studies because it is visualized in the graphical representations, (c) none of the metrics align with determination of experimental control which is the primary question answered via visual analysis, and (d) many common SCD data patterns do not align well with available metrics (e.g., have trends, include few data points, are auto-correlated). Thus, simply because we *can* report quantitative metrics for individual studies does not mean that we *should*. The fact that all data are reported and easily extracted (see below) suggests that if a reader of an SCD study thought that a quantitative metric would be helpful, he or she could independently calculate it. We assert that, at present, *the best summative evaluation of outcomes in a single study is a thorough and systematic explanation of a thorough and systematic visual analysis procedure*. See Appendix 14.1 for a published example of a systematic process for drawing conclusions regarding functional relations.

Researchers and practitioners are often interested in the effects of a given intervention not in a single study, but given *all* of the available research. This interest, evidenced by the number of reviews and syntheses over time, has increased in recent years (Maggin, O'Keeffe, & Johnson, 2011), which is reasonable given the increasing body of studies available for review. Historically, across-study comparisons in SCD research were made via narrative reviews alone; recently, increasingly frequent attempts have been made to develop statistical analyses for this purpose. Using visual analysis and statistical analyses for drawing conclusions *across* studies present equally complex but different problems. However, given the need to assess whether practices are evidence-based or research-supported, the difficulties of synthesis should not preclude attempts to meaningfully and accurately summarize findings in a field of study. Pustejovsky and Ferron (2017) provide three reasons for careful consideration of using syntheses of outcomes across studies: (1) they are critical tools for guiding decision-making regarding evidence-based practices, (2) they can be helpful in identifying variation in treatment effects in different contexts or for different participants, and (3) they can contribute to advancements in SCD methodology.

We will provide suggestions for using both visual and statistical analyses in this chapter; readers should note that we continue to consider visual analysis the primary method by which SCD studies should be analyzed, with the use of secondary statistical

analyses, as appropriate.

Narrative Reviews

Research Questions

Regardless of your use of outcome metrics, all sound syntheses *must* begin with a sound narrative review. First, you should carefully consider your research question. One structured way to do this is to use the PICOS criteria (O'Connor, Green, & Higgins, 2008). Using these criteria will help you to formulate a specific research question, and will also assist in the development of a list of search terms to obtain a comprehensive but appropriately narrowed search. First, determine for what *Participants* you are interested in assessing intervention effects (e.g., school-aged children with ASD who engage in problem behavior [Severini, Ledford, & Robertson, 2017], young children with disabilities [Barton & Wolery, 2008], students with challenging behavior who had no identified disabilities or high incidence disabilities [Maggin, Johnson, Chafouleas, Ruberto, & Berggren, 2012]). Then, determine the specific *Intervention* you are interested in assessing; this may be broad (e.g., social skills interventions; Ledford, King, Harbin, & Zimmerman, 2017), or more narrow (e.g., system of least prompts; Doyle, Wolery, Ault, & Gast, 1988). Importantly, you should consider under what names the intervention might be used (e.g., researchers might use all of these terms relatively interchangeably: naturalistic teaching strategies, incidental teaching, enhanced milieu teaching; Lane, Lieberman-Betz, & Gast, 2016); this will be important for ensuring a comprehensive search. You should also consider the *Comparators*, or as we would more commonly refer to it, the condition to which the intervention is compared (e.g., baseline). For all narrative reviews, but especially when synthesizing outcomes, you should ensure that you are making reasonable comparisons across studies. For example, in one review of group contingencies in early childhood settings (Pokorski, Barton, & Ledford, 2017), many of the studies included primary comparison conditions of two different types of group contingencies (e.g., a multitreatment design with B and C conditions); a determination of the magnitude of differences between these comparisons (e.g., between types of group contingencies) should not be reasonably combined with those of a study including a baseline to group contingency comparison. These types of studies answer two different questions (i.e., Do group contingencies interventions work? and Does one group contingency work better than another?); combining outcomes results in a metric that does not adequately answer either. *Outcomes* of interest should also be identified a priori; you should specifically identify dependent variables you are interested in. These can be broadly defined (e.g., social skills, problem behavior) but can also be more specific (e.g., spoken language, self-injury). Finally, you should designate the *Study designs* that will be considered for your review. It is generally helpful to include both SCD and between groups design that are experimental in nature, although separate analyses will be needed for each group. However, given a specific dependent variable of interest, it is

likely that many reviews will include *either* randomized controlled trials *or* experimental SCDs. For example, if you are interested in teacher report of problem behavior, you are likely to find a number of between group comparisons that experimentally evaluate the effects of interventions on this dependent variable (DV); if you are interested in direct measures of problem behavior in children with ASD, you are likely to find primarily SCD studies that experimentally evaluate the effects of interventions on this DV. There are theoretical and practical reasons for not combining results across these types of different DVs (i.e., because participants serve as their own controls, "treatment effects" have different meanings in SCD and group comparison research); but this is typically not an issue, as it is unlikely that many bodies of evidence exist that measure the same DV and use both between-groups and SCD designs.

Literature Search

After identifying your research question, it is essential to conduct an exhaustive search of the literature to identify all relevant studies *and* to accurately, systematically, and thoroughly document these procedures. The PRISMA guidelines (discussed in Chapter 3) for search reporting are intended to assist researchers in adequately reporting replicable search results; a sample PRISMA flow chart is also presented in Chapter 3. Given differences in databases and the difficulties associated with exact replications of search procedures (for a discussion and example, see Lemons et al., 2016), it is prudent to not only use and report search terms adequately, but to hand search potentially relevant sources (e.g., journals in the area of interest). SCD reviews have historically included primarily peer-reviewed sources, with the rationale that these sources are likely to be of higher quality than unpublished sources (e.g., the unpublished sources were not published given some fatal flaw in study design or implementation). However, given the likelihood of publication bias (Shadish, Zelinsky, Vevea, & Kratochwill, 2016; Tincani & Travers, 2017), researchers should consider including **gray literature**; for review purposes, this generally means including conference abstracts, dissertations, and theses in your search procedures. Note that using a secondary coder and calculating interobserver agreement (IOA) for search procedures reduces the likelihood of error and improves the likelihood that you have developed a replicable search including all relevant sources.

After identifying all articles that meet your PICOS criteria, you should determine whether articles also meet your minimal internal validity criteria for inclusion. For example, some reviews include only studies that meet minimal standards (e.g., three potential demonstrations of effect; a comprehensive review) while others only report outcomes for those with adequate internal validity (sometimes called a best-evidence synthesis; Reichow & Volkmar, 2010).

Coding

Locating all of the relevant sources for a comprehensive review can be a time-consuming part of the synthesis process. The next step is to evaluate each study. We propose three levels of evaluation for reviews including SCD studies: (1) study characteristics, (2) study rigor, and finally (3) study outcomes. All three levels of evaluation should be coded according to specified rules, and a secondary coder should code a proportion (e.g., 20–33%) of the identified studies to ensure reliability. The first step of coding is to identify what information should be coded at each level; the second is to write and pilot a detailed coding manual; the next is to code all studies using the coding manual, including secondary coding and IOA analysis; and the final is to summarize information. Developing the coding manual should be an iterative process. That is, you should develop the codes and definitions, code a few studies, decide whether the coding process accurately and comprehensively portrays the studies' similarities and differences, make changes to the code, and repeat if needed.

Coding study characteristics varies by research question. Examples of items that you might want to code include study-level information like design type, primary dependent variable, number of participants, measurement systems used, setting, implementers, social partners, and arrangement (e.g., small group or 1:1). Depending on how broad your intervention is, you may also need to code intervention components or variations (cf. Ledford et al., 2017). The dynamic nature of SCDs necessitates coding any modifications made to the intervention based on participant data (i.e., you may need to code what changes were made if any participants did not respond to the original intervention). In addition to study-level characteristics, you should also code participant-level characteristics. Again, specific codes will depend on your research questions, but might include: race and ethnicity, gender, age, school placement, diagnoses, pre-baseline assessment information, and reasons for inclusion in the study.

Following coding for study characteristics, you should develop a systematic coding system for rigor, or use a previously developed system (such as one of those described in Chapter 13). The purpose of this systematic coding is two-fold: (a) to determine to what extent individual studies are sufficiently rigorous to be confident that outcomes are causally related to the intervention, and (b) to determine what aspects of rigor and quality are well-represented or missing from a specific group of studies for the purposes of providing suggestions for improving future research. For example, in a group of studies related to sensory-based interventions, we might find that better outcomes (behavior change for participants) are associated with studies with a high risk of bias; and null outcomes (no behavior change) are associated with studies with a lower risk of bias. This would suggest that the *rigorous* evidence we have suggests the interventions are not effective and might provide evidence for the need for more highly rigorous research. Similarly, we might find that larger effects are noted for studies using percentage of intervals rather than counts (a worrisome finding given the properties of interval systems discussed in Chapter 5). These are critical findings, and provide considerably more useful information than simply reporting that outcomes were variable across studies.

Synthesizing Across Studies Using Structured Visual Analysis Guidelines

After coding descriptive information and analyzing internal validity, you should determine for which studies it is reasonable to synthesize outcomes. For example, you may decide that at minimum, studies must meet the WWC standards *with reservations*, must include 80% of the CEC QIs, or must have an average rigor/quality score of at least 2.0 on the SCARF tool (see Chapter 13 for descriptions of these tools). This ensures that you are only assessing outcomes for studies that are sufficiently rigorous for making causal conclusions. Alternatively, you may decide to include studies with low rigor to compare outcomes from studies likely to include bias to more rigorous studies (as described above). In this case, we urge you to explicitly state that outcomes from these less rigorous studies should not be used to draw causal conclusions.

After determining which studies will be included in your outcomes analysis, you should use specific visual analysis guidelines (see Appendix 14.1 for an example) to determine whether a functional relation is present for each design. It is reasonable to develop and pilot your guidelines on non-included studies including training a second observer and ensuring that there is agreement on functional relation determinations. In addition to a "yes/no" determination regarding functional relation, you may also want to classify functional relations as small, medium, or large (based on systematic and pre-specified rules) or classify your confidence in the relation (see Appendix 14.1). Prior to coding the selected set of studies, you should develop, pilot, and modify (as needed) systematic and transparent visual analysis procedures (cf. Common, Lane, Pustejovsky, Johnson, & Johl, 2017).

In one example of a synthesis in which visual analysis was used, Severini, Ledford, and Robertson (2017) conducted a review of interventions designed to reduce problem behavior for individuals with autism in school settings. They independently visually analyzed 100% of studies included in the review and compared both functional relation decisions (yes/no) and confidence rankings (1–4, *not at all confident* to *extremely confident*). Agreement for yes/no decisions was 98%, while exact agreement on confidence rankings was 93%. High agreement scores for visual analysis suggest that the authors were using the same visual analysis rules, improving confidence in decisions and likely replicability. Authors reported the percentage of designs in which a functional relation was demonstrated. Similarly, Ledford et al. (2017) reported a "success rate" (percentage of studies demonstrating a functional relation according to visual analysis) for specific social skills intervention components.

There are a few notable drawbacks to synthesizing outcomes via visual analysis. The first is that developing and using systematic visual analysis for a large body of studies may be perceived as challenging; some studies have shown agreement for visual analysis may be low, especially under certain conditions (e.g., variability; Kahng et al., 2010; Matyas & Greenwood, 1990). However, the reviews mentioned above show that it is possible to have high agreement between observers. We will note that in those two cases, one included an expert visual analyst (PhD, BCBA-D, co-editor of this book) and a

student who she trained; the other included two expert visual analysts (PhDs, BCBA-Ds). This suggests that some level of expertise may be helpful when synthesizing studies using visual analysis; this should not be surprising, and is similar to the expectation that an experienced statistician is needed to synthesize and/or interpret data using statistical methods. Several visual analysis training tools are available, including ones freely available online (e.g., www.singlecase.org); if both analysts "pass" a training or otherwise agree on a set of graphs (e.g., piloted procedures) before analyzing the data, it is more likely that agreement will be acceptable for the review. Regardless, piloting and testing your systematic procedures for visual analysis and carefully reporting these procedures is important.

The second potential drawback to synthesizing outcomes via visual analysis is that there are no individual or omnibus "effect sizes" generated and no standard errors can be calculated (prohibiting "weighting" of data). Thus, even when a success rate is reported (e.g., functional relations were present in 90% of designs), there is generally not an indication regarding the extent to which the average effect was "big" or "small" and whether this size varied across studies, participants, or other characteristics. This is consistent with SCD rationale (e.g., consistency of effect, not size of effect, is critical in determining presence of a functional relation), but may make it difficult to convey to stakeholders how much behavior change occurred and to what extent that behavior change was different across intervention types. Some have also argued that Type I error rates (concluding a treatment effect exists when one does not) is unacceptably high in SCD research; sufficient training and using expert visual analysts may decrease this risk (e.g., expert agreement is often high; Ledford & Wolery, 2013).

Some people have argued that another critical drawback of visual analysis of SCD data for the purposes of synthesis is that it cannot be combined with syntheses of between-group designs. We do not think this is a critical flaw, for two reasons: (1) the logic is different for SCD and between-group designs, and (2) the types of dependent variables measured in SCD and between-group designs is almost always different (Yoder, Bottema-Beutel, Woynaroski, Chandrasekhar, & Sandbank, 2013). We argue that this difference makes combining outcomes across SCD and between-groups studies unreasonable, regardless of metric.

Extracting Data

In SCD studies, data for individual cases are presented via graphs rather than with summary statistics such as means and standard deviations; this is a considerable advantage for data analysts. Before synthesizing outcomes using quantitative metrics, data in published SCD studies must first be extracted to obtain exact coordinate values that can be used for calculating quantitative metrics. **Data extraction** is a multi-step process for determining and recording the X and Y values of data points. WebPlotDigitizer (Rohatgi, 2014), a free data extraction program that can be used on multiple platforms, was evaluated and determined to be a reliable and valid program for extracting graphed data to use in SCD syntheses; other free and costly programs also exist (e.g., PlotDigitizer; for a review, see Moeyaert, Maggin, & Verkuilen, 2016).

After determining which data extraction program to use and downloading (or otherwise obtaining) it, you should create single images of each graph to be included in the synthesis. To do this, you can use the "screen shot" or "page capture" function of an article PDF and save each graph image as a GIF, JPEG, or PNG file with identifying information about the author, year, and dependent variable presented in each graph. Once you have saved the image file, it should be uploaded into a data extraction program that will obtain X-Y values. Generally, you begin by assigning locations and values to each axis (e.g., you "tell" the program where the axis is and what the minimum and maximum values are). Then, you individually select each data point by "clicking" on it with a mouse. Consistently using a higher "zoom" (e.g., 200%) may result in more accurate extractions and minimize variability due to minor errors in placement. Although only the ordinate values will be used to quantify outcomes, the abscissa values should also be saved in a spreadsheet to allow for straightforward comparisons with graphed data and subsequent error correction, if needed. When extracting data from studies using rapid iterative alternation designs (e.g., ATD, AATD), extract data from one data path first, record values in a spreadsheet, and then extract values from the second data path. For time-lagged designs with multiple tiers, save each tier as a distinct image and extract data for each separately. A screenshot of the program is shown in Figure 14.1, illustrating the graph and related x and y values. A screenshot of the data transferred to a spreadsheet appropriate for use to calculate effect sizes is shown in Figure 14.2.

Although there are not universal standards for reporting the data extraction process in syntheses and meta-analyses, methods should be reported with sufficient detail to allow for replication (cf., Rakap, Snyder, & Pasia, 2014). Reliability data should be collected and methods for resolving disagreements or inaccuracies should be reported. Finally, decisions regarding rounding extracted values can be made based on the studies included in the review and should be articulated in final reports. For example, if all studies in a review used interval recording procedures and report total session length, you can

calculate the exact possible values of an outcome and round extracted values to the nearest possible value (e.g., if each session contained 20 intervals, the only possible values are 0%, 5%, 10% [etc.], so a value of 4.89% could be rounded to 5%). Once all data included in the synthesis have been extracted and organized by study and comparison (e.g., tier), quantitative metrics can be calculated. When calculating quantitative metrics, small errors in data extraction (e.g., selecting 1.1 for a data value when the actual value is 1.0) can lead to drastic changes in calculated overlap of data points (Overlap Metrics, below) and relative change (see Log Response Ratio, below). See Appendix 14.2 for a sample worksheet that can be used to guide decision-making about data extraction.

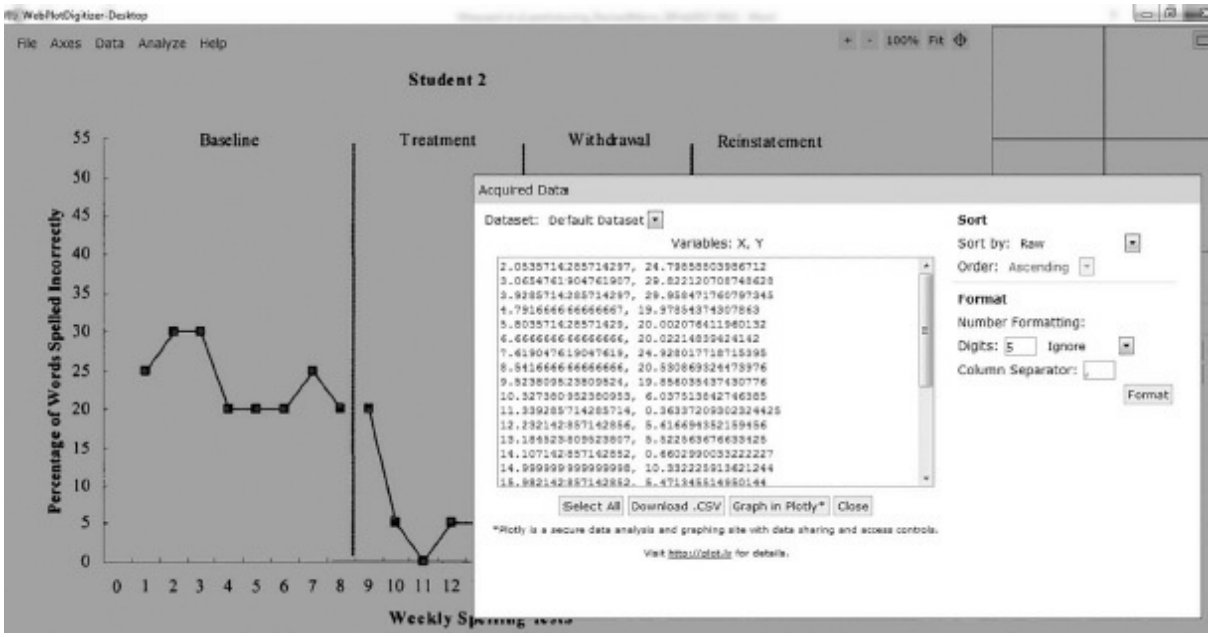


Figure 14.1 Display of the WebPlotDigitizer environment.

	A	B	C
1	Study	Case	Outcome
2	1	1	86
3	1	1	91
4	1	1	58
5	1	1	88
6	1	1	20
7	1	1	25
8	1	1	31
9	1	2	34
10	1	2	87
11	1	2	76
12	1	2	81
13	1	2	32
14	1	2	34
15	2	1	78
16	2	1	98
17	2	1	88
18	2	1	26
19	2	1	24

Figure 14.2 Display of retrieved data from primary SCD studies.

Meta-Analysis of Research Outcomes

The quantitative integration of a large body of results from primary research articles can contribute to practice, research, and theory (Jenson, Clark, Kircher, & Kristjansson, 2007; Manolov & Solanas, 2013; Parker & Brossart, 2003). Meta-analysis is a statistical analysis technique, used often for group-comparison design data, which serves this purpose. It results in the combination of outcomes from several studies addressing the same underlying research question (Cooper, 2010; Glass, 1976) in an objective and systematic way. Benefits of meta-analysis of SCD data include (a) average treatment effects reported in primary studies can be estimated with greater power (i.e., the extent to which a true treatment effect can be identified), and (b) variation in the overall treatment effect between studies can be explored and explained (e.g., differences in magnitudes of effect between students with autism and those with intellectual disabilities).

Before being combined, study results should be converted to a common standardized effect size measure, which makes it possible to compare and synthesize results across similarly focused studies (Lipsey & Wilson, 2001). By **effect size**, we mean an estimation of the overall magnitude of behavior change. For instance, Shogren, Faggella-Luby, Bae, and Wehmeyer (2004) conducted a meta-analysis summarizing nine SCD studies (with a total of 21 cases) all investigating the influence of student choice as a treatment to decrease problem behaviors. However, challenging behavior was not measured on the same outcome scale across studies. In one of the studies (Cole & Levinson, 2002), challenging behavior was measured as the percentage of steps that included the behaviors, on a scale from 0 to 100. In another study (Dibley & Lim, 1999) challenging behavior was reported as a count, with a range from 0 to 20. Therefore a standardized effect size is needed to allow comparison of scores across different studies.

In group-comparison studies, there is widespread agreement about how these standardized effect sizes should be expressed, what the statistical properties of the estimators are (e.g., underlying sampling distribution), and how to translate from one measure (e.g., a correlation) to another (e.g., *Hedges' g*). However, group-comparison methods generally involve only one (post-test only) or two (pre-test/post-test) measurements of participant response. Therefore, in the meta-analysis of group-comparison designs, important information on the dynamic nature of participant response to treatment is missed—individual client responses are lost in the group averaging process and important findings may be obscured. Inferences about causes of changes (when they can be made) are made at the level of the group, which neglect effects of the intervention on any individual participant. The particulars of who responded to an intervention under which conditions might be obscured when reporting only group means and associated effect sizes (Horner et al., 2005). This limits the applicability of results to specific clients (Barlow, Nock, & Hersen, 2009).

Summarizing single-case experimental design studies, on the other hand, afford the

researcher an opportunity to provide detailed documentation of the characteristics of participants who responded and those who did not (i.e., non-responders). By meta-analyzing SCD studies, an overall average treatment effect across participants and across studies is obtained in addition to detailed information about variations in the treatment effect related to specific participants under investigation. SCDs allow behaviors of the individual to be measured at various points in time, thereby allowing the treatment effect to be evaluated with more than a single observation, which allows meta-analysts to summarize how the treatment effect changes over time (i.e., identifying trends). As a consequence, important research questions can be addressed (which cannot be answered by summarizing group-comparison designs) such as: (1) What is the magnitude of the average treatment effect across cases and across studies? (2) How does the effectiveness of the treatment change over time across cases and across studies? (3) What is the magnitude and direction of the case-specific treatment effect and trends? (4) How much do the treatment effect and trends vary within cases, across cases and/or across studies? and (5) Does a (case and/or study level) predictor influence the treatment's effect? Because synthesis of outcomes has the potential to answer these interesting questions, there is a growing interest in synthesizing SCD studies.

Meta Analysis of SCDs

The field of meta-analyses of SCDs is much less well-developed compared to group-comparison designs, and there are no agreed-upon methods or standards for effect size estimation (Kratochwill et al., 2010). This is due to the complex nature of SCDs. Shadish and Sullivan (2011), among others, noted that modeling SCD data at the case level presents a substantial challenge. Based on previous review studies summarizing characteristics of SCDs (Beretvas & Chung, 2008; Ferron, Farmer, & Owens, 2010; Maggin, O'Keeffe, & Johnson, 2011; Schlosser, Lee, & Wendt, 2008; Shadish & Sullivan, 2011), the following challenges have been identified:

1. Observed data series tend to be fairly short, with lengths of 30 or fewer data points being common. Ferron et al. (2010) found a median series length of 24 and according to the survey of Shadish and Sullivan (2011), the number of data points per case ranged from 2 to 160, with median and mode equal to 20, and 90.6% of the cases having 49 or fewer data points.
2. SCD data are serially dependent, exhibiting small to modest autocorrelation. **Autocorrelation** refers to the lack of independence in SCD data, such that data points that are closer together in time are more similar than those that are farther apart in time. Shadish and Sullivan (2011) noted that the average autocorrelation is close to but significantly different from zero and significantly heterogeneous (i.e., the degree to which autocorrelation occurs may be dependent on factors such as the dependent variable of interest and the time between two consecutive measures).
3. SCD data may have an outcome scale that differs from case to case (e.g., percentages, counts, interval scale).
4. SCD data have heterogeneous outcomes, including continuous, rates or counts, or ordinal outcomes. Rates and counts are especially common (Shadish & Sullivan, 2011). This in turn implies data are likely to be **heteroscedastic** (variability differs between baseline and intervention conditions), have notable apparent outliers, or notable floor/ceiling effects. Standardization is more difficult for these outcomes.
5. Researchers are generally interested in research questions that have been answered using a variety of condition ordering types (e.g., time-lagged designs such as multiple baseline designs and sequential introduction and withdrawal designs such as A-B-A-B designs). Each design type must be handled separately with an appropriate design matrix (for additional information, see Moeyaert et al., 2014b).
6. Data for some cases may have anomalies such as gaps in time (Ferron, Moeyaert, Van den Noortgate & Beretvas, 2014; Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013a).

These features represent substantial difficulties in developing a universal standardized effect size measure that considers all of the SCD complexities. During the last decade, a variety of different effect sizes measures (ranging from simple to more complex) have been proposed and empirically validated. Each has flaws, but some methods are likely to be more useful than others (depending on the research question and specific SCD data and design characteristics, Manolov & Moeyaert, 2017). The field of meta-analysis of SCDs is still under development and rapidly changing. Thus, although we share some contemporary methods for conducting meta-analysis of SCD work, we expect that researchers interested in these methods will need to closely follow published research in the area over time. Current methods for synthesizing SCDs can be categorized based on three approaches: comparing overlap across conditions (overlap- based metrics), comparing means across conditions (mean-based metrics), and modeling data patterns across conditions (regression-based approaches). Each approach will be briefly summarized below, including recommendations for using each method when meta-analyzing outcomes from SCDs. A guidance document to assist readers in planning and conducting an SCD synthesis or meta-analysis can be found in the web-based resources for this chapter.

Overlap-Based Metrics

Overlap metrics (also called nonoverlap statistics) are one type of quantitative index suggested for use to synthesize effects across SCDs. Designed for use in conjunction with visual analysis to purportedly describe the magnitude of intervention effectiveness (Vannest & Ninci, 2015), **overlap metrics** measure the degree of nonoverlap between adjacent conditions via pairwise comparisons or evaluations of overlap in the highest or lowest levels of a condition (Parker, Vannest, & Davis, 2011). Authors of the overlap metrics argue they are feasible for use as measures of effectiveness since they are nonparametric statistics designed to avoid assumptions associated with parametric statistics that may not be met given the parameters of SCD data (e.g., independence, normality; Parker et al., 2011). Specifically, creators of the measures argue overlap metrics provide a standardized metric of behavior change that may be aggregated across studies to produce an overall magnitude of intervention effectiveness (Parker et al., 2011; Vannest & Ninci, 2015). Additionally, many overlap metrics are easy to calculate using paper and pencil or online calculators (Pustejovsky, 2016d), increasing likelihood of use (cf. Whalon, Conroy, Martinez, & Werch, 2015). Given the ease of calculation, and that one characteristic assessed via visual analysis is between-condition overlap, it is not surprising that overlap metrics are the most frequently used quantitative statistics in behavioral research (Maggin, O'Keefe, & Johnson, 2011).

Purposes and Examples of Overlap Metrics

Many overlap metrics have been created to quantify the amount of nonoverlap in graphed data as a measure of intervention effectiveness. These include:

- Percentage of non-overlapping data (PND; Scruggs, Mastropieri & Castro, 1987)
- phi coefficient (Rosenthal, 1994; Burns, Zaslofsky, Kanive, & Parker, 2012)
- Percentage of data points exceeding the median (PEM; Ma, 2006)
- Percentage of nonoverlapping data (PAND; Parker, Hagan-Burke, & Vannest, 2007)
- Improvement rate difference (IRD; Parker, Vannest, & Brown, 2009)
- Nonoverlap of all pairs (NAP; Parker & Vannest, 2009)
- Pairwise data overlap squared (PDO^2; Parker & Vannest, 2007)
- Kendall's tau for nonoverlap between groups (Tau_{novlap} or Tau; Parker, Vannest, Davis, & Sauber, 2011)
- Percentage of nonoverlapping corrected data (PNCD; Manolov & Solanas, 2009)
- Tau-U (Parker et al., 2011).

Each metric uses a different approach to calculate and quantify the amount of

nonoverlap between conditions (see Parker et al., 2011 for a detailed review).

Drawbacks of Overlap Metrics

Evaluations have noted considerable weaknesses of using overlap metrics, some of which have been recognized by authors and resulted in the creation of novel overlap metrics (e.g., Tau-U; Parker et al., 2011). However, most weaknesses detailed below are problematic for all or most of the overlap metrics. They include: (a) failure to account for replication, (b) procedural sensitivities, and (c) mischaracterization as an estimate of magnitude.

Overlap metric outcomes may not align with conclusions drawn via visual analysis (Barton et al., 2016; Ledford et al., 2016, 2017; Ma, 2006; Wolery et al., 2010; Zimmerman et al., 2017). Inconsistencies between overlap metrics and visual analysis are likely to occur given the failure of overlap metrics to account for all characteristics of data (e.g., stability, variability, consistency, immediacy of change; Barton et al., 2016; Wolery et al., 2010), magnitude of differences, or consistency of effects (replication logic; Wolery et al., 2010). However, in one study, the NAP metrics (and the associated small, medium, and large effects) aligned with visual analysis results (Severini, Ledford, & Robertson, 2017). Although two overlap metrics are reported to account for trend (PEM-T and Tau-U), further investigations suggest they might be associated with a higher likelihood of error compared to visual analysis (Tarlow, 2016; Wolery et al., 2010).

In addition to providing information on only one aspect of SCD data, overlap metrics are also sensitive to procedural variations such as the number of data points in a condition (Tau-U; Pustejovsky, 2016c). Specifically, values of overlap metrics can be increased by changing the length of the baseline (Pustejovsky, 2016b) or treatment condition (Pustejovsky, 2016b; Wolery et al., 2010). Overlap metrics are also influenced by other procedural variations such as design type (Chen et al., 2016; Pustejovsky, 2016b), session length (Pustejovsky, 2016b), behavior of interest (Pustejovsky, 2016b), and measurement system (Ledford, Lane, Ayres, & Lam, 2014; Pustejovsky, 2015; Pustejovsky, 2016b; Pustejovsky & Ferron, 2017). As a result, conclusions regarding the magnitude of the effectiveness of an intervention may be attributable to characteristics of the design, amount of data, or measurement system used rather than true behavior change that occurred as a result of the treatment. More alarmingly, the sensitivity to the number of data points in a condition could result in the potential for researchers to manipulate condition lengths to increase the size of an overlap metric (Pustejovsky, 2016b; Pustejovsky & Ferron, 2017; Wolery et al., 2010).

Variability in conclusions regarding the effectiveness of interventions using overlap metrics may result from the failure of indices to discriminate the *magnitude* of change between conditions when visual inspection of the data suggest differential effectiveness is present across designs (Campbell, 2012; Chen et al., 2016; Ma, 2006; Rakap, Snyder, & Pasia, 2014; Wolery et al., 2010). Comparisons of overlap metrics to each other and visual analysis yielded (a) no discrimination between very small to very large magnitudes of

effect (Campbell, 2012, Chen et al., 2016; Wolery et al., 2010), (b) no differentiation between non-effects and contra-therapeutic effects (Chen et al., 2016; Zimmerman et al., 2017), and (c) 100% nonoverlap values for graphs with varying magnitudes of effectiveness as determined by visual analysis (Wolery et al., 2010). Figure 14.3 displays four simulated graphs with perfect overlap scores (e.g., no data points overlap across adjacent conditions); however, considerable differences exist in the visual analysis conclusions regarding both the presence of behavior change and the magnitude of effect. Similar to the presented simulation data, recent reviews of speech generating device interventions (Chen et al., 2016), parent-implemented functional assessment-based interventions (Barton et al., 2016), and social skills interventions (Ledford et al., 2017) note overlap metrics did not accurately capture the range of effectiveness present in reviewed studies.

Evaluations suggest some overlap metrics are highly correlated with other overlap metrics (Chen et al., 2016; Parker & Vannest, 2009). Despite their correlations with each other, conclusions regarding the overall effectiveness of an intervention have been demonstrated to vary based on the overlap metric selected (Chen et al., 2016; Pustejovsky, 2015; Rakap, Snyder, & Pasia, 2014; Zimmerman et al., 2017). This could be considerably problematic, introducing bias if authors "shopped" for a metric that indicated larger or smaller effects.

Recommendations Regarding Overlap Metrics

Given the limitations, we recommend overlap metrics not be used to synthesize SCD since they are not accurate indicators of the presence of a functional relation *and* are not estimates of magnitude. At minimum, if they are used, we strongly caution against drawing causal conclusions *or* reporting overlap metrics as *effect sizes*. When conducting meta-analytic summaries, overlap metrics cannot be combined with effect sizes from group design studies (Wolery et al., 2010) and should not be used to aggregate outcomes across studies with different measurement systems (Pustejovsky, 2015). Furthermore, sampling variability (e.g., standard errors and confidence intervals) cannot be estimated for all overlap indices (Pustejovsky, 2015, 2016a, 2016c; Tarlow, 2016), thus prohibiting the use of fixed-effect meta-analytic syntheses (Pustejovsky, 2015) and potentially increasing the likelihood of Type I error (Tarlow, 2016). Overlap metrics should not replace visual analysis as a tool to describe behavior change in syntheses of SCD studies (Barton et al., 2016; Ledford et al., 2016; Ledford et al., 2017; Rakap, Snyder, & Pasia, 2014), although they can be helpful for describing *overlap* between conditions.

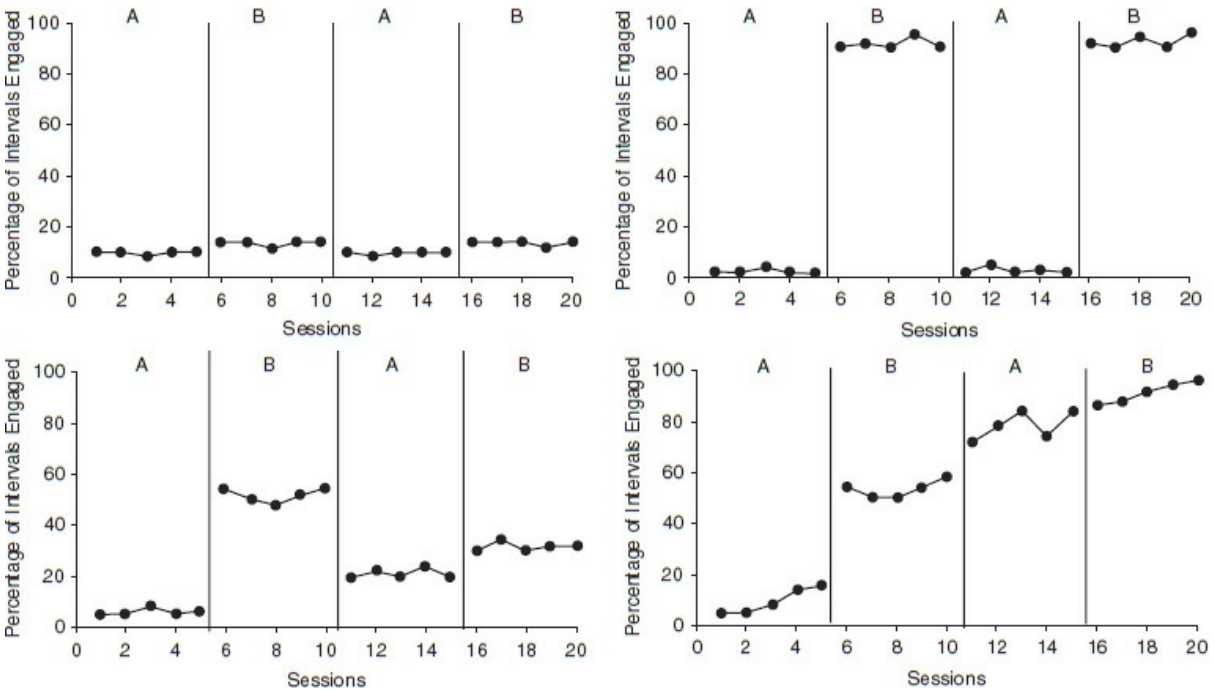


Figure 14.3 Simulated data across four A-B-A-B withdrawal designs demonstrating variability in the magnitude of effects despite identical values of nonoverlap. Each design has 100% nonoverlap of data points between each A-B comparison. The top two graphs show that overlap values do not assess magnitude of change (i.e., are not "effect sizes"). The bottom two graphs show that data patterns indicating potential threats to internal validity (inconsistent effects, maturation, history) are not accounted for using these metrics.

Combining Overlap-Based and Descriptive Indices

Simple averages across studies can be calculated from overlap indices (e.g., PND, NAP, Tau, IRD, PNCD) reported in the initial primary study. Similarly, you can combine descriptive indices quantifying changes in level and slope (such as slope and level change [SLC]; Solanas, Manolov, & Onghena, 2010), mean phase difference (MPD, Manolov & Solanas, 2013), percentage change index (PCI; Hershberger, Wallace, Green, & Marquis, 1999), mean baseline reduction (MBLR, Campbell, 2003), and percentage of zero data (PZD; Scotti, Evans, Meyer, & Walker, 1991). Although these techniques are very simple and do not require standardization (as they are percentage indices), they are not recommended because a sound method for performing a meta-analysis on non-overlap indices is not developed (e.g., Schlosser et al., 2008). Moreover, simply calculating an average of overlap or non-overlap does not include a consideration of precision, and there is no weight assigned. This means that all studies are weighted equally, regardless of the number of participants or data points. Moreover, as described above, those average overlap or non-overlap indices do not reflect the size of the effect and thus are difficult to interpret.

Mean-based Metrics

Mean-based metrics are an alternative approach to synthesizing SCD outcomes that provide traditional standardized effect size measurements of the magnitude of behavior change (Shadish, Hedges, Horner, & Odom, 2015). Mean-based metrics can be used to evaluate within-case effects (cf. log response ratio; Pustejovsky, 2017) and between-case effects (cf. between-case standardized mean difference; Hedges, Pustejovsky, & Shadish, 2012, 2013). Within-case effect sizes are used to calculate the magnitude of change in the dependent variable by analyzing effects within a single case, whereas between-case effect sizes are used to calculate the magnitude of change in the dependent variable across multiple cases.

Log Response Ratio

The log response ratio (LRR) is a mean-based effect size index that quantifies the magnitude of behavior change between two adjacent conditions (Pustejovsky, 2017). LRR is calculated using a natural logarithm transformation to quantify effects as a proportion of change between conditions; the transformation allows outcomes to be evaluated on a scale that is less restricted than a typical ratio (Pustejovsky & Ferron, 2017). Designed to model effects measured via direct, systematic observational recording (see Chapter 5), LRR is most appropriate for use with dependent variable outcomes using a ratio scale where a score of zero indicates the absence of an outcome (e.g., count, percentages, rates, and continuous duration recording; Pustejovsky, 2017; Pustejovsky & Ferron, 2017). LRR can be calculated for outcomes with positive (LRR increasing) or negative (LRR decreasing) valence (i.e., whether predicted behavior change is an increase or decrease in level; see Pustejovsky, 2017 for detailed computations).

When synthesizing outcomes across SCDs, LRR has many features that make the index desirable for use as an effect size. Most notably, effect size outcomes using LRR may be more readily interpreted by researchers and consumers since the metric closely aligns with the percentage of change between conditions, a commonly used construct (Pustejovsky, 2017). Once calculated, LRR can be transformed to describe a percentage of change between conditions. Unlike overlap-based metrics, LRR is not sensitive to procedural variations in data (e.g., length of session, measurement system; Pustejovsky, 2015), nor is the magnitude of LRR values impacted by changes in measurement units. It is also less sensitive to procedural variations than other common metrics. Additionally, LRR may be used to synthesize outcomes across studies that use different measurement systems (e.g., timed event recording and partial interval recording to estimate count) since the index compares mean-levels of outcomes in each condition rather than variability between conditions (Pustejovsky, 2017).

Like other mean-based and regression metrics, LRR assumes independence of data points, an assumption that may be unlikely to be met for typical variables measured in the context of SCDs. Moreover, models used to estimate the index assume data stability in each condition; thus LRR is not appropriate for data when trends are likely in either condition (e.g., when behavior change is expected to occur gradually rather than abruptly; Pustejovsky, 2017). Since the index is calculated as a proportion, it cannot be calculated for data with condition mean values of zero (e.g., acquisition behaviors not expected to be present in baseline). Auto-correlation can be addressed when conducting meta-analyses using LRR via the use of robust variance estimation, but no corrections to date have been modeled to address the presence of trend (see Pustejovsky, 2017 for a detailed explanation). Additionally, given the novelty of the index, benchmark values categorizing the size of LRR outcomes have not been established; LRR values are likely to be large when the base rates are low, even when mean differences are relatively small (e.g., a change of 1 to 1.1 receives the same effect size as a mean change of 100 to 110); thus, when LRR is used, analysts should attend to confidence intervals in addition to effect sizes. Current evaluations comparing outcomes obtained via LRR and other established within-case effect sizes are ongoing (Zimmerman et al., 2017), thus applications of LRR will likely evolve rapidly as the field of SCD meta-analysis continues to grow. Unlike between-case standardized mean difference (SMD), within-case LRR effect size estimates are not comparable to effect sizes calculated for group designs; estimates can only be compared to effect sizes obtained via other SCD syntheses (cf. Common, Lane, Pustejovsky, Johnson, & Johl, 2017 for a sample SCD synthesis using within-case LRR and between-case estimates). LRR estimates, standard errors, and confidence intervals can be calculated using open-source software and packages obtained via the following link: https://github.com/jepusto/SingleCaseES.

Standardized Mean Differences

Calculating SMD indices is another option for characterizing behavior change in SCD. *Glass's Δ* (*delta*), *Cohen's d*, and averaging *HPS d* (initials of the surnames of authors of this version of d statistic: Hedges, Pustejovsky, & Shadish, 2012; 2013) statistic fall within this category. An initial application of SMD to SCD data focused on the between-group designs *d* statistic using pooled standard deviation in the denominator (Cohen, 1992):

$$d = \bar{X}_A - \bar{X}_B \, s_p$$

$\bar{X}_A$ and $\bar{X}_B$ refer to the mean of the baseline and treatment observations respectively and s_p is the within-case standard deviation (i.e., a measure of variability of data in baseline and intervention conditions). Instead of the pooled within-case standard deviation, only the baseline standard deviation can be used to obtain the standardized mean difference. This is referred to as *Glass's Δ* (Busk & Serlin, 1992; Glass, McGaw, & Smith, 1981):

Δ = X ¯ A – X ¯ B s b

with s_b referring to the baseline standard deviation. We recommend using the standard deviation of outcome scores in the baseline condition. If the pooled standard deviation is used, the standard deviation is overestimated because the change in outcome scores due to the intervention is reflected. This means that the standard deviation is not only reflecting the outcome scale, but also the value of the treatment effects (and this varies across studies). On the other hand, when only baseline data are used to estimate the standard deviation, there is a smaller n and thus it can be less accurate. If the data in baseline are not variable, then the standard deviation will be 0, rendering an incalculable equation.

Hedges g (1981) is an extension of *Cohen's d* and recommended in context of SCDs as this involves a bias correction for small sample sizes:

g = d (1 – (3 4 m – 1))

where n is the number of data points in the series and m = degrees of freedom = $(n - 2)$.

Once the SMD (*Glass's Δ, Cohen's d or Hedges g*) is calculated, a weight can be assigned to each study effect size by a function of the sample size. For the simplest weight calculate $1/v$ (where v is the known variance of the sampling distribution). The following link can be used to have access to open-source software to calculate these classical mean differences: http://cran.r-project.org/web/packages/RcmdrPlugin.SCDA/index.html.

Given that the inferential use of these *SMD* indices in SCD is problematic (Beretvas & Chung, 2008), an alternative was proposed by Hedges et al. (2012, 2013, *HPS d* statistics) specifically for SCD (also called between-case standardized mean difference). These indices were developed to take into account autocorrelation and between-participants variability, apart from within-participant variability. The *HPS d* statistics have the advantage that they are comparable to *Cohen's d* as obtained from between-group design studies. Note that *HPS d* is only applicable for combining multiple-baseline design studies with at least three cases, which comprises a relatively small proportion of available SCD research. Moreover, this metric, like others, is subject to variability based on procedural variations (Pustejovsky, 2016b). An open-source user-friendly tool can be used to calculate *HPS d* and it can also be calculated using an R program (directions available here: http://jepusto.github.io/getting-started-with-scdhlm).

Regression Based Approaches

Meta-analytic techniques based on regression analysis, namely the *PHS d* statistic (Pustejovsky, Hedges, & Shadish, 2014) and hierarchical linear modeling (Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2014; Shadish, Kyse, & Rindskopf, 2013; Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003a, 2003b, 2007, 2008) are promising methods for combining SCD research (Kratochwill et al., 2010). The *PHS d* statistic (Pustejovsky et al., 2014) has the same underlying idea as the *HPS d* statistic, but the former offers the possibility of obtaining a standardized mean difference from a variety of fitted multilevel models (e.g., controlling for baseline trend and taking into account change in slope). This index is applicable for multiple baseline designs, A-B-A-B withdrawal designs, and alternating treatments designs.

The three-level hierarchical linear model takes the natural multilayered SCD data structure into account—measurement occasions are nested within subjects and subjects in turn are nested within studies. That is, we know that data points from the same participants are related (more likely to be similar than data points from different participants) and that data from a particular study are related (more likely to be similar than data points from different studies). Ignoring the multilayered nature can have a substantial impact on the conclusions of a multilevel analysis (Hox, 2002; Van den Noortgate, Opdenakker, & Onghena, 2005) as standard error estimates will be too small resulting in an inflated number of Type I errors when used in statistical tests (i.e., the statistical test indicates a treatment effect, whereas in reality there is none). By using the piecewise regression model introduced by Center, Skiba and Casey (1985–1986), the change in level (β_{2jk}) and change in slope (β_{3jk}) can be calculated for each case j within study k:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} T_{ijk} + \beta_{2jk} D_{ijk} + \beta_{3jk} T_{ijk} D_{ijk} + e_{ijk} \text{ with } e_{ijk} \sim N(0, \sigma_{ejk}^2) \quad (1)$$

Y_{ijk} is the outcome score on measurement occasion i for case j from study k (can be a baseline or a treatment observation) and is regressed on a time variable (T_{ijk}), a dummy coded variable (D_{ijk} = 0 if i belongs to the baseline, 1 otherwise), and an interaction between T_{ijk} and D_{ijk}. The ordinary least square estimate of β_{2jk} and β_{3jk}(i.e., b_{2jk} and b_{3jk} respectively) are the effect sizes of interest (change in level and change in slope respectively) and equal an unknown population effect size, indicated by the β coefficients, plus a random deviation from this population parameter, indicated by the error terms:

$$b_{2jk} = \beta_{2jk} + e_{2jk} \text{ with } e_{2jk} \sim N(0, \sigma_{e2jk}^2) \; b_{3jk} = \beta_{3jk} + e_{3jk} \text{ with } e_{3jk} \sim N(0, \sigma_{e3jk}^2) \quad (2)$$

The effect sizes can be standardized and combined across cases and across studies to obtain overall treatment effect estimates. The standardized effect sizes,b'_{1jk} and b'_{2jk} are obtained by dividing the estimated effect sizes, b_{1jk} and b_{2jk}, by the residuals' standard deviation, σ ^ e 2 (obtained by equation):

b ' 1 j k = b 1 j k σ ^ e j k 2 and b ' 2 j k = b j k 2 σ ^ e j k 2

For more details about standardization, we refer the reader to Van den Noortgate and Onghena (2008) and Ugille, Moeyaert, Beretvas, Ferron and Van den Noortgate (2012). To combine the estimated effect sizes across subjects and across studies, a multilevel meta-analysis can be performed, one for each kind of effect size (i.e., b'_{1jk} and b'_{2jk}, Ugille et al., 2012). This approach is the most flexible, given its ability to model complexities such as autocorrelation; predictors at the case (e.g., age, gender, SES, school type) and study level (e.g., age, gender, SES, school type, study quality); heterogeneous within-subject, between-subject, and between-study (co)variance; and allowance for combining different SCD types. It allows for estimation of average treatment effects across studies in addition to subject-specific and study-specific treatment effects (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013b, 2013c, 2015; Shadish & Rindskopf, 2007; Van den Noortgate & Onghena, 2003a, 2003b, 2008). A workshop including a step-by step tutorial to perform the multilevel meta-analysis (include software applications and illustrations) is available on the following website: www.single-case.com.

Readers should note that the basic regression-based approach assumes the errors are independent and normally, identically distributed. However, work is currently being done to extend the model by (1) including autocorrelation to deal with dependent errors and (2) modeling heterogeneous variance (Joo, Ferron, Moeyaert, Beretvas, & Van den Noortgate, 2017). In addition, continuous outcomes are assumed (when the data are counts, a generalized multilevel regression model would be more appropriate). Many SCDs do not include sufficient data for regression approaches, and the use of a combination of continuous and non-continuous outcomes is also problematic.

Summary

Given the emphasis on evidence-based practice and the ever-expanding amount of available research, scholars have been increasingly interested in synthesis of SCD outcomes. Over the past decade or so, many indices have been developed and validated for summarizing groups of SCD studies; we expect that this area of research will continue to grow. We suspect that no one metric will be appropriate for all analyses; thus, secondary statistical analyses should be chosen based on research questions and contemporary information about the strengths and weaknesses of each approach. We urge readers to use visual analysis to determine whether functional relations exist in each study; then, to use an appropriate metric to summarize the magnitude of behavior change, always explicitly reporting what characteristics of the data the metric relies on (e.g., Is it a measure of mean difference or overlap? Do the assumptions for the metric hold for the specific data included?).

Appendix 14.1

Visual Analysis Worksheet

Characteristic	Questions	+	—
Level	Is a consistent level established in each condition prior to condition change?	Yes	No
	Is there a consistent level change between conditions, in the expected direction?	Yes	No
Trend	Are unexpected trends present that make determination of behavior change difficult?	No	Yes
	Is there a consistent change in trend across conditions, in the expected direction?	Yes	No
Variability	Does unexpected variability exist in one or more conditions?	No	Yes
	Does within-condition variability impede determinations about level changes between conditions?	***No***	***Yes***
Consistency	Are data within conditions and changes between conditions consistent?	Yes	No
	If changes are inconsistent with regard to level, trend, or variability, was that expected?	Yes	No
	Does inconsistency impede confidence in a functional relation?	***No***	***Yes***
Overlap	Are data highly overlapping between conditions? (e.g., are there many points in the intervention condition that are not improved relative to baseline?)	No	Yes
	If overlapping, does the degree of overlap improve over time? (e.g., initial intervention data points are overlapping, but later ones are not)	Yes	No
	Is overlap consistent across comparisons? (e.g., Do approximately the same number or percent of data points overlap across A-B comparisons?)	Yes	No
	Was overlap expected a priori? (e.g., Was variability or a delay in treatment effect expected, given knowledge about participant behavior and past research?)	Yes	No
	Does presence of overlap impede confidence in a functional relation? *(Does the degree to which data are similar between conditions result in lower confidence for ≥1 comparisons?)*	***No***	***Yes***
Immediacy	Are changes between tiers immediate, in the intended direction?	Yes	No
	If no, are delays in change consistent across tiers (e.g., if there is a 3 session delay in Tier 1, is there a 2–4 session delay in Tier 2?)	Yes	No
	Does lack of immediacy impede confidence in a functional relation?	***No***	***Yes***

What is your determination regarding the presence of a functional relation?			Present	Not Present
How confident are you in your determination?	Not at all confident	Not very confident	Quite confident	Extremely confident

Appendix 14.2

Data Extraction Decision Worksheet

Authors:
Synthesis:

Data Extraction Program:

Image Naming Rule:

Image Capture Rules

Design	Zoom
A-B-A-B withdrawal design	
Multitreatment design	
Changing criterion design	
Multiple baseline design	
Multiple probe design	
Alternating treatments design	
Adapted alternating treatments design	
Multielement design	
Combination design (e.g., ATD in ABAB)	
Other (e.g., Author name, multiple baseline with 4 panels and 3 data paths on each panel)	

Rounding Rules for Studies								
Study ID	Author	Design	Interval length	Session length	Total number intervals	Only whole numbers	Decimal possible	Rounding Rule
1	*Adams*	*ATD*	*10 s*	*10*	*60*	–	–	*Round to closest possible value*
2	*Adams*	*ABAB*	–	*10*	–	*No*	*Yes*	*Round to nearest hundredth*

References

Barlow, D. H., Nock, M. K., & Hersen, M. (2009). *Single case experimental designs: Strategies for studying behavior change* (3rd ed.). Boston, MA: Pearson.

Barton, E. E., Meadan, H., Fettig, A., & Pokorski, B. (2016, February). *Evaluating and comparing visual analysis procedures to non-overlap indices using the parent implemented functional assessment based intervention research*. Poster presented at the Conference on Research Innovations in Early Intervention, San Diego, CA.

Barton, E. E., & Wolery, M. (2008). Teaching pretend play to children with disabilities: A review of the literature. *Topics in Early Childhood Special Education*, *28*, 109–125.

Beretvas, S. N., & Chung, H. (2008). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention*, *2*, 129–141.

Burns, M. K., Zaslofsky, A. F., Kanive, R., & Parker, D. C. (2012). Meta-analysis of incremental rehearsal using phi coefficients to compare single-case and group designs. *Journal of Behavioral Education*, *21*, 185–202.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research designs and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Erlbaum.

Campbell, J. M. (2003). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: A quantitative synthesis of single-subject research. *Research in Developmental Disabilities*, *24*, 120–138.

Campbell, J. M. (2012). Commentary on PND at 25. *Remedial and Special Education*, *34*, 20–25.

Center, B. A., Skiba, R. J., & Casey, A. (1985–1986). A methodology for the quantitative synthesis of intra-subject design research. *Journal of Special Education*, *19*, 387–400.

Chen, M., Hyppa-Martin, J. K., Reichle, J. E., & Symons, F. J. (2016). Comparing single case design overlap-based effect metrics from studies examining speech generating device interventions. *American Journal on Intellectual and Developmental Disabilities*, *121*, 169–193.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*, 155–159.

Cole, C. L., & Levinson, T. R. (2002). Effects of within-activity choices on the challenging behavior of children with severe developmental disabilities. *Journal of Positive Behavior Interventions*, *4*, 29–37.

Common, E. A., Lane, K. L., Pustejovsky, J. E., Johnson, A. H., & Johl, L. E. (2017). Functional assessment-based interventions for students with or at-risk for high-incidence disabilities: Field testing single-case synthesis methods. *Remedial and Special Education*. doi:10.1177/07419325176933

Cooper, H. M. (2010). *Research synthesis and meta-analysis: A step-by-step approach*. London: Sage.

Dibley, S., & Lim, L. (1999). Providing choice making opportunities within and between daily school routines. *Journal of Behavioral Education*, *9*, 117–132.

Doyle, P. M., Wolery, M., Ault, M. J., & Gast, D. L. (1988). System of least prompts: A literature review of procedural parameters. *Journal of the Association for Persons With Severe Handicaps*, *13*, 28–40.

Ferron, J. M., Farmer, J. L., & Owens, C. M. (2010). Estimating individual treatment effects from multiple- baseline data: A Monte Carlo study for multilevel-modeling approaches. *Behavior Research Methods*, *42*, 930–943.

Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014). Estimating causal effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods*, *19*, 493–510.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, *5*, 3–8.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Newbury Park: Sage.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*, 107–128.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods*, *3*, 224–239.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, *4*, 324–341.

Hershberger, S. L., Wallace, D. D., Green, S. B., & Marquis, J. G. (1999). Meta-analysis of single-case designs. In R. H. Hoyle (Ed.), *Statistical strategies for small sample research* (pp. 109–132). Newbury Park, CA: Sage.

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.

Hox, J. (2002). *Multilevel analysis. Techniques and applications*. Mahwah, NJ: Erlbaum.

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools*, *44*, 483–493.

Joo, S., Ferron, J., Moeyaert, M., Beretvas, S., & Van den Noortgate, W. (2017). Model specification approaches for the level-1 error structure when synthesizing single-case data with multilevel models. *Journal of Experimental Education*.

Kahng, S. W., Chung, K. M., Gutshall, K., Pitts, S. C., Kao, J., & Girolami, K. (2010). Consistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis*, *43*, 35–45.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_scd

Lane, J. D., Lieberman-Betz, R., & Gast, D. L. (2016). An analysis of naturalistic

interventions for increasing spontaneous expressive language in children with autism spectrum disorder. *The Journal of Special Education*, *50*, 49–61.

Ledford, J. R., Ayres, K. A., Lane, J. D., & Lam, M. F. (2015). Accuracy of interval-based measurement systems in single case research. *Journal of Special Education*, *49*, 104–117.

Ledford, J. R., King, S., Harbin, E. R., & Zimmerman, K. N. (2017). Social skills interventions for individuals with ASD: What works, for whom, and under what conditions? *Focus on Autism and Other Developmental Disabilities*. doi: 10.1177/1088357616634024

Ledford, J. R., Lane, J. D., Zimmerman, K. N., & Shepley, C. (2016, February). *Bigger, better, & more complex: To what extent do newer overlap-based metrics adequately describe single case data?* Poster presented at the Conference on Research Innovations in Early Intervention. San Diego, CA.

Ledford, J. R., & Wolery, M. (2013). The effects of graphing a second observer's data on judgments of functional relations when observer bias may be present. *Journal of Behavioral Education*, *22*, 312–324.

Lemons, C. J., King, S. A., Davidson, K. A., Berryessa, T. L., Gajjar, S. A., & Sacks, L. H. (2016). An inadvertent concurrent replication: Same roadmap, different journey. *Remedial and Special Education*, *37*, 213–222.

Lipsey, M. W., & Wilson, D. B. (2001). P*ractical meta-analysis. Applied social research methods series* (Vol. 49). Thousand Oaks, CA: Sage.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. *Behavior Modification*, *30*, 598–617.

Maggin, D. M., Johnson, A. H., Chafouleas, S. M., Ruberto, L. M., & Berggren, M. (2012). A systematic evidence review of school-based group contingency interventions for students with challenging behavior. *Journal of School Psychology*, *50*, 625–654.

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta- analysis of single-subject research for students with disabilities: 1985–2009. *Exceptionality*, *19*, 109–135.

Manolov, R., & Moeyaert, M. (2017). Recommendations for choosing single-case data analytical techniques. *Behavior Therapy*, *48*, 97–114.

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods*, *41*, 1262–1271.

Manolov, R., & Solanas, A. (2013). A comparison of mean phase difference and generalized least squares for analyzing single-case data. *Journal of School Psychology*, *51*, 201–215.

Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis*, *23*, 341–351.

Moeyaert, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-subject experimental data. *Journal of*

School Psychology, *52*, 191–211.

Moeyaert, M., Maggin, D., & Verkuilen, J. (2016). Reliability, validity, and usability of data extraction programs for single-case research designs. *Behavior Modification*, *40*, 874–900.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van Den Noortgate, W. (2013a). Modeling external events in the three-level analysis of multiple-baseline across-participants designs: A simulation study. *Behavior Research Methods*, *45*, 547–559.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2013b). The three-level synthesis of standardized single-subject experimental data: A Monte Carlo simulation study. *Multivariate Behavioral Research*, *48*, 719–748.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van Den Noortgate, W. (2013c). Modeling external events in the three-level analysis of multiple-baseline across-participants designs: A simulation study. *Behavior Research Methods*, *45*, 547–559.

Moeyaert, M., Ugille, M., Ferron, J., Beretvas, S., & Van den Noortgate, W. (2014a). The influence of the design matrix on treatment effect estimates in the quantitative analyses of single-case experimental design research. *Behavior Modification*, *38*, 665–704.

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2015). Estimating intervention effects across different types of single-subject experimental designs: Empirical illustration. *School Psychology Quarterly*, *30*, 50–63.

Moeyaert, M., Ugille, M., Ferron, J., Onghena, P., Heyvaert, M., & Van den Noortgate, W. (2014b). Estimating intervention effects across different types of single-subject experimental designs: Empirical illustration. *School Psychology Quarterly*, *25*, 191–211.

O'Connor, D., Green, S., & Higgins, J. P. T. (2008). Defining the review question and developing criteria for including studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions version 5.0.0*. London: The Cochrane Collaboration.

Parker, R. I., & Brossart, D. F. (2003). Evaluating single-case research data: A comparison of seven statistical methods. *Behavior Therapy*, *34*, 189–211.

Parker, R. I., Hagan-Burke, S., & Vannest, S. (2007). Percentage of all nonoverlapping data (PAND): An alternative to PND. *Journal of Special Education*, *40*, 194–204.

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single case research: Nonoverlap of All Pairs (NAP). *Behavior Therapy*, *40*, 357–367.

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate differences for single-case research. *Exceptional Children*, *75*, 133–150.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification*, *35*, 303–322.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, *42*, 284–299.

Parker, R. I., & Vannest, S. (2007). *Pairwise data overlap for single case research*. Unpublished manuscript. Pokorski, E. A., Barton, E. E., & Ledford, J. R. (2017). A

review of the use of group contingencies in preschool settings. *Topics in Early Childhood Special Education*, *36*, 230–241.

Pustejovsky, J. E. (2015). Measurement-comparable effect sizes for single-case studies of free-operant behavior. *Psychological Methods*, *20*, 342–359.

Pustejovsky, J. E. (2016a). *Standard errors and confidence intervals for NAP*. Retrieved from http://jepusto.github.io/NAP-SEs-and-CIs

Pustejovsky, J. E. (2016b). *Procedural sensitivities of effect sizes for single-case designs with behavioral outcome measures*. Retrieved from http://jepusto.github.io/working_papers/

Pustejovsky, J. E. (2016c). *Tau-U*. Retrieved from http://jepusto.github.io/Tau-U

Pustejovsky, J. E. (2016d). *SCD-effect-sizes: A web application for calculating effect size indices for single-case designs* (Version 0.1) [Web application]. Retrieved from https://github.com/jepusto/SingleCaseES

Pustejovsky, J. E. (2017). *Using response ratios for meta-analyzing single-case designs with behavioral outcomes*. Retrieved from https://osf.io/4fe6u/

Pustejovsky, J. E., & Ferron, J. M. (2017). Research synthesis and meta-analysis of single-case designs. In J. M. Kaufmann, D. P. Hallahan, & P. C. Pullen (Eds.), *Handbook of special education* (2nd ed.). New York, NY: Routledge.

Pustejovsky, J. E., Hedges, L. V., & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics*, *39*, 368–393.

Rakap, S., Snyder, P., & Pasia, C. (2014). Comparison of nonoverlap methods for identifying treatment effect in single-subject experimental research. *Behavioral Disorders*, *39*, 128–145.

Reichow, B., & Volkmar, F. R. (2010). Social skills interventions for individuals with autism: Evaluation for evidence-based practices within a best evidence synthesis framework. *Journal of Autism and Developmental Disorders*, *40*, 149–166.

Rohatgi, A. (2014). *WebPlotDigitizer user manual version 3.4*. Retrieved from http://arohatgi.info/WebPlotDigitizer/userManual.pdf

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The handbook of research synthesis* (p. 239). New York, NY: Sage.

Schlosser, R. W., Lee, D. L., & Wendt, O. (2008). Application of the percentage of nonoverlapping data (PND) in systematic reviews and meta-analyses: A systematic review of reporting characteristics. *Evidence-Based Communication Assessment and Intervention*, *2*, 163–187.

Scotti, J. R., Evans, I. M., Meyer, L. H., & Walker, P. (1991). A meta-analysis of intervention research with problem behavior: Treatment validity and standards of practice. *American Journal on Mental Retardation*, *96*, 233–256.

Scruggs, T. E., Mastropieri, M. A., & Castro, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education*, *8*, 24–33.

Severini, K. E., Ledford, J. R., & Robertson, R. (2017). A synthesis of interventions

designed to decrease challenging behaviors of individuals with ASD in school settings. *Manuscript under review.*

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). *The role of between-case effect size in conducting, interpreting, and summarizing single-case research* (NCER 2015–2002) Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education. This report is available on the Institute. Retrieved from http://ies.ed.gov/

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: new applications and some agenda items for future research. *Psychological Methods*, *18*, 385–405.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation*, *113*, 95–109.

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*, 971–980.

Shadish, W. R., Zelinsky, N. A., Vevea, J. L., & Kratochwill, T. R. (2016). A survey of publication practices of single-case design researchers when treatments have small or large effects. *Journal of Applied Behavior Analysis*, *49*, 656–673.

Shogren, K. A., Faggella-Luby, M. N., Bae, S. J., & Wehmeyer, M. L. (2004). The effect of choice-making as an intervention for problem behavior: A meta-analysis. *Journal of Positive Behavior Interventions*, *4*, 228–237.

Solanas, A., Manolov, R., & Onghena, P. (2010). Estimating slope and level change in N = 1 designs. *Behavior Modification*, *34*, 195–218.

Tarlow, K. R. (2016). An improved rank correlation effect size statistic for single-case designs: Baseline corrected tau. *Behavior Modification, 41,* 427–467.

Tincani, M., & Travers, J. (2017). Publishing single-case research design studies that do not demonstrate experimental control. *Remedial and Special Education.* doi: 10.1177/0741932517697447

Ugille, M., Moeyaert, M., Beretvas, S., Ferron, J., & Van Den Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Research Methods*, *44*, 1244–1254.

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly*, *18*, 325–346.

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers*, *35*, 1–10.

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *The Behavior Analyst Today*, *8*, 196–209.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence Based Communication Assessment and Intervention*, *2*, 142–151.

Van den Noortgate, W., Opdenakker, M., & Onghena, P. (2005). The effects of ignoring a

level in multilevel analysis. *School Effectiveness and School Improvement*, *16*, 281–303.

Vannest, K. J., & Ninci, J. (2015). Evaluating intervention effects in single-case designs. *Journal of Counseling & Development*, *93*, 403–411.

Whalon, K. J., Conroy, M. A., Martinez, J. R., & Werch, B. L. (2015). School-based peer-related social competence interventions for children with autism spectrum disorder: A meta-analysis and descriptive review of single case research design studies. *Journal of Autism and Developmental Disorders*, *45*, 1513–1531.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education*, *44*, 18–28.

Yoder, P. J., Bottema-Beutel, K., Woynaroski, T., Chandrasekhar, R., & Sandbank, M. (2013). Social communication intervention effects vary by dependent variable type in preschoolers with autism spectrum disorders. *Evidence-Based Communication Assessment and Intervention*, *7*, 150–174.

Zimmerman, K. N., Pustejovsky, J. E., Ledford, J. R., Barton, E. E., Severini, K. E., & Lloyd, B. P. (2017). Synthesis tools part two: Comparing overlap metrics and within-case effect sizes to evaluate outcomes when synthesizing single case research designs. *Manuscript under review*.

Index

Page numbers in italic indicate figures. Page numbers in bold indicate tables on the corresponding pages.