Methods Guide for Effect Estimation and Synthesis of Single-Case Studies

January 15, 2024

Contents

Authors

John M. Ferron, Megan Kirby, and Lodi Lipien *University of South Florida*

James Pustejovsky, Man Chen, and Paulina Grekov *University of Wisconsin - Madison*

Wendy Machalicek *University of Oregon*

Disclaimer

This report was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R324U190002 to the University of Oregon. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Citation

Preface

We developed the *Methods Guide* to support educational researchers interested in estimating effect size measures and synthesizing findings from single-case design studies. To do so, we aim to do two things: (a) provide guidance and decision rules to simplify the process of selecting effect size estimation and synthesis methods and (b) illustrate the step-by-step application of appropriate methods using readily available tools. The guide is not meant to cover all effect

size methods that have been proposed for single-case design studies, nor is it meant to summarize all of the research on effect size estimation and synthesis of single-case research. Rather, it is meant to provide easy-to-follow decision rules, along with step-by-step instructions and illustrations of user-friendly tools, so that researchers aiming to conduct a synthesis involving single-case studies are more readily able to do so.

The guide is organized so that researchers can go immediately to the sections and chapters that are relevant to their immediate task, rather than having to read the guide sequentially from start to finish. The first chapter provides background and a flow chart (also shown below) to help researchers select among three broad approaches to synthesizing results from single-case research.

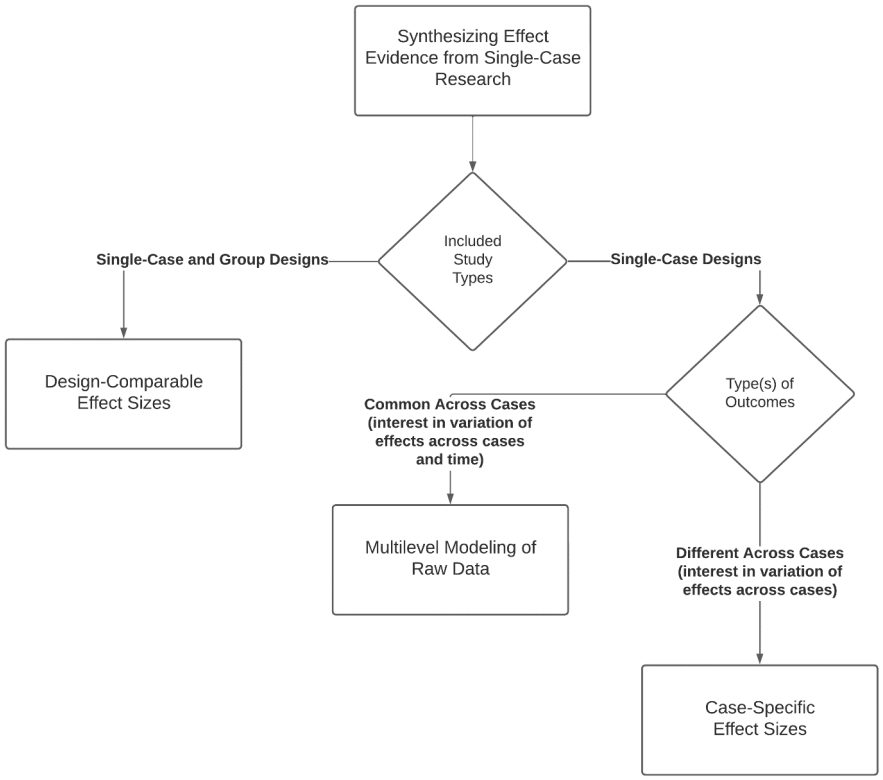


Figure 1: Approaches to synthesizing findings from single-case research

Those interested in synthesizing both single-case and group design studies should consider the design-comparable effect size approach and start by reading Chapter 2, which introduces this approach and its assumptions. They can then turn to one of the more detailed chapters providing step-by-step illustrations of how to estimate design comparable effect sizes with a user-friendly app

(Chapter 3 if there are no trends, Chapter 4 if there are trends only in the treatment phases, or Chapter 5 if there are trends in both the baseline and treatment phases).

Those interested in synthesizing single-case design studies that share a common outcome, and particularly if they are interested in examining changes in the effect over time, should consider the multilevel modeling approach and read Chapter 6, which introduces this approach and its assumptions. They can then turn to one of the more detailed chapters providing step-by-step instructions along with illustrations of how to estimate multilevel models to synthesize single-case data with a user-friendly app (Chapter 7 if there are no trends, or Chapter 8 if there are trends).

Finally, those interested in synthesizing single-case design studies where the outcome measures vary across cases, and especially if they are interested in examining how the size of the effect varies from case to case, should consider the case-specific effect size approach. They can turn to Chapter 9, which introduces this approach and its assumptions, and Chapter 10, which provides step-by-step instructions along with illustration of how to estimate case-specific effect sizes with a user-friendly app.

Datasets

Acknowledgements

We are grateful for the guidance from Katherine Taylor, who served as our project officer at the Institute of Education Sciences. We are also grateful for the suggestions and comments we received from the members of our external advisory panel: Jennifer Ledford, Joel Levin, Daniel Swan, Tai Collins, Katie Wolfe, and Kimberly Vannest.

Chapter 1

Approaches for Estimation and Synthesis of Single-Case Studies

This chapter provides the background and purpose for the Methods Guide for Effect Estimation and Synthesis of Single-Case Studies. We provide an overview of the three major approaches for effect size estimation and synthesis of single-case studies:

(a) design-comparable effect sizes,
(b) multilevel modeling of raw individual participant interrupted time-series data, and
(c) case-specific effect sizes. We describe the general motivation and rationale for each approach and provide a series of decision rules to guide researchers in their selection. Subsequent chapters in this guide provide more detailed illustrations of the three major approaches.

1.1 Background

Educational decisions made at the state, district, school, and student levels are expected to be informed by empirical evidence (**??**). These expectations create a compelling need for synthesis, or the integration of research findings from multiple, existing sources of evidence—including findings from single-case designs (SCDs). The need for methods to synthesis findings from SCDs is all the greater because some educational disciplines have historically relied upon this methodology to collect evidence about interventions, such that the bulk of available evidence about some interventions comes from SCDs.

This methods guide responds to three key developments in educational research

related to the documentation of evidence-based practices. The first is the expanded use of SCD research methods across disciplines over the last 50 years in both general and special education contexts (**?**). Although the history of SCD methodology is outside our present scope, SCD design innovations have allowed researchers to advance its use beyond quasi-experimental and behavior analysis origins (see **?**, for more detailed information on the history of SCD). Over the past 20 years, researchers' commitment to using rigorous procedures to identify evidence-based educational practices affirms not only the importance of randomized control trials (RCTs), but also the utility and efficiency of SCDs.

The second key development in education research is the emergence of effect size measures and synthesis methods for use with SCD and other interrupted time-series data (**??**). Interest in the synthesis of SCDs is long-standing (e.g., **???**), with methods arising concurrently with the meta-analyses of group design research becoming more statistically rigorous and sophisticated (**??**). The interest and intensity of methodological research focused on SCDs has increased over time, resulting in a substantial expansion in the diversity, flexibility, and accessibility of available analytic methods. The current state of SCD effect size estimation methods can be categorized into three strands: (a) approaches that summarize the effect for each case and then synthesize these case-specific effect sizes (e.g., **?**), (b) methods that use multi-level models to analyze the raw or standardized data from one or multiple SCD studies (**?**), and (c) techniques that use design-comparable effect size metrics (**?????**). We organize the present methods guide around these three broad methodological strands.

The third development in education research is the increased use of systematic review and meta-analysis procedures to identify and affirm evidence-based practices in education (**??**). In examining the production of systematic reviews and meta-analyses of SCDs from 1985 to 2009, **?** found a marked increase in such products between 2000 and 2009. Similarly, **?** found increasing production through 2015, as well as improvements in the quality of published reviews. However, even contemporary reviews frequently fail to use appropriate methods for combining findings across studies (**?**). Thus, there remains a need for guidance about how to select and apply methods for synthesizing SCD research results.

Effect size measures are a valuable complement to visual analysis for the interpretation of single-case research results and are a key input in two of the available approaches for meta-analytic synthesis of SCDs: case-specific and design-comparable effect size estimation methods. Researchers have a variety of technically sound effect size metrics to select from when interpreting SCD findings, but relatively few published meta-analyses use design-comparable effect sizes, multilevel modeling, or more advanced case-specific effect sizes (cf. **???**). One reason for their lack of widespread use by researchers may be the conceptual and procedural complexity associated with advances in effect size measures and meta-analysis techniques. The complexity of data extraction and calculation of effect sizes for SCDs may also hinder a more rapid uptake of effect size estimation methods. In addition, researchers may lack software tools for

effect size estimation for single-case studies when using some common statistical packages such as SPSS and SAS (**???**).

1.2 Purpose of the Methods Guide

The purpose of this methods guide is to improve educational researchers' practice of estimating effect size measures and synthesizing findings from SCDs. We recognize that no single method is ideal for all research goals. Furthermore, methods that have the most to offer can be complex and may appear difficult to carry out. Through the use of this methods guide, we aim to (a) provide guidance and decision rules to simplify the process of selecting effect size estimation and synthesis methods, and (b) illustrate the step-by-step application of appropriate methods using readily available tools, such as web-based software to calculate case-specific effect sizes or design-comparable effect sizes for SCD studies.

1.3 Study Quality

Conducting a research synthesis—composed of group design studies, SCDs, or both—involves several stages: formulating the research aims, specifying inclusion criteria, conducting a systematic search, and screening for eligible studies (**??**). Additionally, before carrying out effect size calculations or meta-analysis, it is critical to consider the methodological quality and potential biases of studies one includes in the synthesis. Several distinct sets of standards are available for SCD studies to assist researchers with assessing methodological quality (e.g., **?????**). After examining the methodological quality of studies eligible for inclusion in a review, researchers can use one of two strategies to guide their estimation and synthesis of effect sizes. One strategy incorporates consideration of study quality as an inclusion criterion. Researchers can screen studies and exclude low-quality studies so that the synthesis is based on a subset of studies with quality deemed adequate or high enough that outcome(s) can be attributed to the intervention. For example, the **?** indicates that for researchers to include an SCD study in a meta-analysis, the study must meet minimum design criteria (e.g., a multiple baseline study has at least three temporally spaced opportunities for the effect to be demonstrated, along with phases of at least three observations) and must provide minimum evidence of experimental control over extraneous factors (e.g., baselines do not document a therapeutic trend). Other screening approaches that rely on visual analysis have also been suggested (**??**). Alternatively, researchers can use broader inclusion criteria, but then carefully code for aspects of study quality, so that they can examine study quality codes as potential moderators in a meta-analysis. With this approach, researchers can estimate effect sizes for studies of varying degrees of quality and empirically explore whether the variation in effect size across studies is dependent on aspects of study quality.

We assume that researchers who use this methods guide will have already selected a method to assess study quality and an approach for incorporating those assessments into their synthesis. Thus, we do not focus on SCD study quality assessment methods, but rather provide guidelines that researchers can apply to a collection of studies meeting their inclusion criteria, which potentially include study quality. In this guide we focus on the final stages of a research synthesis—those involving questions of how to select a method for estimating effects, how to compute effect size estimates (or otherwise prepare data for synthesis), and how to synthesize findings in the form of effect size estimates or individual-level data.

1.4 Selecting an Approach for Effect Estimation and Synthesis

To select an approach for estimating and synthesizing effects from SCDs, researchers should first reflect on the research aims that motivate their synthesis. In some contexts, researchers' primary aims may be summarizing evidence to arrive at statements about the average efficacy of a class of interventions. In other circumstances, researchers might be interested in understanding variation in effects and the extent to which such variation is associated with participants' characteristics, specific intervention features, or other contextual factors. When summarization is a priority, researchers may find it more useful to use design-comparable effect sizes that describe average effects. If individual-level variation is the focus, then approaches using case-specific effect sizes or multilevel modeling may be more advantageous.

Another overarching consideration for selecting a synthesis approach pertains to the features of the included studies. Quantitative synthesis requires choosing an effect size metric that permits comparisons of the magnitude of effects across individual participants and studies. Consequently, the extent to which eligible studies use different types of designs or different outcome measures constrains how effects from those studies can be described or compared. For instance, if all eligible studies in the review used multiple baseline designs across participants (or another common type of SCD), then several different synthesis approaches are feasible. In contrast, if eligible studies include both single-case and group design studies (e.g., small randomized experiments, each with a single pre-test and a single post-test), researchers may seek a synthesis approach that permits comparisons across both design types. If all eligible studies used SCDs with very similar methods for assessing the dependent variable, then synthesis based on multilevel modeling of raw data is possible. In contrast, if eligible studies used non-equivalent assessments, then researchers may need to use a synthesis approach based on case-specific effect sizes that are suitable for comparison across studies involving different assessments. These two broad considerations—the aims of the synthesis and the features of eligible studies—should guide the selection of an approach for synthesis of SCDs. We now detail how the three

broad synthesis approaches fit within these considerations.

1.5 Design-Comparable Effect Sizes

In some situations, researchers aim to synthesize the evidence for intervention effectiveness using both single-case and group design studies. For example, a meta-analysis by **?** analyzed 22 single-case and between-group studies to examine the effects of text-to-speech and other read-aloud tools on reading comprehension outcomes for students with reading disabilities. The authors used the standardized mean difference to estimate read-aloud intervention effects in the group design studies and a comparable standardized mean difference to estimate effects from the included SCD research, resulting in an overall average weighted effect size of d= 0.35, 95% confidence interval (CI) [0.14, 0.56]. Because the purpose of the study involved the comparison and averaging of effects across single-case and group designs, it was critical that **?** used an effect size metric that is theoretically comparable across the designs. Researchers should select from the design-comparable effect size options (e.g., accounting for the absence or presence of baseline trends) when the aim is to compare and synthesize effects across eligible SCD and group design studies (**??????**). However, if researchers aim to synthesize findings from only SCD studies (i.e., not to integrate findings across single-case and group design studies), it may be feasible and preferable to use other options, such as synthesizing effects using case-specific effect sizes or multilevel modeling.

1.6 Case-Specific Effect Sizes

In addition to averaging effects across studies, researchers may also be interested in exploring variation in treatment effects by categorical differences or individual participant characteristics (e.g., Do effects vary across settings? Are effects consistent across ethnic and racial groups?). When included studies use different outcome measures (e.g., included studies report outcomes measured on different scales such as a rate per session, occurrence/count, or ratio), it is important for researchers to use an effect size metric and synthesis approach that accounts for such. For example, **?** examined five potential moderators (emotional and behavioral disorder risk status, reinforcement frequency, target behaviors, intervention format, and grade level) in their synthesis of 21 SCDs to obtain an overall effect of the Good Behavior Game in promoting positive behavior in the classroom. Results of their meta-analysis suggested that the intervention was more effective in reducing problem behaviors among students with or at risk for emotional and behavioral disorders. Another meta-analysis by **?** first calculated and aggregated the effect sizes across all included studies, and then investigated the moderating effects of participant characteristics, targeted outcomes, and implementation components on the efficacy of video self-modeling, in which a learner with disabilities watches a video of a model

engaged in targeted skills or tasks. They found that intervention effects were stronger for younger participants with autism spectrum disorders compared to those not identified as autistic or having an autism spectrum disorder.

Because these syntheses focused on investigating variation across individuals in the effect of treatment, it was important that the effect size estimation and synthesis approach produced effect estimates for each individual participant (rather than a study-level summary effect estimate). Design-comparable effect size options are not viable for studying within-participant effects or variation in effects between individuals within the study because these effect sizes produce estimates at the study level (i.e., the effect averaged across the cases), not the individual level. Furthermore, outcome measures differed widely among studies included in the aforementioned reviews, so the researchers needed an individual-level effect size metric that was not scale-dependent (e.g., not based on simple raw score mean differences). When included studies use outcome measurements that cannot simply be re-scaled to be equivalent, case-specific effect size estimation and synthesis options may be best suited. However, the case-specific effect size options are not viable for meta-analysts wanting a single overall effect size after synthesizing both single-case design and group design studies because case-specific effects are not comparable to group design effects.

1.7 Multilevel Modeling of Individual Participant Interrupted Time-Series Data

When a set of SCDs use the same or very similar outcome measures with the aim of studying the variation in effects over time within and between individuals, the multilevel modeling approach should be considered. For example, **?** meta-analyzed 15 single-case studies totaling 79 students to examine the effects of an intervention on the level and trend in correct writing sequences per minute for students with disabilities. The researchers found that effect sizes were greater when students received intervention for longer durations (i.e., there was a positive effect on the trend) and that this temporal change in effect was more pronounced with younger students. In contexts like this, we suggest researchers avoid a meta-analytic approach that relies on a single effect size estimate for a study (e.g., design-comparable effect sizes) or even a single effect estimate for a case (e.g., case-specific effect sizes) and consider options for multilevel modeling of individual participant data series instead.

1.8 Summary of Options for Effect Estimation and Synthesis

The flow chart in Figure **??** illustrates a set of heuristic decision rules for selecting among the three general approaches to synthesizing results from single-case research. If the primary purpose of one's research is to integrate findings from

both single-case and group-design studies, researchers should consider design-comparable effect sizes, contextually noting that effects are likely to be different for group design studies than from SCD studies **?**. Alternately, if the researchers plan to only include SCD studies, then they can use two other approaches (i.e., multilevel modeling and case-specific effect sizes). The choice between the latter two is related to the measurement of the dependent variable. Where there is interest in variability in effects across cases and over time, researchers should consider multilevel modeling of the raw data series, so long as the outcome is measured consistently across cases. However, if the aim is to examine how effects vary across the cases but the outcome measurements are non-equivalent and cannot be easily equated, then researchers should consider the options for estimating and synthesizing case-specific effect sizes.

1.9 Limitations in Selecting an Approach for Effect Estimation and Synthesis

We emphasize that Figure **??** presents a heuristic, simplified procedure for selecting among the three general approaches to effect size estimation and synthesis, which cannot and does not cover every possible research context. We anticipate and acknowledge that there will be situations where researchers' aims and contexts differ from those we have described and thus do not align perfectly with one of our primary approaches to estimating and synthesizing single-case effect sizes. For example, researchers who are synthesizing findings from a set of SCDs may wish to compare their results to a previously published meta-analysis of group design studies, but not to investigate individual-level variation in treatment effects. They might therefore elect to use design-comparable effect sizes even though they are not formally integrating results from group design studies within their review.

A further possibility is that researchers might elect to use multiple approaches to synthesis to address different aims or research questions. For example, consider a project in which researchers have identified both single-case and group design studies. They might want to integrate findings across design types while also exploring the variation in effects among individuals. In this scenario, researchers could estimate design-comparable effect sizes for their first aim and case-specific effect sizes from the subset of SCDs for their second aim.

We also acknowledge that situations may arise that fall between those we described for case-specific effect sizes and those for multilevel modeling of the raw data series. For example, researchers may want to examine how effects vary over time and across cases, using studies with different outcomes. For this purpose, the researchers will need to identify and apply extensions of the primary approaches we present in this guide. For example, the researchers could either standardize the raw data before estimating a multilevel model, or they could synthesize case-specific effect sizes using multiple standardized effects for each

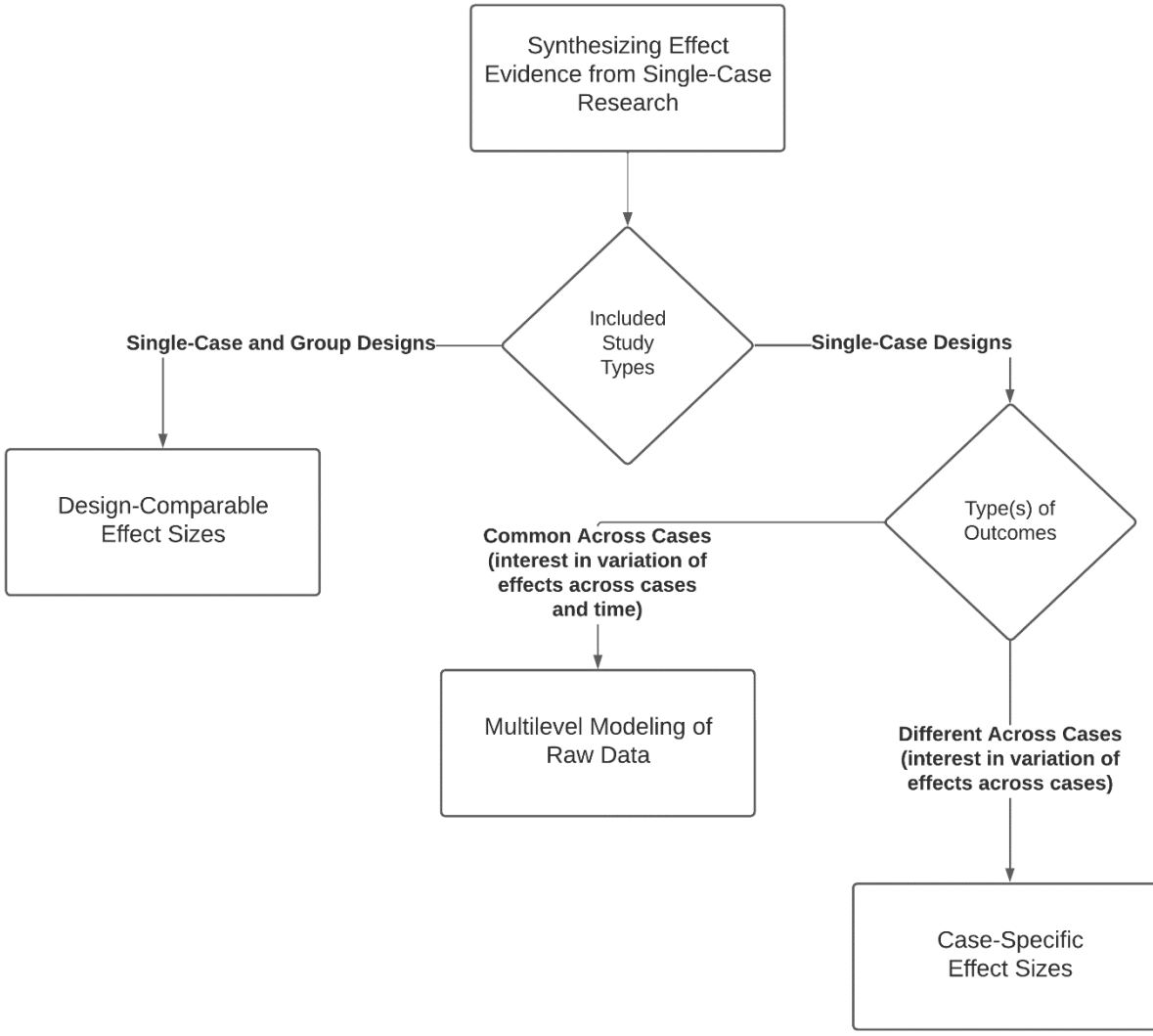


Figure 1.1: Approaches to Synthesizing Results from Single-Case Research

case (e.g., an effect that indexes the immediate shift in level, and another effect that indexes a change in slope). In this guide, we aim to address what we perceive to be the most common scenarios, rather than conduct an exhaustive review of all possibilities.

Finally, we anticipate that the heuristic guidance we provide here will need to be refined over time as further methodological innovations become available. We anticipate that research presently underway will provide even more meta-analytic options, with implications for how to select an approach for synthesis. At some point it may be possible to compute case-specific effect sizes that are also design-comparable, or it may be possible to standardize the data for multilevel models in a way that leads to parameter estimates from the model that correspond to design-comparable effect estimates. When such methods become available, some distinctions made here will become artificial. However, even as methodology continues to advance, researchers need guidance that acknowledges the complexity of research purposes and contexts and is dynamic in its accommodation of such variation, while also being concrete and straightforward enough for widespread implementation. To support the uptake of advanced methods for meta-analytic synthesis by educational researchers, the remainder of this guide follows the proposed heuristics for selecting among the three major approaches to effect size estimation and synthesis.

1.10 Structure of the Methods Guide

We divide this guide into three major sections: (a) design-comparable effect size estimation and synthesis, (b) multilevel modeling, and (c) case-specific effect size estimation and synthesis to estimate and synthesize effects. Each section is independent; they are not in a particular chronological order, nor do they build upon each other. Rather, we expect those using this guide to follow the decision rules in Figure **??** to determine which approach is most appropriate for them, and then reference the corresponding relevant section.

We divide each section into chapters. Each initial chapter introduces the specific approach to synthesizing results from SCD research, its assumptions, and determination for use. We then provide additional decision rules for selecting among the specific techniques and options available within a given broad approach. This is followed by an illustration where we:

1. Describe the purposes for estimating and synthesizing effects, and the available data.
2. Demonstrate how to use the decision rules in Figure **??**, along with the additional decision rules specific to the initial section chapter, to arrive at the option being illustrated.
3. Present the illustrated data to show how it needs to be structured for the analysis.

4. Provide a step-by-step illustration of how to estimate and synthesize effects using readily available analysis tools.

Chapter 2

Introduction to Design-Comparable Effect Sizes

This chapter provides background on design-comparable effect sizes, describes when to use them, and explains the key assumptions behind their use. We highlight both the assumptions for using design-comparable effect sizes in meta-analytic synthesis and the assumptions for estimation of these effect sizes from primary study data. We then describe available options for estimating design-comparable effect sizes. These options allow for different assumptions regarding trends in baseline or treatment phases and different assumptions about the variation in the treatment effect across cases. This chapter concludes with a set of decision rules for selecting among the options for design-comparable effect sizes.

2.1 Background

Design-comparable effect sizes are effect size indices for single case studies that are on the same metric as effect size indices used in group comparison studies (**??????**). These design-comparable effect sizes are valuable in a variety of synthesis contexts, particularly those that involve both single-case design (SCD) and group comparison studies. However, these methods also have several limitations.

First, design-comparable effect sizes rely on an estimate of the variance between participants, requiring at least three distinct participants for estimation. This constrains their use to SCDs involving multiple participants (e.g., multiple baseline design across participants) or close replications of SCDs, each of which has a single participant (e.g., ABAB design replicated across several students). Fur-

ther, existing research on design-comparable effect size methods is only available for multiple baseline, multiple probe, or replicated treatment reversal designs. Researchers have yet to develop extensions for alternating treatment designs and other types of SCDs. However, ongoing research is likely to increase the designs for which design-comparable effect sizes can be estimated.

A second conceptual limitation is that design-comparable effect sizes produce a single summary effect size per outcome—which represents an *average* effect across participants—just as between-group design study effect sizes are summaries of average effects across participants. As a result, the design-comparable effect size might conceal heterogeneity of effects across participants in an SCD. Meta-analytic researchers are then limited to moderator analyses focused on study-level characteristics; it is not feasible to examine potential moderators that vary across cases in a study or across time points within a case.

A third limitation is that design-comparable effect sizes are based on a hierarchical model that involves specific assumptions about the distribution of outcomes measured in the study. Developing a reasonable model requires care and attention to the plausibility of its assumptions. It is not a trivial or automatic process (as effect size calculations for between-group experimental designs are sometimes treated). Moreover, for some types of outcomes, the distributional assumptions of the model may be inappropriate, which further limits the applicability of the design-comparable effect size.

To address these methodological limitations, we use this chapter to provide researchers with guidance on the selection and use of design-comparable effect size estimates. We describe six of the most common modeling options and provide guidance on how to select among these options when calculating design-comparable effect size for use in a research synthesis.

2.2 When to Use Design-Comparable Effect Sizes

When choosing an effect size for single-case design data, researchers should begin by considering the broader purpose for computing effect sizes. In some cases, researchers may want to synthesize results across different types of studies, as in a comprehensive synthesis of all studies conducted on a topic. For example, researchers conducting a meta-analysis might include all studies examining the effects of social skills interventions on the social and academic outcomes of elementary-aged students with disabilities. In some areas of education research, it is likely that the literature identified for synthesis includes both group and single-case experimental studies. To average the effect across studies with different designs, researchers must pick an effect size index that has a comparable interpretation for each of the included designs (**????**). For exactly this purpose, methodologists developed the design-comparable effect size for SCDs, providing an effect size on a common metric by answering the question, "*What*

would the standardized mean difference effect size be if one could somehow perform a between-group randomized experiment based on the same population of participants, intervention protocol, and outcome measures?"

2.3 General Definition of Design-Comparable Effect Sizes

To understand the logic of the design-comparable effect size, it is helpful to consider how effect sizes are defined in group design studies. In a between-groups randomized experiment comparing a treatment condition (Tx) to a control condition (C) for a specified population, researchers commonly summarize results using the standardized mean difference effect size index. In Equation (**??**), we define this effect size parameter as

$$\delta = \frac{\mu_{Tx} - \mu_C}{\sigma_C}, \tag{2.1}$$

where μ_{Tx} is the average outcome if the entire population received the treatment condition, μ_C is the average outcome if the entire population received the control condition, and σ_C is the standard deviation of the outcome if the entire population received the control condition. The effect size may be estimated by substituting sample means and sample standard deviations in place of the corresponding population quantities (**?**), or by pooling sample standard deviations across the intervention and control conditions under the assumption that the population variance is equal. Alternately, the mean difference in the numerator of the effect size can be estimated based on a statistical model, such as an analysis of covariance that adjusts for between-group differences in baseline characteristics (**?**). Researchers often apply the Hedges g small-sample correction, which reduces the bias of the effect size estimator that arises from estimating σ_C based on a limited number of observations (**?**).

Using data from a multiple baseline, multiple probe, or replicated treatment reversal design, the design-comparable effect size for SCDs aims to estimate the same quantity as the standardized mean difference from a between-groups experiment. This task poses challenges because the data from such SCDs involve repeated measurements taken over time. To precisely define the design-comparable effect size, researchers must therefore be specific about the timing of both intervention implementation and outcome assessment. Hypothetically, if a between-groups experiment uses the same study procedures as the SCD, researchers would still need to determine and specify *when to begin intervention and when to collect outcome data*. Furthering this example, suppose that the SCD takes place over times $t = 1, ..., T$. In our hypothetical between-groups experiment, intervention starts at time A for $1 \leq A < T$ and collection of outcome data for all participants occurs at time B for $A < B$. The standardized mean difference from such an experiment would contrast the average outcome at time B if the entire population had started intervention at time A [i.e., $\mu_B(A)$] to the

average outcome at time B if the entire population had remained in baseline through time B and then started intervention later at time T [i.e., $\mu_B(T)$]. The standardized mean difference would then correspond to

$$\delta_{AB} = \frac{\mu_B(A) - \mu_B(T)}{\sigma_B(T)}, \tag{2.2}$$

where $\mu_B(A)$ is the average outcome at follow-up time B if the entire population were to receive the intervention at time A. Then, $\mu_B(T)$ is the average outcome at follow-up time B if the entire population were to receive the intervention at time T. Finally, $\sigma_B(T)$ is the standard deviation of the outcome at follow-up time B if the entire population were to receive the intervention at time T. Note that $\mu_B(T)$ corresponds to the average outcome under the control condition (μ_C, above), because participants would not yet have received intervention as of time B. Similarly $\sigma_B(T)$ is the analogue of σ_C, the standard deviation of the outcome under the control condition because participants would not yet have received the intervention as of time B.

? described a strategy for estimating δ_{AB} using data from an SCD study. Broadly, the strategy involves specifying a multilevel model for the data, estimating the component quantities $\mu_B(A)$, $\mu_B(T)$, and $\sigma_B(T)$ based on the specified model, and applying a small-sample correction analogous to Hedges g. However, because of the need to estimate the standard deviation of the outcome across the participant population [$\sigma_B(T)$], this strategy only works if the SCD study includes data from *multiple* participants. The approach involves a multilevel model for the data because SCDs involve repeated measurements collected for each of several participants. The first level of the model describes the pattern of repeated measurements over time nested within a given participant and the second level of the model describes how the first-level parameters vary across participants. As a result, the model involves deconstructing $\sigma_B(T)$ into two components: within-participant variation and between-participant variation. This process is not typically possible in a between-groups randomized experiment unless researchers collect repeated outcome measures for each participant.

2.4 What We Assume When We Synthesize Design-Comparable Effect Sizes

The motivation for using the design-comparable effect size from SCD studies is strongest when researchers intend to synthesize SCD and group design studies in the same meta-analysis. If assumptions needed for synthesis are not reasonably met, it may be more appropriate to analyze the SCD and group design studies separately. Then, researchers may want to consider alternative effect size metrics for the SCD studies. For this reason, this chapter presents the broader synthesis assumptions prior to the specific assumptions needed for estimating design-comparable effect sizes.

The random effects model is the predominant statistical model for synthesizing effect sizes across studies. With this statistical model, we do not assume that the population effect size estimated in one study is identical to the population effect size estimated from another study. Rather, we assume that the effect size estimated for Study j may differ from the effect size estimated for Study k (e.g., $\sigma_j \neq \sigma_k$). There are several general frameworks for explaining why the effect size may vary from one study to the next. One such framework posits that differences can arise from variation across studies in the units, treatments, observing operations, and settings (UTOS; **?**). The inclusion and exclusion criteria for the meta-analysis can be set up to constrain (but not completely eliminate) the variation from study to study on these dimensions. Use of a random effects model for the summary meta-analysis and the exploration of moderators of the treatment effects is warranted because some degree of variation in effects is anticipated.

There are several different ways to understand the assumptions underlying the random effects model. One way is to imagine that the included studies in a synthesis represent a random sample from a super-population of possible studies on the topic of interest. In a Bayesian framework, the model can also be motivated by the assumption of exchangeability, meaning that the effect size of studies included in a synthesis are on a common metric that permits judgements of similarity and that their relative magnitude cannot be systematically predicted a priori (**?**). For brevity, we refer to both suppositions (i.e., the super-population and Bayesian motivations) as the exchangeability assumption. Crucially, the exchangeability assumption depends on the effect size metric used for synthesis; for a given set of studies, the assumption may be reasonable for one effect size metric but unreasonable for another.

Thus far, we have defined the standardized mean difference metric (δ) for design-comparable effect sizes for SCDs. Therefore, we now consider the exchangeability of δs. When intending to synthesize a set of studies where there is considerable variation among the study outcomes, sampled units, or treatments (i.e., they differ greatly from one another on one or more of these UTOS characteristics), then the δs from these studies are likely not exchangeable. In contrast, when there is similarity in the UTOS, exchangeability is more reasonable. As an example, for the standardized mean difference metric, exchangeability is more plausible when one study's population of participants closely mirrors the participant population characteristics from another study (i.e., similar, if not same, inclusion criteria). Further, when the populations are similar and studies use the exact same operational measure of the dependent variable, we can assume that the distribution of outcomes in the control condition has similar variance. Alternatively, if the two studies drew from populations with very different characteristics so that the disbursement of the study results (distribution of the dependent variable) varied widely, an intervention that produces identical effects on the scale of the dependent variable would have quite different effect sizes on the scale of the standardized mean difference. When populations are distinctly different, the exchangeability assumption is less tenable. Thus,

we encourage researchers to examine the studies they plan to include in their synthesis for the potential lack of exchangeability.

Researchers can explore this aspect of the exchangeability assumption by examining the sampling methods and measurement procedures of the included studies. When subsets of studies use the same operational measure of the dependent variable, the between-participant (case) variance in those studies can be compared. To illustrate the exploration of between-case variability, consider the sampling procedures used in the following two studies extracted from **?** that examined the effect of interventions on writing performance as measured by correct word sequences (CWS). In one study, the sample consisted of three 7-year-old White males identified by their teachers as struggling with writing (**?**). In the other study (**?**), the sample comprised three 10-year-old students who exhibited poor writing skills. The first student was a Black male identified with an emotional disturbance, the second student was a White male identified with a specific learning disability, and the third student was a White female identified with a specific learning disability. Presented with this information, we then seek to answer the following questions: (a) Are these samples similar enough to satisfy the exchangeability assumption with the standardized mean difference metric? (b) Might the second sample with variation in differences in participant characteristics (age, race, gender, and educational disability category), be so much more variable in writing performance that it is not reasonable to use the standardized mean difference metric to judge similarity of effects across both studies?

In the first study (**?**), the mean number of CWS during baseline for the three participants were 8.29, 15.0, and 10.8. In the second study (**?**), participants' mean CWS baseline levels were 14.6, 29.1, and 22.1. In Study 1, the between-case variation, as indexed by the standard deviation (SD) of the three baseline means, is 3.4, whereas the between-case SD is 7.3 in Study 2. Now we must consider whether this difference is large enough to distort the design-comparable effect size.

To address our questions, we first consider the raw score effect size for each study by specifying a multilevel model that assumes no trends in baseline or treatment phases, and variation in the effect across cases. In Study 1, the shift in the expected number of CWS when moving from baseline to intervention, or raw score effect size, is 10.7. In Study 2, the CWS raw score effect size is 9.6, about 1.1 times smaller than the raw score effect size for Study 1. Next, we consider the design-comparable effect size computed using the same model used for the raw score effect size. The design-comparable effect size is 0.962 for Study 1 and 0.827 for Study 2; the effect size for Study 1 is approximately 1.2 times greater than Study 2, like the ratio of the observed raw score effect sizes.

We anticipate that future research will provide guidance regarding how much difference in sampling (and resulting between-case variability) researchers can accommodate without creating notable problems for the synthesis of design-comparable effect sizes. Similarly, researchers will continue to investigate, and

hopefully establish consensus about, the degree to which differences among outcomes and treatments between studies are tolerable to have confidence in their effect size results. Until then, we suggest that meta-analysts be aware of the underlying exchangeability assumption (that the effect sizes expressed on a given metric are exchangeable across studies) and be forthright and transparent about their findings when reporting results of their synthesis.

If the differences in the units, treatments, observing operations, and settings between the SCD studies and the group studies are much larger than the differences among either the SCD or group studies, it may be preferable to meta-analyze the SCD and group studies separately. Conversely, if the differences are negligible from one set of studies to the next, the exchangeability assumption is more tenable, and attention can be turned to the assumptions necessary to estimate the design-comparable effect size.

2.4.1 What We Assume When We Estimate Design-Comparable Effect Size

Design-comparable effect sizes for SCD studies rely on multilevel models. Estimation is based on several key assumptions, including distributional assumptions about the errors. The models assume that observations for each case are normally distributed around the case-specific trend line, and the variation of observations around the trend line is homogeneous from one phase to the next and from one case to the next. In addition, there are assumptions about the underlying structural model, including whether there are trends in phases. Below, we explain these assumptions in greater detail and provide exemplars of each.

2.4.2 Normality

Use of design-comparable effect sizes assumes that the experimental observations are normally distributed around the case-specific trend lines. The data may be consistent with this assumption of normality, but not always. For example, **?** utilized a multiple baseline design across three participants to examine the effects of a math software intervention with a game component on the off-task behaviors of students with attention-deficit hyperactivity disorder (ADHD). To examine the normality assumption, we extracted the study data using Webplot Digitizer (**?**) and present it in Figure **??**. When examining the baseline phases, observations appear to be distributed somewhat normally around the baseline means, with a pooled skewness value near zero ($sk = 0.15$) and a pooled kurtosis value near zero ($ku = -0.77$). However, we observe non-normality in the treatment phases because the outcome is a percentage that has remained near the floor of 0% for much of the treatment phase ($sk = 1.86$; $ku = 4.99$).

With count-based variables, we often anticipate some departures from the traditional normality assumption. These departures tend to be more pronounced

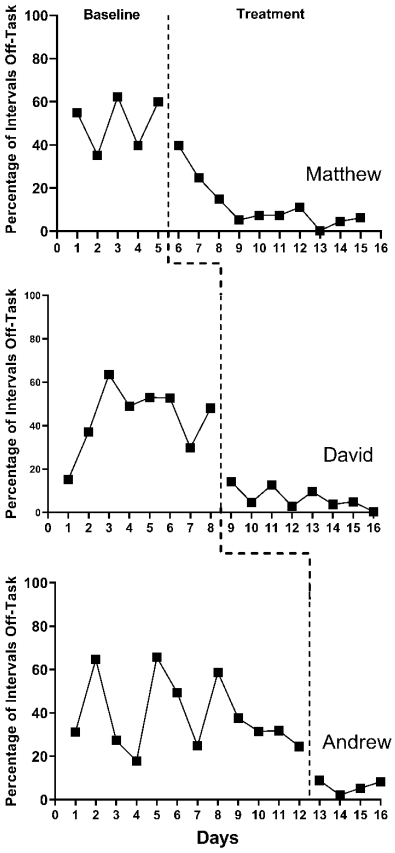


Figure 2.1: Multiple Baseline Design Across Three Participants (Ota & DuPaul, 2002)

when a count variable has a phase mean close to zero. Departures from the traditional normality assumption are more pronounced when a percentage variable has a phase mean close to either end of the $0\% - 100\%$ range. These situations are common in published SCDs because the quality of an SCD study is often judged on the stability and level of observations in baseline and treatment phases. For example, if researchers design an intervention to increase a non-reversible behavior (e.g., academic skill), it would be ideal to only recruit participants without the target skill in their repertoire so that observations in the baseline phase reflect such-they would have baseline observations at or near 0. Conversely, an SCD recommendation for behavior-reduction interventions is to seek low rates or non-occurrence of the target behavior in the treatment phase, communicating that the intervention effectively led to the amelioration or extinction of a problem behavior. However, distributions may more closely align with the normality assumption when treatment phase counts are higher than 0, or the percentage variable has a mean closer to 50%. Available evidence indicates that design-comparable effect sizes can tolerate a moderate degree of non-normality for relatively simple model specifications (**?**). However, additional research is needed to determine how much non-normality can be present before there are substantial consequences for the design-comparable effect size.

2.4.3 Homogeneity of Variance

In addition to assuming normality, the illustrations of design-comparable effect size estimation provided in this guide all assume that the within-case variation is homogeneous across phases and cases within a study. Although assumptions of homogeneity across cases and phases may be reasonable, there are situations when researchers should not assume homogeneity of variance. Consider again the **?** study, with the multiple baseline design graphs in Figure **??**. Results of our visual analysis suggest that the variance differs between the baseline and treatment phases, with less variation in the treatment phase as the percentage of off-task behaviors decreased and approached 0 (i.e., extinction). With count-based variables (e.g., raw counts or counts converted to percentages), variability often depends on the variable mean. Thus, treatments that shift the mean tend to change the variance. If studies have unstructured baselines, substantial variability is common. If there is tight experimental control in the intervention phase (e.g., researchers predict and control for interventionist/peer attention, strong treatment fidelity), we might expect some reduction in variance. With studies like these, where the variance differs between the baseline and treatment phases, we recommend estimating a more complex multilevel model that yields separate variance estimates for the two phases rather than assuming homogeneity across phases. Such models are feasible to estimate using the tools presented in subsequent chapters, but are beyond the scope of the guide. Until future research provides more concrete guidance about the best ways to proceed when encountering between-case heterogeneity, meta-analysts must remain aware of their assumptions and transparent about analytic decisions when reporting methods and results.

Upon discovering substantial violations to the normality or homogeneity assumptions, we encourage researchers to consider whether violations to the exchangeability assumption needed for synthesis are also present. For example, imagine that meta-analysts interested in synthesizing the effects of oral narrative interventions select a multiple baseline design study that has all cases reporting baseline observations consistently at or near 0. It is likely that both the normality and homogeneity assumptions are violated for this study. It is also likely that the outcome used in the hypothetical study differs greatly from the outcomes used in the included group design studies. This leads to questions about the exchangeability of the effect sizes. In such circumstances, we advise against trying to force the computation of a design-comparable effect size and the synthesis of SCD and group design studies together. A more appropriate synthesis option may be to meta-analyze the SCD and group design studies separately, allowing for the use of a more appropriate effect size metric for the SCD studies separate from that used for the group design studies. However, we expect that normality and homogeneity of variance assumption violations will often be more modest (or non-existent) than in the above example, so that researchers can continue to entertain the use of the design-comparable effect size.

2.4.4 Appropriate Structural Model

The estimation of design-comparable effect sizes requires assumptions about the structural model for the data series collected in the SCD. Ideally, these assumptions are based on content expertise, knowledge of the intervention domain, and understanding of the dependent variable(s) under review. The assumptions also rely on visual analysis and calculation of descriptive statistics from the studies. Regarding baseline trend, given our knowledge and understanding of the behavior(s), context(s), and participants included in the studies, we can assume one of three things: no trend in baseline, a linear trend in baseline, or some form of nonlinear trend. Similarly, we can use the same knowledge to make assumptions about data trends in the treatment phase: no trend, linear trend, or nonlinear trend. Furthermore, we can make assumptions about the parameters defining the baseline trajectory (e.g., level and slope) and the change in trajectory with treatment (e.g., the change in level and change in slope)—they either differ across cases within the study, or we can assume that some of these parameters are the same across cases. Purposeful consideration of our included SCDs is likely to lead to more accurate effect size estimation. If we do not select a structural model consistent with our data, we can expect biased design-comparable effect size estimates.

As an example, we present the study of a writing intervention for post-secondary adults with intellectual and developmental disabilities that targeted the improvement of sentence construction (**?**). Figure **??** is a graphical depiction of the study design and outcome data. Visual analysis of the baseline phases suggests potential baseline trends in accurate writing sequences. Therefore, it

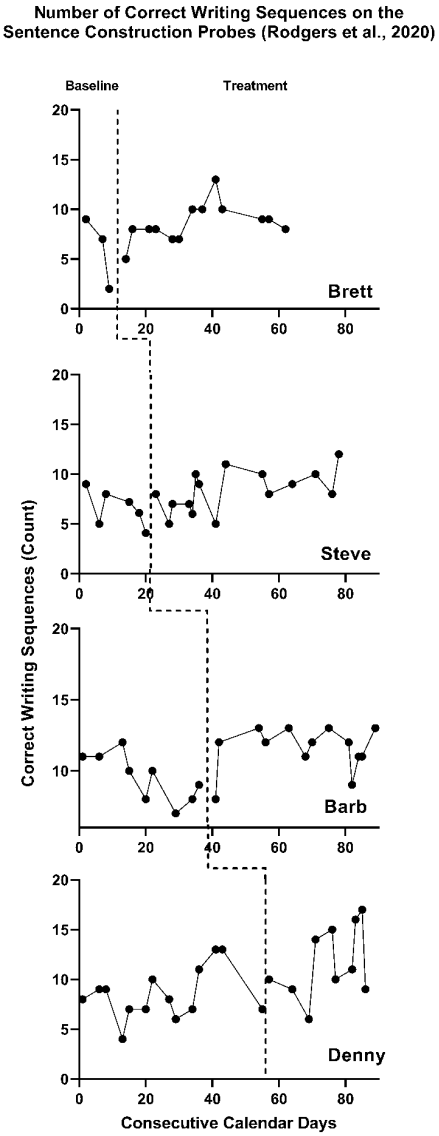


Figure 2.2: Multiple Baseline Across Participants (Rodgers et al., 2020)

would be appropriate to specify a model that assumes a trend in baseline. For participant Denny, our visual analysis of observed correct writing sequences suggests an increasing baseline trend, represented by the solid line in Figure **??** (estimated using ordinary least squares regression). In contrast, if we selected a baseline model assuming no trend, Denny's projected baseline of no trend is considerably different than the projected baseline assuming a linear trend. (In Figure **??**, note the difference between the dotted lines representing the baseline trends projected through the treatment phase). The effect estimate would be larger if we did not model the baseline trend in Figure **??** because the observed treatment values are further above the projection based on no trend than the projection based on trend. Thus, assumptions about trends have consequences for the effect size estimates.

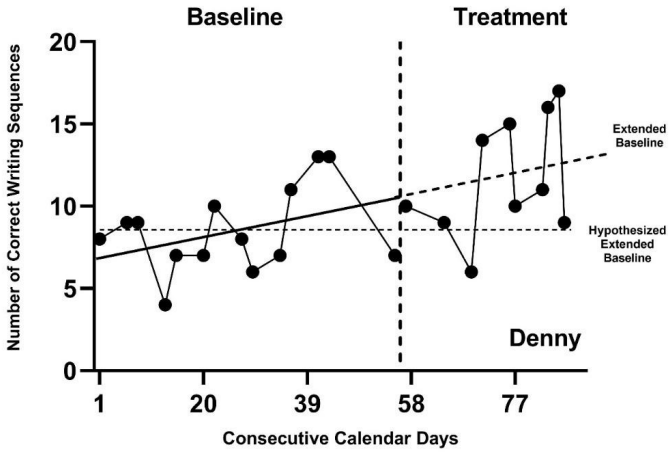


Figure 2.3: Hypothesized and Projected Baselines for Denny (Rodgers et al., 2020)

To select an appropriate structural model, we recommend that meta-analysts start by carefully considering the discipline or intervention under study for the research synthesis, including but not limited to the outcome of interest, the participants included in the studies, and what prior research says about what can be expected regarding trends in absence of intervention and data trends in the treatment phase. For example, a review of research on the use of positive behavioral interventions and supports (PBIS) over time show some efficacy in the reduction of both in-school and out-of-school student suspensions for Black students and students with disabilities (see **?**, for a review). Using this as a basis for understanding the nature of the interventions, contexts, and populations, we can reasonably assume future PBIS research would report similar responses to the interventions if the research contexts are similar. However, rather than make such a broad assumption without verifying it, we recommend

that meta-analysts visually analyze all data for studies included in the synthesis and consider the degree to which the data from the studies are reasonably consistent with prior expectations. If data from the studies are consistent with trend expectations (e.g., a learning curve associated with the development of a new skill), these assumptions can be used to select among the design-comparable effect size modeling options. Misalignment between a priori assumptions and actual observations across the studies' included data is likely to compromise the degree of confidence that we place in estimating effect sizes. If substantial inconsistencies exist between researchers' expectations for and actual observations of study data, results from any single modeling option become more suspect. In these situations, we encourage researchers to estimate the effects for each of the competing sets of trend assumptions to provide information on the sensitivity of the findings to the modeling decisions.

2.5 Modeling Options for Design-comparable Effect Size Estimation

After meta-analysts consider the tenability of the exchangeability assumption, we suggest consideration of normality and homogeneity of variance assumptions. If violations to these distributional assumptions are severe, researchers should reconsider whether the SCD and group design study outcomes are similar enough to assume exchangeability and warrant synthesis using the standardized mean difference effect size metric. If outcomes are substantially different across design types, it may be more reasonable to meta-analyze the SCD and group design studies separately and to use a different effect size metric for the SCD studies. When outcomes appear more similar and violations are not too severe, we suggest meta-analysts proceed with the design-comparable effect sizes and note findings. In their decision to proceed with design-comparable effect sizes, researchers should describe the range of characteristics of the SCD as well as the predominant SCD used in the area of synthesis. For example, do most SCD studies in the research area tend to use reversal designs (e.g., ABAB designs) or do studies predominantly use designs of multiple baseline and/or multiple probe across participants? When synthesizing reversal designs with design-comparable effect sizes, researchers are currently limited to models that assume stability (i.e., no trend). For multiple baseline or multiple probe designs, a variety of trend assumptions are feasible such as linear or quadratic.

When the synthesis includes predominantly multiple baseline and/or multiple probe designs, researchers should state their expectations about trends given their understanding of the participants, context, and outcome under study. We also recommend they visually inspect the graphs of the data from the primary studies, analyzing them for consistency with the trend expectations. Based on these considerations, it is helpful to determine which of the following sets of trend assumptions are most reasonable for the set of studies to be synthesized: (a) no trends in baseline or treatment phases, (b) no trends in baseline, but

trends in treatment phases, or (c) trends in baseline and differential trends in treatment. After clarifying the trend assumptions, researchers next need to clarify assumptions about variability in treatment effect across cases (e.g., Is the treatment effect expected to be the same for each case or are between-participant differences anticipated in the response to intervention?). Again, we rely heavily on logic models for the area of research and visual analyses of primary study data to determine if these data are reasonably consistent with the expectations.

Figure **??** arranges these considerations into a series of decision rules that researchers can use to select from one of six common models for design-comparable effect sizes. Although there are other possibilities for specifying a design-comparable effect size estimation model, these six models cover a wide range of scenarios that can be estimated with the *scdhlm* software application (**?**). In addition, we included models that have received the most attention in the methodological literature, as well as those that have been applied in meta-analyses of single-case data.

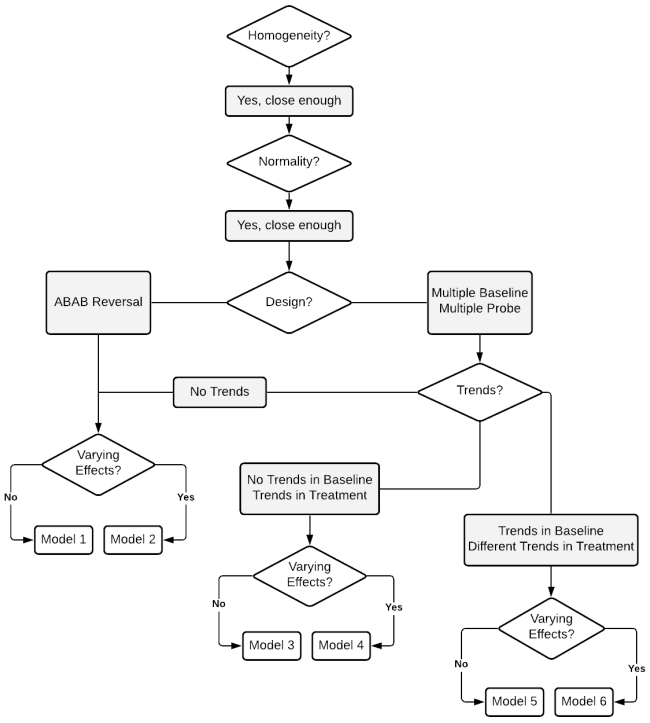


Figure 2.4: Flow Chart for the Selection of Design-comparable Effect Sizes

Chapter 3

Illustration of Design-Comparable Effect Sizes When No Trends Are Assumed

This chapter illustrates the computation of design-comparable effect sizes in contexts where we assume no trend in either baseline or treatment phases. We provide step-by-step instructions to demonstrate the selection and estimation of design-comparable effect sizes, using two multiple baseline studies, a replicated ABAB study, and a group design study. We use the scdhlm *app to show how to estimate the design-comparable effect sizes for the single-case studies and discuss estimating the effect size for the group study. Then, we illustrate how to combine the effects using a fixed effect meta-analysis.*

In this chapter, we illustrate the computation of design-comparable effect sizes based on models that do not include time trends (i.e., Models 1 and 2 in Figure **??**). These models assume that the dependent variable has a stable level throughout the baseline phase for each case and that introduction of treatment leads to an immediate shift in the level of the dependent variable. In this chapter, we pretend we are researchers who want to synthesize evidence from several single-case design (SCD) and group design studies that examine intervention effects on the improvement of math problem solving for students with disabilities. To model the process succinctly, we limit our illustration to four included studies comprised of two multiple baseline designs, a replicated ABAB design, and one group design. In one of the SCDs, **?** used a multiple baseline design across three small groups to examine the effect of a mathematics intervention for 12 students struggling with mathematical word problems. In another SCD study,

? used a multiple baseline design across four students with math word problem solving difficulties to examine the impact of a mathematics intervention. The third SCD study by **?** used an ABAB design replicated across nine participants with disruptive behaviors from two classrooms to examine intervention effects on math problem solving. The fourth included study by **?** used a group design where students with learning disabilities were randomly assigned to intervention and control conditions and students' math word problem solving accuracy was measured pre- and post-intervention.

Because our goal is to aggregate effects across single-case and group design studies, we consider design-comparable effect sizes (see the decision rules in Figure **??**). We break down this illustration into four stages: (1) selecting a design-comparable effect size for the SCD studies, (2) estimating the design-comparable effect size for the SCD studies using the *scdhlm* app (**?**), which is a web-based calculator for design-comparable effect sizes, (3) estimating the effect size for the group study, and (4) synthesizing the effect sizes across the included studies. We concentrate primarily on the first two steps, because well-developed methods for group studies are illustrated in detail elsewhere (e.g., **?**; **?**; **?**; **?**).

3.1 Selecting a Design-Comparable Effect Size for the Single-Case Studies

We first select an appropriate design-comparable effect size using the decision rules in Figure **??**. To formulate hypotheses about potential trends in baseline and/or treatment phases (i.e., if and why we expect those trends), we rely on previous experience and existing knowledge of the population, study contexts, and outcome of interest. Based on such, we assume that students in the studies would not improve mathematics skills without intervention. We also expect the interventions to abruptly change the outcomes, so we anticipate an immediate shift in the level of performance from baseline to treatment and no trend in the treatment phase. Finally, because we assume no trends in either baseline or treatment phases, we tentatively consider Model 1 or Model 2 from Figure **??**. These models are the only options for the ABAB design and are appropriate for multiple baseline designs when not anticipating trends.

Ideally, we make the choice between Model 1 and Model 2 based on our conceptual understanding and our a priori expectation that either the treatment effect would not vary from case to case (i.e., Model 1) or would vary from one case to the next (Model 2). Due to the participants' differential learning histories and existing skills or abilities, we anticipate some variation in the effect across cases. We tentatively select Model 2 as the best fit, as it is most consistent with our understanding of this context. To verify the tenability of our assumptions, we conduct a visual analysis of the SCD study graphs. We present the data extracted from each SCD study, with data from **?** shown in Figure **??**, **?** data in Figure **??**, and data extracted from **?** in Figure **??**.

Prior to examining the data to evaluate its consistency with the Model 2 trend assumptions, we consider whether the data are reasonably aligned with the homogeneity and normality assumptions underlying all design-comparable effect size models. Note that in Figures **??** and **??**, we generally see similar variation from one case to the next, and similar within-case variation across phases. We also fail to see any outliers that would call into question the normality assumption. However, in Figure **??**, there is a tendency for the treatment phases to have more variability than the baseline phases. Further, in Figure **??**, all observations for Student A4 have values near zero in the first baseline phase—inconsistent with our homogeneity and normality assumptions. Although it appears that some of the design-comparable effect size assumptions were violated, we proceed because the violations appear relatively minor and our combining effect sizes across single-case and group studies requires that we compute a design-comparable effect size for each study (or drop the study from the synthesis).

Next, given the data observed in the primary studies, we consider if a model absent of baseline trends appears reasonable. Across the cases in Figures **??**-**??**, the typical baseline pattern is one of no trend and is consistent with our expectations. However, there are a few exceptions. In Figure **??**, Ben's data appear to have an upward baseline trend with observed values of 2, 2, and 3. This visual trend is ambiguous; it could be an artifact of Ben just happening to solve one more math problem on Day 3 (i.e., chance). Yet, based on our previously stated expectations and an absence of baseline improvement for the other participants, we find it more reasonable to assume that Ben would continue to solve two or three problems per session (i.e., no trend) as opposed to continuing to improve by about one problem per day (i.e., linear trend). The former assumption is also consistent with the decision made by **?** to intervene with Ben. In other words, had the researchers thought Ben's math problem solving was improving in baseline, there would be less reason for Ben to receive the intervention.

We also consider other case examples. In Figure **??**, Gary's data appear to have an upward baseline trend and Kyrie's data appear to have a downward trend in baseline. These trends are inconsistent with each other, and with the typical pattern seen across the **?** study cases. As with Ben (see Figure **??**), we are not confident that the apparent trend would continue. If Gary's trend were to continue, there would be no need for intervention. If Kyrie's trend were to continue, we would predict negative percentages correct in the upcoming sessions, which is not possible. Thus, we proceed with the assumption of no baseline trend because the typical pattern for most cases in each study is consistent with our expectations of no trend, and because of the ambiguity surrounding the few possible exceptions.

Next, focusing on the treatment phases, we conduct similar analyses. In Figures **??** and **??**, where there is an immediate shift in performance and no trend in the treatment phases, we find patterns that match our hypothesized treatment trend expectations. In Figure **??**, we again see a typical pattern where the shift in performance is immediate and stable over time. While here there are a few

exceptions where cases appear to have a decline in performance toward the end of the intervention phase (e.g., Andy and Ellie), we find it best to select a model consistent with the typical and expected pattern. We proceed with a model that assumes no trend in baseline and treatment phases (i.e., Model 1 or Model 2 from Figure **??**) because the design-comparable effect sizes assume a common model across cases and estimate an average effect across cases.

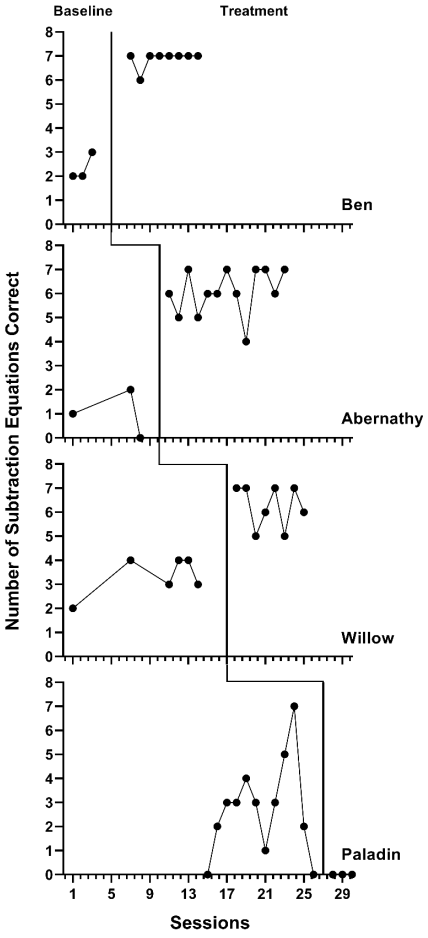


Figure 3.1: Multiple Baseline Data Extracted from Case et al. (1992)

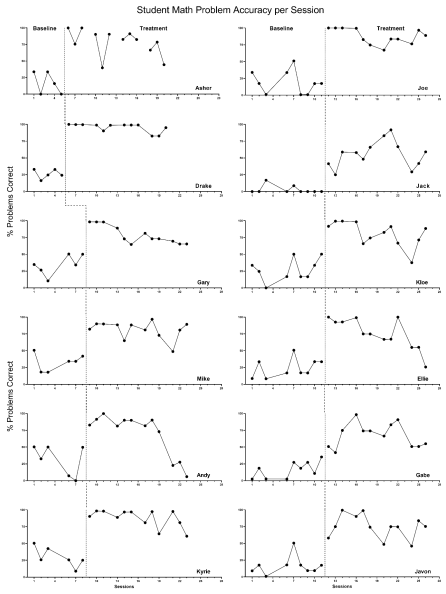


Figure 3.2: Multiple Baseline Data Extracted from Peltier et al. (2020)

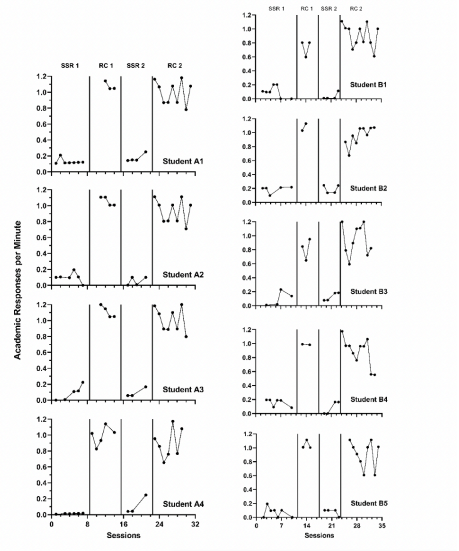


Figure 3.3: Replicated ABAB Data Extracted from Lambert et al. (2006)

3.2 Details of the No Trend Models for Design-Comparable Effect Sizes

To fully differentiate between Model 1 and Model 2, we present the formal specification of each. For both Model 1 and Model 2, we can write the within-case model as:

$$Y_{ij} = \beta_{0j} + \beta_{1j} Tx_{ij} + e_{ij}, \tag{3.1}$$

where Y_ij is the score on the outcome variable Y at measurement occasion i for case j, and Tx_{ij} is dummy coded with a value of 0 for baseline observations and a value of 1 for the treatment phase observations. The mean baseline level for case j is β_{0j} (see Figure **??** for a visual representation of β_{0j} and β_{1j}). The raw score treatment effect for case j is indexed by β_{1j}, which is the difference between the treatment phase outcome mean and the baseline phase mean. The error (e_{ij}) is time- and case-specific and assumed normally distributed and first-order autoregressive with variance σ_e^2.

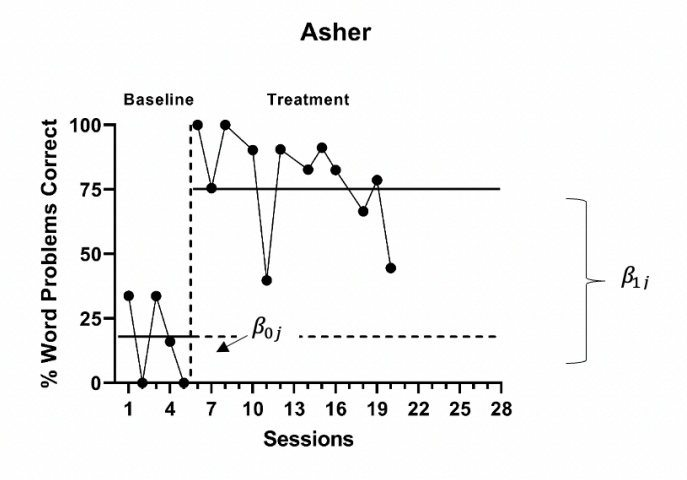


Figure 3.4: Illustration of Treatment Effect for Asher (Peltier et al., 2020)

For Model 1, the between-case model is:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{3.2}$$

$$\beta_{1j} = \gamma_{10} \tag{3.3}$$

where γ_{00} is the across-case average baseline mean and u_{0j} is a case-specific error, which is the deviation from the overall average for case j. We assume the error is normally distributed with variance $\sigma_{u_0}^2$. We assume the across-case average raw score treatment effect, γ_{10}, to be constant for all cases. Thus, there is no error term in the equation for β_{1j}. Based on this model, the design-comparable effect size is defined as the average raw score treatment effect (γ_{10}) divided by

a SD that is comparable to the SD used to standardize mean differences in group-design studies (**?**):

$$\delta = \frac{\gamma_{10}}{\sqrt{\sigma_{u_0}^2 + \sigma_e^2}} \tag{3.4}$$

The Model 2 specification is like Model 1, with the only difference being an error term (u_{1j}) added to Equation (**??**) to account for between-case variation in the treatment effect. More specifically:

$$Y_{ij} = \beta_{0j} + \beta_{1j} Tx_{ij} + e_{ij}, \tag{3.5}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{3.6}$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \tag{3.7}$$

Again, the across-case average baseline mean is γ_{00} and the across-case average raw score treatment effect (i.e., difference in treatment and baseline phase means) is γ_{10}. The case-specific errors (u_{0j} and u_{1j}) account for between-case differences in baseline level and response to treatment. They are assumed multivariate normal with covariance $\Sigma_u = \begin{bmatrix} \sigma_{u_0}^2 & \\ \sigma_{u_1 u_0} & \sigma_{u_1}^2 \end{bmatrix}$. We define the design-comparable effect size exactly as in Equation (**??**), because the effect size is scaled by the SD of the outcome in the absence of intervention and not dependent on the addition of u_{1j}, which only impacts between-case variance in the treatment phase.

Because we tentatively selected Model 2 based on our a priori considerations, we use it to illustrate the estimation of design-comparable effect sizes for this data set. For several reasons, we also contrast the Model 2 results to what we obtain from Model 1. First, contrasting Model 1 and 2 results provides us with an additional way to examine the empirical support for our model selection. For instance, we could report that visual analyses of the model-implied individual trajectories for Model 2 fit the raw data better than the model implied individual trajectories from Model 1. Second, the contrast allows us to examine the sensitivity of our effect size estimates to the model chosen. For instance, this contrast may lead us to rule out between-case effect size variation as having a large impact on the estimated design-comparable effect size. Finally, contrasting allows us to illustrate a method of selecting between models in circumstances where a priori information is not sufficient.

3.3 Estimating the Design-Comparable Effect Size for the Single-Case Studies

3.3.1 Example 1: Multiple Baseline Study by ?

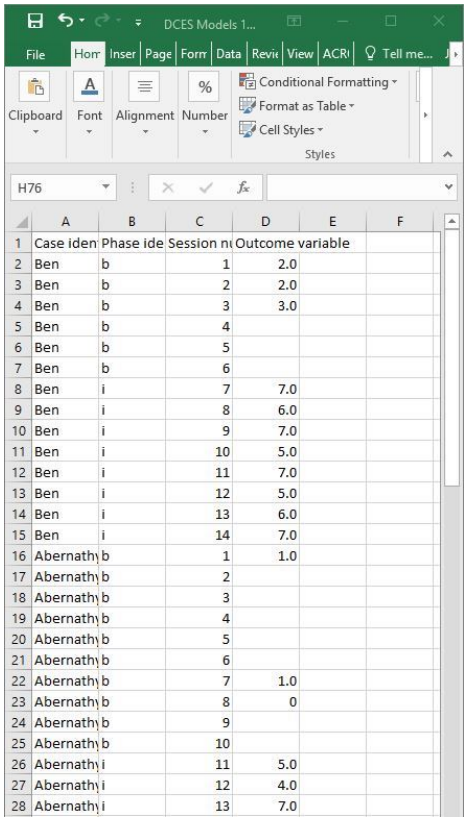
We can estimate design-comparable effect sizes for all models suggested in Figure **??**, using a web-based calculator for design-comparable effect sizes (*scdhlm*;

?). The scdhlm app is available at https://jepusto.shinyapps.io/scdhlm/. To use this app, researchers must store their dataset in an Excel file (.xlsx), comma delimited file (.csv), or text file (.txt). In addition, we recommend that users inspect their data to ensure the inclusion of the following variables: case identifier, phase identifier, session number, and the outcome. Although not required, researchers may want to arrange the data columns by order of variable appearance in the app. We show this arrangement for the **?** study in Figure **??**. There, we demonstrate the following order of column headers with case identifier representing data in the first column, followed by variables in this order (left-to-right): phase identifier, session number, and the outcomes in the fourth column. Researchers have the flexibility to use any labeling scheme that clearly distinguishes between baseline and intervention conditions. For example, for the phase identifier, one can use b or 0 to indicate baseline observations and i or 1 to indicate intervention observations. However, the app requires that numerical values be used for both session number and outcome. Finally, we recommend that users arrange the data first by case (i.e., enter all the rows of data for the first case before any of the rows of data for the second case) and then by session number.

After starting the app, we use the Load tab to load the data file, as illustrated in Figure **??**. As mentioned previously, the data file can be a .txt or .csv file that includes one dataset, or an Excel (.xlsx) file that has either one (e.g., a data set for one study) or multiple spreadsheets (one spreadsheet for each of several studies). If using a .xlsx file with multiple spreadsheets, the *scdhlm* app allows us to select the spreadsheet containing the data for the study of interest from the *Load* tab. Then, we use the drop-down menus on the right of the screen to indicate the study design (*treatment reversal* versus *Multiple Baseline/Multiple Probe across participants*) and which variables in the data set correspond to the case identifier, phase identifier, session, and outcome (see Figure **??**).

After loading our data, we use the *Inspect* tab to ensure the accurate import of raw data into the app and assigned variable names are accurate (Figure **??**). In addition, we can use the *Inspect* tab to view a graph of the data (Figure **??**). We recommend that researchers compare these data with the graphed data from the original studies to ensure accuracy in uploading the data and specifying the design and variable names on the *Load* tab. Check the graphed data again for consistency with the (tentatively) selected model for the design-comparable effect size.

After inspecting the data, we next specify the model for the design-comparable effect size using the *Model* tab. Figure **??** shows our specification for Model 2 (i.e., the model that assumes no trends and an effect that varies across cases). The specification process begins with the selection of trend type for the baseline phase. For this example, under *Type of time trend*, we select *level* because we assume no time trends in the baseline phases. Then, we opt to include *level* as a fixed effect and check the box to enable the model estimation of the average baseline level (i.e., γ_{00} from Equation (**??**)). We also include *level* as a random



	A	B	C	D	E	F
1	Case iden	Phase ide	Session n	Outcome variable		
2	Ben	b	1	2.0		
3	Ben	b	2	2.0		
4	Ben	b	3	3.0		
5	Ben	b	4			
6	Ben	b	5			
7	Ben	b	6			
8	Ben	i	7	7.0		
9	Ben	i	8	6.0		
10	Ben	i	9	7.0		
11	Ben	i	10	5.0		
12	Ben	i	11	7.0		
13	Ben	i	12	5.0		
14	Ben	i	13	6.0		
15	Ben	i	14	7.0		
16	Abernathy	b	1	1.0		
17	Abernathy	b	2			
18	Abernathy	b	3			
19	Abernathy	b	4			
20	Abernathy	b	5			
21	Abernathy	b	6			
22	Abernathy	b	7	1.0		
23	Abernathy	b	8	0		
24	Abernathy	b	9			
25	Abernathy	b	10			
26	Abernathy	i	11	5.0		
27	Abernathy	i	12	4.0		
28	Abernathy	i	13	7.0		

Figure 3.5: Illustration of Treatment Effect for Asher (Peltier et al., 2020)

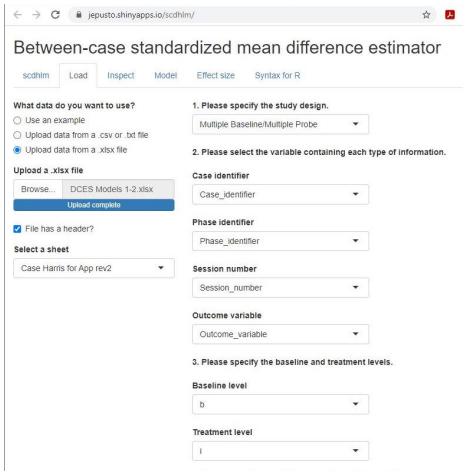
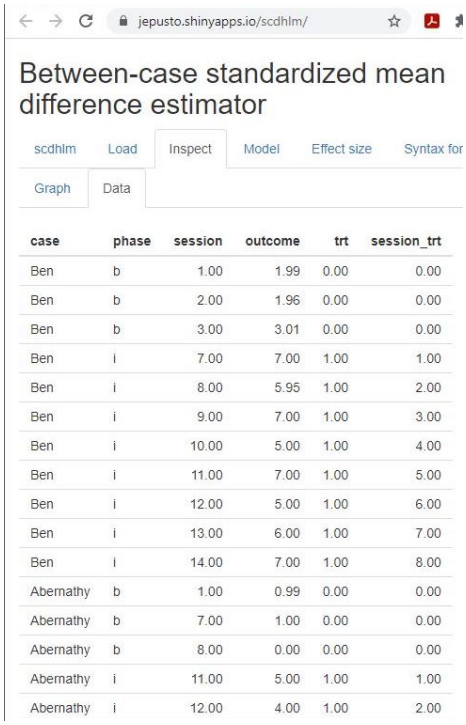


Figure 3.6: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Load Tab



case	phase	session	outcome	trt	session_trt
Ben	b	1.00	1.99	0.00	0.00
Ben	b	2.00	1.96	0.00	0.00
Ben	b	3.00	3.01	0.00	0.00
Ben	i	7.00	7.00	1.00	1.00
Ben	i	8.00	5.95	1.00	2.00
Ben	i	9.00	7.00	1.00	3.00
Ben	i	10.00	5.00	1.00	4.00
Ben	i	11.00	7.00	1.00	5.00
Ben	i	12.00	5.00	1.00	6.00
Ben	i	13.00	6.00	1.00	7.00
Ben	i	14.00	7.00	1.00	8.00
Abernathy	b	1.00	0.99	0.00	0.00
Abernathy	b	7.00	1.00	0.00	0.00
Abernathy	b	8.00	0.00	0.00	0.00
Abernathy	i	11.00	5.00	1.00	1.00
Abernathy	i	12.00	4.00	1.00	2.00

Figure 3.7: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Data Tab within Inspect Tab for Case et al. (1992)

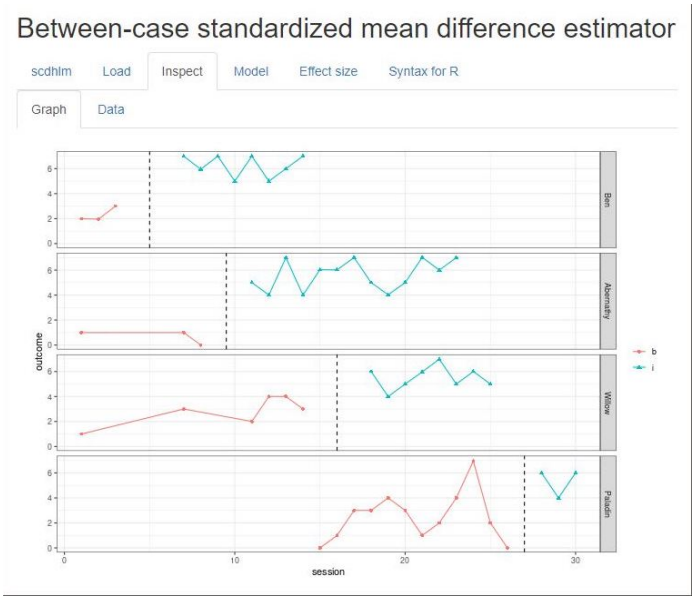


Figure 3.8: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Graph Tab Within Inspect Tab for Case et al. (1992)

effect, so that the baseline level can vary from case to case. Focusing on the treatment phase next, we select *change in level* as the *Type of time trend*, because we assume that the treatment will only change the level of the outcome (not trend) and include *change in level* as a fixed effect. As a fixed effect, we can obtain an estimate of the average shift in level between baseline and treatment phases (i.e., γ_{10} from Equation (**??**). Finally, we choose to include *change in level* as a random effect to allow the change in level (i.e., treatment effect) to vary from case to case. Note the *scdhlm* app allows us to make different potential assumptions about the correlation structure of the session-level errors. Shown in Figure **??** are the default options of auto-regautoregressive and constant variance across phases. These defaults match the model presented in Equation (**??**). At this point, our model for the design-comparable effect size matches Model 2 from Figure **??**.

At the bottom of the screen, the *scdhlm* app provides a graph of the data with trend lines based on the specified model. We recommend that users inspect this graph to ensure that the trend lines fit the data reasonably well. If the trend lines do not fit the data well, a question about model selection arises. For example, we find the baseline trend line for Abernathy is relatively high compared to their actual baseline observations. However, other model specifications (e.g., Model 1 shown later) did not improve the fit of Abernathy's baseline trend line. So, we decide to proceed with estimation because most other trend lines look appropriate, and the model is consistent with our a priori expecta-

tions. Throughout this process, researchers should maintain a high standard for transparency in decision-making when reporting methods and results.

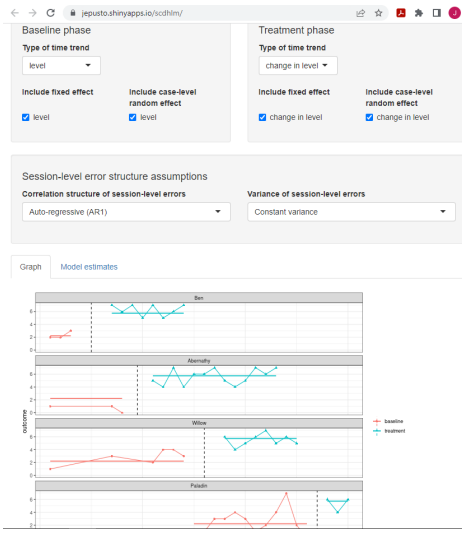


Figure 3.9: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Model Tab Showing Model 2 Specification for Case et al. (1992)

For the **?** data set, the a priori-identified Model 2 provides trajectories that fit the data reasonably well[1]. Thus, we proceed to the *Effect size* tab. As shown in Figure **??**, the estimated design-comparable effect size for this study is 2.57 with a standard error (SE) of 0.45 and 95% confidence interval (CI) [1.65, 3.50].

Additional information reported on the *Effect size* tab include estimates of other quantities from the model, information about the model specification, and the assumed time-points used in calculating the design-comparable effect size. The reported degrees of freedom are used in making a small-sample correction to the effect size estimate, analogous to the Hedges' g correction used with group designs (**?**). Larger estimated degrees of freedom imply more precision in estimating the denominator of the design-comparable effect size, making the small-sample correction less consequential. Conversely, smaller degrees of freedom are indicative of imprecise design-comparable effect size denominator estimation, making the small-sample correction more consequential. The *Effect size* tab, shown in **??**, also reports autocorrelation, which is the estimate of the correlation between level-1 errors of the model for the same case, differing by one time point (i.e., session) based on a first-order autoregressive model. Given the selected follow-up time, the reported intra-class correlation is an estimate of the between-case variance of the outcome as a proportion of the total variation in the outcome (including both between-case and within-case variance). Larger

[1] Meta-analysis must specify their criteria for "reasonably well".

values of the intra-class correlation indicate that more of the variation in the outcome is between participants. The remaining output information (*Study design*, *Estimation method*, *Baseline specification*, *Treatment specification*, *Initial treatment time*, *Follow-up time*) describe the model specification and assumptions used in the effect size calculations. The *scdhlm* app includes such to allow for reproducibility of the calculations.

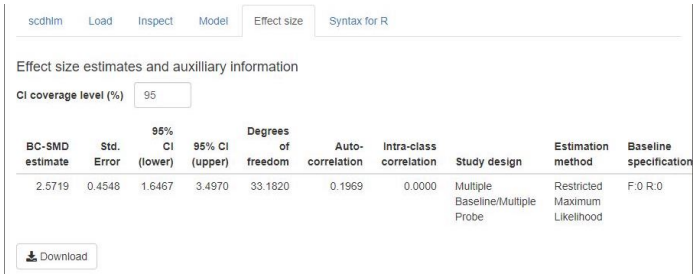


Figure 3.10: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Effect size Tab Showing Model 2 Estimate for Case et al. (1992)

If our a priori arguments for selecting Model 2 were weak, or we wanted to consider Model 1, we can easily compute the results of a different model by going back to the *Model* tab and changing our specification. The only difference is that Model 1 does not include *change in level* as a random effect for the treatment phase. For illustration purposes, we present Model 1 specification results for this study in Figure **??**. As before, we keep the *level* option selected as the *Type of time trend* for the baseline phase, as both a fixed effect and a random effect. In the treatment phase, we keep *change in level* as the *Type of time trend* but select only *change in level* as a fixed effect. The Model 1 trend lines fit similarly to those from Model 2; the Model 1 design-comparable effect size is 2.59 with an SE of 0.41 and 95% CI [1.76, 3.42]. This suggests that the effect size estimate is not sensitive to our decision about whether the treatment effect varies across cases. Despite this information, we proceed with the Model 2 estimate because it is consistent with our expectations for the research in this area.

3.3.2 Example 2: Multiple Baseline Study by ?

We now have a design-comparable effect size for the first single-case study by **?**. Next, we repeat these steps for all other SCD studies in our synthesis. For this illustration, we include a second multiple baseline study (**?**). For this second study, we ran through the same sequence of steps: 1. load the data; 2. inspect the data in both tabular and graphic form; 3. specify our selected model for the data (i.e., Model 2 for this illustration); and 4. estimate the design-comparable effect size.

After performing these steps, we found that the estimated model trajectories

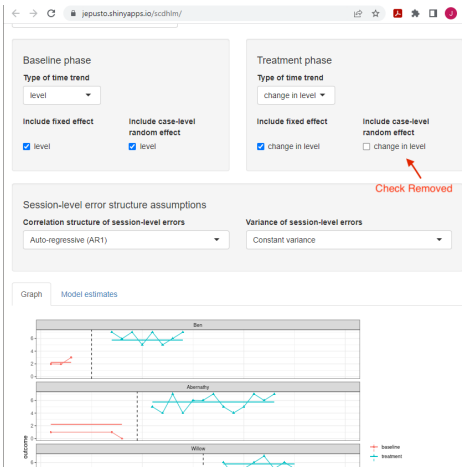


Figure 3.11: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Model Tab Showing Model 1 Specification for Case et al. (1992)

for the Model 2 specification fit the **?** data well, as shown in Figure **??**. For the **?** study, the design-comparable effect size is 2.95 with an SE of 0.28 and 95% CI [2.42, 3.53]. We then estimated the effect for Model 1 (i.e., we remove the check next to *change in level* under *include random effect*), which is the more restrictive model that does not allow the treatment effect to vary across cases. Model 1 produced a design-comparable effect size estimate of 2.83 with an SE of 0.24 and 95% CI [2.34, 3.31], like the Model 2 design-comparable effect size.

3.3.3 Example 3: Replicated ABAB Design by ?

For the third SCD study, we illustrate use of the *scdhlm* app using data from a replicated ABAB design by **?**. As with the previous two SCD studies, we load our spreadsheet containing the study data in the usual manner. However, unlike **?** and **?**, the **?** study does not utilize a multiple baseline design. Therefore, on the right side of the *Load* tab, using the drop-down menu under *Please specify the study design*, we must select *Treatment Reversal* (as opposed to *Multiple baseline/Multiple probe across participants*). After loading the data, we again use the *Inspect* tab to visually inspect the data in both tabular and graphic form. Then using the *Model* tab (shown in Figure **??**), we continue to specify Model 2 by checking the option under the *Baseline phase* to include *level* as both a fixed effect and a random effect and checking under the *Treatment phase* to include *change in level* as both a fixed effect and a random effect. Users of the app will note that for reversal designs, there is no drop-down menu from which they can potentially add trends (i.e., *level* is the only option for baseline specification, and *change in level* is the only option for treatment phase specification). This makes specification using the *Model* tab simpler than in our previous examples of multiple baseline studies, although it does also constrain the user's ability to

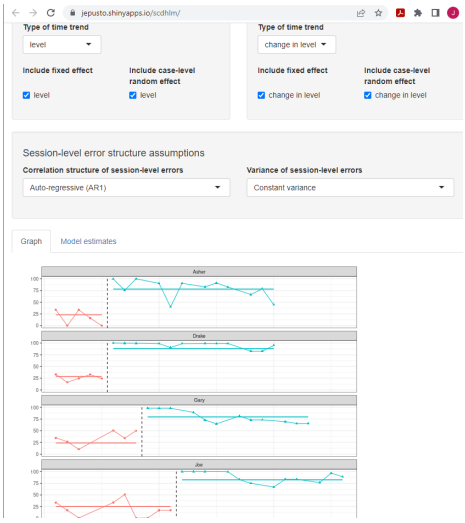


Figure 3.12: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Model 2 Specification for Peltier et al. (2020)

specify a well-fitting model. On the *Model* tab, we see that the fitted trajectories fit these data well.

After specifying the model, we view the Effect size tab to obtain the design-comparable effect size (see Figure **??**). For **?**, the estimated Model 2 design-comparable effect size is 6.37 with an SE of 0.39 and 95% CI [5.60, 7.14]. Like the previous SCD studies, we also estimate the effect for Model 1, which is the more restrictive model that does not allow the treatment effect to vary across cases. Model 1 produces an estimate of 6.34 with an SE of 0.37 and 95% CI [5.61, 7.08], a similar result to the Model 2 design-comparable effect size.

3.4 Estimating the Design-Comparable Effect Size for the Group Studies

After estimating the design-comparable effect size for each SCD study, we turn our attention to estimating the design-comparable effect size for each group study. **?** used random assignment of individual students with learning disabilities to either the intervention ($n = 12$) or control ($n = 8$) conditions. After intervention, both groups of students completed 25 math word problems selected from the *British Columbia Mathematics Achievement Test*. Because details for estimating standardized mean difference effect sizes from group studies are readily available from a variety of sources including chapters (**?**), books (e.g., **?**) and journal articles (e.g., **??**), we do not model the calculation procedures and simply report the results. Using the **?** group design study summary

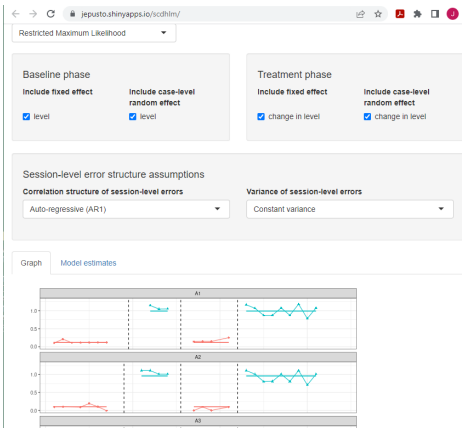


Figure 3.13: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Model 2 Specification for Lambert et al. (2006)

statistics at posttest, we calculate a standardized mean difference in math word problem solving as 0.71, with an SE of 0.45 based on Hedges' g to correct for small-sample bias.

3.5 Analyzing the Effect Sizes

After we have obtained effect sizes from each single-case and group study, we can proceed with synthesizing the effect sizes. Depending on our synthesis goals, we have a variety of tools and approaches available. We can: (a) create graphical displays (e.g., forest plots) to show the effect size for each study along with their confidence interval, (b) average the effect sizes and create a confidence interval for the overall average effect, (c) estimate the extent of variation in effects across studies, (d) examine the effect sizes for evidence of publication bias, and (e) explore potential moderators of the effect. Since the use of design-comparable effect sizes for the SCD studies produce estimates having the same metric as the commonly used standardized mean difference effect sizes from group studies, researchers can accomplish these goals (e.g., averaging the effect sizes or running a moderator analysis) using methods developed for group studies. Details on these methods are readily available elsewhere (e.g., **??**). We illustrate the averaging of the effect sizes from our studies here using a fixed effect meta-analysis, so that this illustration is consistent with the approach used in What Works Clearinghouse intervention reports (**?**).

Table **??** reports the effect size estimates, SEs, and fixed effect meta-analysis calculations for our four included studies. In fixed effect meta-analysis, the overall average effect size estimate is a weighted average of the effect size estimates from the individual studies, with weights proportional to the inverse of the sampling variance (squared SE) of each effect size estimate. Further, the SE

Table 3.1: Fixed Effect Meta-Analysis Calculations for Example Math Invervention Studies

Study	Effect Size Estimate (A)	Standard Error (B)	Inverse-variance Weight (%) (C)
Case et al. (1992)	2.57	0.45	4.94 (16.9)
Peltier et al. (2020)	2.95	0.28	12.76 (43.7)
Lambert et al. (2006)	6.37	0.39	6.57 (22.5)
Hutchinson (1993)	0.71	0.45	4.94 (16.9)
Fixed effect meta-analysis	3.28	0.19	29.21 (100)

of the overall effect size is the square root of the inverse of the total weight. In Table **??**, column C reports the inverse variance weight assigned to each study, with the percentage of the total weight listed in parentheses. For instance, the effect size estimate from **?** receives 43.7% of the total weight, while the effect size estimates from **?** and **?** each receive 16.9% of the total weight. The total inverse variance weight is 29.21. The overall average effect size estimate is 3.28 with an SE of 0.18 and an approximate 95% CI [2.91, 3.64]. The Q-test for heterogeneity is highly significant, $Q(3) = 99.3$, $p < .0001$, indicating that the included effect size estimates are more variable than we would expect due to sampling error alone. In other words, it is unlikely that we would observe such a wide dispersion of effect size estimates if the studies were all estimating the same true effect size parameter.

In fixed effect meta-analysis, the overall average effect size estimate is a summary of the effect size estimates across the included studies, where studies are treated as fixed. Therefore, the SE and CI in fixed effect meta-analysis account for the uncertainty in the process of effect size estimation that occurs in each of the individual studies. However, they do not account for uncertainty in the process of identifying studies for inclusion in the meta-analysis (**??**), nor do they provide a basis for generalization beyond the included studies. When conducting syntheses of larger bodies of literature—and especially of studies with heterogeneous populations, design features, or dependent effect sizes—researchers will often prefer to use random effects models (**?**) or their further extensions (**??**).

Of the studies we use as illustrative examples in this chapter, the dependent variable of the group study (**?**) was the broadest. Conversely, the **?** and **?** MB studies used more narrowly defined dependent variables. Finally, we note the outcome from the replicated ABAB design (**?**) included a behavioral component (i.e., it is not purely academic). In a synthesis of many studies, researchers might use moderator analysis (i.e., meta-regression analysis) to explore the extent to which variation in effect size is related to dependent variable characteristics or other study features. Methods for conducting such moderator analysis are described elsewhere (**?**, Chapters 19-21; **?**).

Chapter 4

Illustration of Design-Comparable Effect Sizes When Assuming Only Trends in The Treatment Phases

This chapter illustrates the computation of design-comparable effect sizes in contexts where one assumes trends in the treatment phase only, with no trends in the baseline phase. Using data from two multiple baseline studies and a group study, we provide step-by-step instructions for selecting design-comparable effect sizes for the single-case studies and estimating them using the scdhlm *app. We also briefly discuss estimating the effect size for the group study and synthesizing the effect sizes across the single-case and group and design studies.*

In this chapter, we demonstrate the computation of design-comparable effect sizes using models with trends in the treatment phase(s) online (i.e., no baseline trend), as represented by Models 3 and 4 in Figure **??**. These models assume that: (a) baseline observations of the dependent variable for each case occur at a stable level and (b) intervention leads to a change in both level and *trajectory* or growth rate in the treatment phase.

Let us consider a hypothetical scenario where we are interested in synthesizing single-case design (SCD) and group design studies that report intervention effects for reducing the amount of time it takes individuals to fall asleep at night (i.e., sleep onset latency). For purposes of illustration, we consider a review comprised of just three studies, including two multiple baseline designs

and one group design. The first SCD study (**?**) investigated the effects of two parent-implemented interventions (positive routines versus fading routines) on reducing sleep onset latency for six child participants with autism. Our second included SCD study, **?**, evaluated the effects of a sleep intervention on reducing the sleep latency of seven adult participants with intellectual disabilities. Our third study, a group design by **?**, compared the efficacy of two sleep intervention delivery methods (face-to-face or booklet) for 66 children with learning disabilities (aged 2–8 years) randomly assigned to either treatment or waitlist control groups.

We recommend that researchers use the decision rules in Figure **??** to select a method to estimate and synthesize effects, and to consider design-comparable effect sizes when the purpose is to aggregate effects across SCD and group design studies. For Models 3 and 4 and using the aforementioned set of studies for our hypothetical synthesis, we break down the process of model selection and estimation into four steps: 1. Selecting a design-comparable effect size for the SCD studies. 2. Estimating the design-comparable effect size for the SCD studies using the *scdhlm* app (**?**). 3. Estimating the effect sizes for the group studies. 4. Synthesizing the effect sizes across single-case and group studies.

We focus on the first two steps because guidance and illustrations of methods for group studies are readily available from other sources (**??**).

4.1 Selecting a Design-Comparable Effect Size for the Single-Case Studies

We begin by using the decision rules in Figure **??** to guide our selection of an appropriate design-comparable effect size. Across the studies, we use our knowledge of the population, context, and outcome of interest to inform any assumptions about the presence or absence of trends in either baseline and/or treatment phases. For our included studies, we assume that all participants generally have stable but high levels of behavior impeding a healthy nighttime routine (e.g., long latency to onset of sleep, frequent nighttime awakenings, etc.). We also assume no therapeutic trend in the baseline phase for each participant because otherwise there would be no need for the participants to receive the intervention. Further, we postulate that intervention will result in improved sleep behavior over time, as noted by a reduction in sleep problems in the treatment phase. Because we predict trend in only the treatment phase but not the baseline phase, we tentatively consider using Model 3 or Model 4 (see Figure **??**).

Ideally, we would make the choice between Model 3 and Model 4 based on a logic model and a priori expectation of treatment effect variation (or lack thereof) within and across cases. For example, if we hypothesize that treatment effects do not vary from one case to the next, we will specify our design-comparable effect size using Model 3. Alternatively, if we assume that effects do vary across

cases, we will use Model 4. For this synthesis, we note that all studies investigated sleep problems in the context of a participant's own home. Furthermore, primary caregivers delivered the interventions, and we assume they have different learning histories and backgrounds. Therefore, we believe it is reasonable to anticipate some variation in the effect across cases and tentatively select Model 4.

The next step in the selection of design-comparable effect sizes for SCDs requires us to conduct visual inspection of each SCD study graph to see if our trend assumptions hold. We extracted data from the **?** study and present the multiple baseline across participant graphs in Figure **??** (three participants who received the bedtime fading intervention) and Figure **??** (three participants who received the positive routines intervention). In Figure **??**, we extracted data from the **?** study and present the multiple baseline design graph for the seven participants with the target outcome of reducing sleep onset. When visually analyzing the graphs, we consider whether the data are reasonably consistent with the homogeneity and normality assumptions underlying any of the models for design-comparable effect sizes. In Figures **??**-**??**, we observe similar variation between cases and variation across phases within each case and conclude that the assumption of homogeneous level-1 variance is tenable. In addition, we find no severe outlying data to suggest clear departures from non-normality. Therefore, our initial decision to use design-comparable effect sizes appears appropriate.

Next, for our included studies, we consider the appropriateness of Models 3 and 4 that assume no trends in baseline. Visual inspection of the participants' graphs in Figure **??** (**?**) appears to reinforce our assumption of no trend in baseline data. Similarly, as shown in Figure **??** (**?**), sleep onset latency neither increased nor decreased in the baseline condition for most participants. However, there are some exceptions to baseline stability observed. In Figure **??**, Case D appears to show some improvement during baseline. In Figure **??**, Niahm, Mary, and Thomas appear to have worsening sleep problems during the baseline phase. These trends are inconsistent with the typical pattern seen across cases in Figures **??** and **??**. Whether there are truly trends that would continue is questionable. Perhaps the trends are artifacts related to fidelity or reliability of parental data collection, and thus would not continue.

In similar situations, we recommend that researchers reflect upon and reconsider their model selection. Using this specific example, it seems odd to assume that simply enrolling in a study would affect sleep patterns well established within the participant's home routine. Rather, we find it more reasonable to assume that sleep onset latency for individuals with disabilities would vary around an average level across baseline observations. Therefore, we proceed with Model 4 because most cases in the included studies have baseline data consistent with the expectation of no trend, and because of the ambiguity and questions surrounding the few possible exceptions.

We proceed with visual inspection of observations in the treatment phase after examining baseline phase observations. Across most cases, we note an immedi-

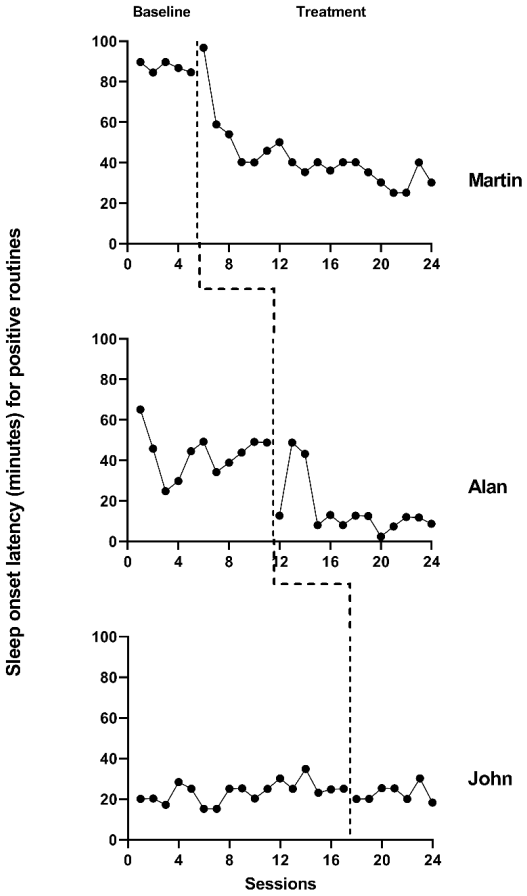


Figure 4.1: Multiple Baseline Data for Martin, Alan, and John (Delemere & Dounavi, 2018)

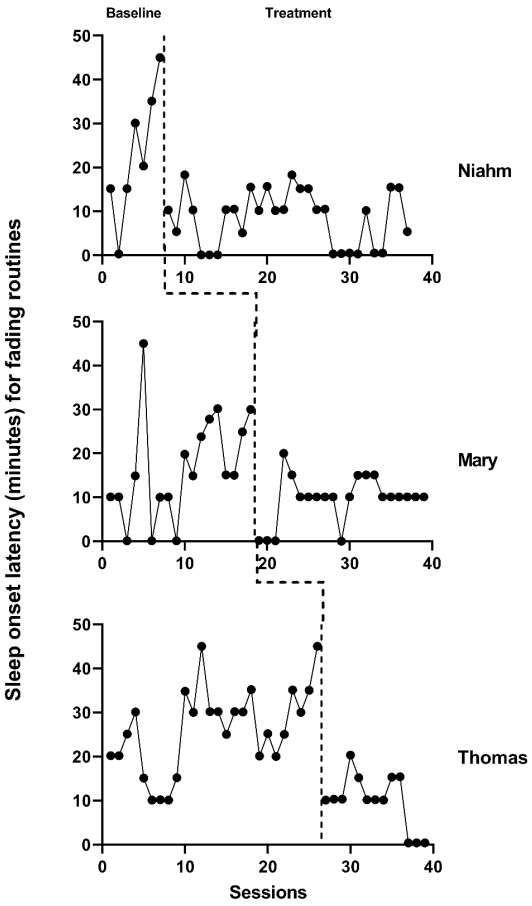


Figure 4.2: Multiple Baseline Data for Niahm, Mary, and Thomas (Delemere & Dounavi, 2018)

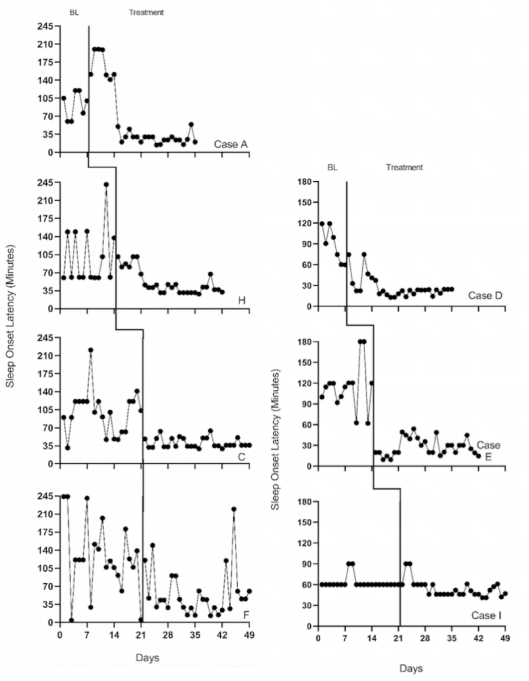


Figure 4.3: Multiple Baseline Design Data from 7 Participants with the Target Outcome of Reducing Sleep Onset Latency (Gunning & Espie, 2003)

ate change in the time it takes participants to fall asleep and a downward trend in treatment phase observations. While there are a few exceptions where participants appear to respond immediately to intervention, the downward trend does not continue across the remainder of the treatment phase (e.g., see Niahm, Mary, and Thomas in Figure **??**). Because the design-comparable effect size assumes a common model across cases and estimates an average effect across the cases, we find it best to select a model that is consistent with the typical and expected pattern. Thus, it appears reasonable to proceed with a model that assumes no trend in baseline conditions and a trend in the treatment phase (i.e., Model 3 or Model 4 from Figure **??**.

4.2 Details of the Models for Design-Comparable Effect Sizes

To appreciate how the treatment effect is defined in Models 3 and 4, let us examine the models in more detail. For both Model 3 and Model 4, we can write the within-case model as:

$$Y_{ij} = \beta_{0j} + \beta_{1j} Tx_{ij} + \beta_{2j} Tx_{ij} \times (Time_{ij} - k_j - D) + e_{ij}, \tag{4.1}$$

where Y_{ij} is the score on the outcome variable Y at measurement occasion i for case j, Tx_{ij} is dummy coded with a value of 0 for baseline observations and a value of 1 for the treatment phase observations, k_j is the last time-point before case j enters the treatment phase, and D is a centering constant that defines the focal time for indexing the treatment effect. The raw score treatment effect for case j is indexed by β_{1j}, which is the distance between the treatment phase trend line and the baseline mean at the time where $(Time_{ij} - k_j - D) = 0$. In Figure **??**, we portray the raw score treatment effect for Martin at a time of 5 observations into treatment. If we set $D < 5$, the focal treatment effect would be smaller; if we set $D > 5$, the focal treatment effect would be larger. The other coefficients of the within-case model are β_{0j}, which is the mean level of the outcome during baseline for case j, and β_{2j}, which is the slope of the treatment phase trend line for case j. The error (e_{ij}) is time-specific and case-specific and assumed normally distributed and first-order autoregressive with variance σ_e^2.

Because the effect varies over time in the treatment phase and we estimate it at a specific focal time (i.e., when $(Time_{ij} - k_j - D) = 0$), it is important for us to consider what focal time we should select. For example, is it better to estimate the effect at the beginning of the treatment phase or between 3-10 observations into the treatment phase? Researchers should carefully consider their focal time selection by examining patterns in the typical treatment phase length in each SCD and group design study. For example, we chose to estimate the effect for the SCD studies 10 observations into the treatment phase for two reasons. First, 10 observations into treatment is closely aligned with the group design study time at posttest. Second, SCD studies in this area, like those used for this illustration, tend to have at least 10 treatment observations; a single

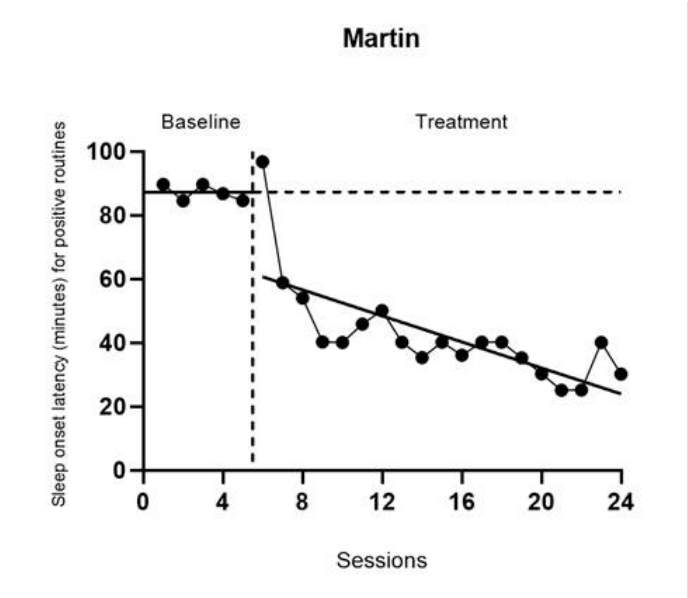


Figure 4.4: Martin's Treatment Effect Five Observations into Treatment (Delemere & Dounavi, 2018)

exception in our studies can be seen for John (**?**), who has only seven treatment observations.

For Model 3, the between-case model specification is:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{4.2}$$

$$\beta_{1j} = \gamma_{10} \tag{4.3}$$

$$\beta_{2j} = \gamma_{20} \tag{4.4}$$

where γ_{00} is the across- case average baseline mean and u_{0j} is a case-specific error, which accounts for variation between cases in their mean baseline levels and is assumes assumed to follow a normal distribution with variance $\sigma^2_{u_0}$. We assume the same average treatment effect (γ_{10}) and the average slope for the treatment phase (γ_{20}) across cases. Thus, there are no error terms in the latter two equations. We show the design-comparable effect size in Equation (**??**), defined as the average raw score treatment effect (γ_{10}) divided by a standard deviation (SD) comparable to the SD used in group design studies.

$$\delta = \frac{\gamma_{10}}{\sqrt{\sigma^2_{u_0} + \sigma^2_e}}, \tag{4.5}$$

Model 4 is specified in a very similar way as Model 3, apart from adding error terms (u_{1j} and u_{2j}) to the final two equations to account for between-case

variation in the treatment effect and treatment phase slopes. More specifically:

$$Y_{ij} = \beta_{0j}+\beta_{1j}Tx_{ij}+\beta_{2j}Tx_{ij}\times(Time_{ij}-k_j-D)+e_{ij}, <!--MC: I use Tx instead of Tx and use e_{ij} instead of r_{ij} in Equat \tag{4.6}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{4.7}$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \tag{4.8}$$

$$\beta_{2j} = \gamma_{20} + u_{2j} \tag{4.9}$$

Again, the across-case average baseline mean is γ_{00} and the average raw score treatment effect at the focal time is γ_{10}. The case-specific errors (u_{0j}, u_{1j}, u_{2j}), which account for between-case differences in baseline level and response to treatment, are assumed multivariate normal with covariance $\Sigma_u = \begin{bmatrix} \sigma^2_{u_0} & & \\ \sigma_{u_1u_0} & \sigma^2_{u_1} & \\ \sigma_{u_2u_0} & \sigma_{u_2u_1} & \sigma^2_{u_2} \end{bmatrix}$. Note that one could opt for a model that is intermediate between Model 3 and Model 4, by including only u_{1j} or u_{2j}, and thereby allowing random variation in either (but not both) the level or slope changes. We proceed with random variability in each, because we think the treatment could lead to both different shifts in level and different treatment phase slopes for different participants. However, the inclusion of three random effects with an unstructured covariance matrix can be challenging to estimate with a limited number of cases. A simpler model with fewer random effects may be preferable if estimation difficulties arise.

Regardless of whether we allow both β_{1j} and β_{2j} to vary randomly as shown in Equations (**??**) and (**??**) or allow only one of these effects to vary randomly, we define the design-comparable effect size exactly as in Equation (**??**). This is because the effect size is scaled by the SD of the outcome in the absence of intervention (i.e., during baseline); it is not impacted by assumptions about the between-case variation during the treatment phase.

In the following sections, we will illustrate the estimation of design-comparable effect sizes for the SCD studies using Model 4 based on a priori considerations and the differences noted between cases in their treatment phase slopes. After obtaining the design-comparable effect sizes using Model 4, we repeat the process using Model 3 for several reasons. First, contrasting the models allows us an additional way to examine the empirical support for our chosen model. For instance, visual analyses of the model-implied individual trajectories for Model 4 might show a better fit than those for Model 3. Second, the contrast allows us to examine the sensitivity of the effect size estimates to our selected model. For instance, whether we assume the effect size varies across cases could have little to no effect on the design-comparable effect size estimate. Last, the contrast between Model 3 and Model 4 design-comparable effect sizes allows us to illustrate a method of selecting between models in circumstances where a priori information is not sufficient to select a model.

4.3 Estimating the Design-Comparable Effect Size for the Single-Case Studies

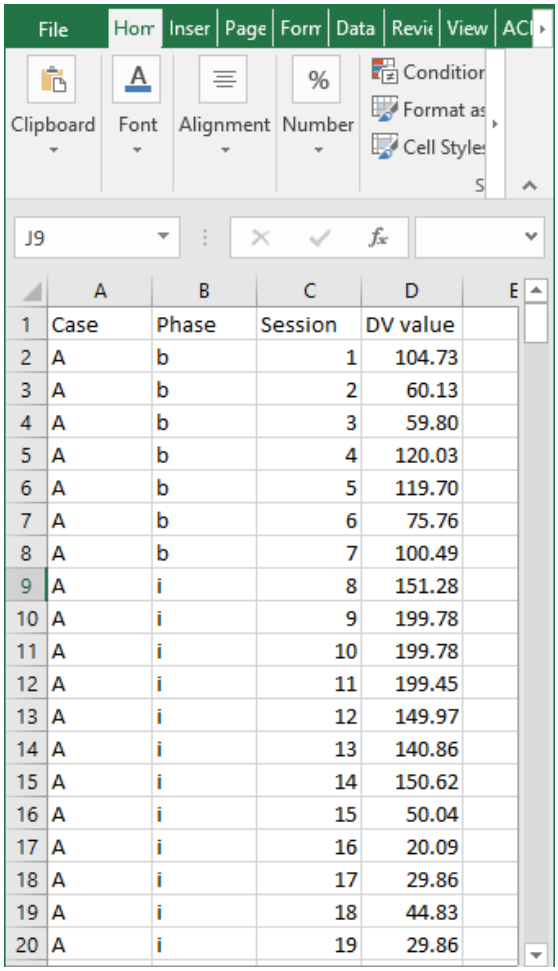
4.3.1 Example 1: Multiple Baselise Study by ?

We can estimate design-comparable effect sizes for any model in Figure **??** using a web-based calculator for design-comparable effect sizes (**?**). The *scdhlm* app is available at https://jepusto.shinyapps.io/scdhlm/. To use this app, researchers must store the data in an Excel file (.xlsx), comma delimited file (.csv), or text file (.txt). In addition, the data must include columns for the *case identifier*, +phase identifier_, *session number*, and *outcome*. Although not required, we suggest that researchers arrange their data columns by order of variable appearance in the *scdhlm* app, putting the *case identifier* in the first column, *phase identifier* in the second column and so on. We present this layout in Figure **??**, representing data extracted for the study by **?**. In this illustration, for *phase identifier*, we use *b* to indicate baseline observations and *i* to indicate intervention observations. However, researchers can use any other labeling scheme that clearly distinguishes between baseline and intervention conditions (e.g., 0 and 1, respectively). Unlike the *phase identifier* variable, *session number* and *outcome* variables must contain numerical values. We also recommend entering data into the spreadsheet first by case (e.g., enter all the rows of data for the first case before any of the rows of data for the second case), then by *session number*.

After starting the app, we use the *Load* tab to load the data file, as illustrated in Figure **??**. The data file could be a .txt or .csv file that includes one dataset or could be an Excel (.xlsx) file that has either one spreadsheet (e.g., a data set for one study), or multiple spreadsheets (one spreadsheet for each of several studies). If using a .xlsx file with multiple spreadsheets, we can select the spreadsheet containing the data for the study of interest from the *Load* tab. Then, we use the drop-down menus on the right of the screen to indicate the study design (*Treatment Reversal* versus *Multiple Baseline/Multiple Probe across participants*) and define which variables in the data set correspond to the case identifier, *phase identifier*, session, and *outcome* (see Figure **??**).

After we load our data, we use the *Inspect* tab to ensure that the raw data imported correctly and mapped to their corresponding variable names (Figure **??**). In addition, we can use the Inspect tab to view a graph of the data (Figure **??**). At this point, we recommend that researchers compare these data with the graphed data from the original studies as an additional measure that ensures the study data uploaded to the app correctly (according to the selections on the *Load* tab). Later, these graphed data can also be checked again for consistency with the tentatively selected model for estimating the design-comparable effect size.

Upon completion of data inspection, we next specify the model for the design-comparable effect size using the *Model* tab. Figure **??** shows the specification for Model 4, the model that assumes no baseline trend, a trend in the treatment



	A	B	C	D
1	Case	Phase	Session	DV value
2	A	b	1	104.73
3	A	b	2	60.13
4	A	b	3	59.80
5	A	b	4	120.03
6	A	b	5	119.70
7	A	b	6	75.76
8	A	b	7	100.49
9	A	i	8	151.28
10	A	i	9	199.78
11	A	i	10	199.78
12	A	i	11	199.45
13	A	i	12	149.97
14	A	i	13	140.86
15	A	i	14	150.62
16	A	i	15	50.04
17	A	i	16	20.09
18	A	i	17	29.86
19	A	i	18	44.83
20	A	i	19	29.86

Figure 4.5: Snapshot of Spreadsheet Containing Extracted Gunning & Espie (2003) Data

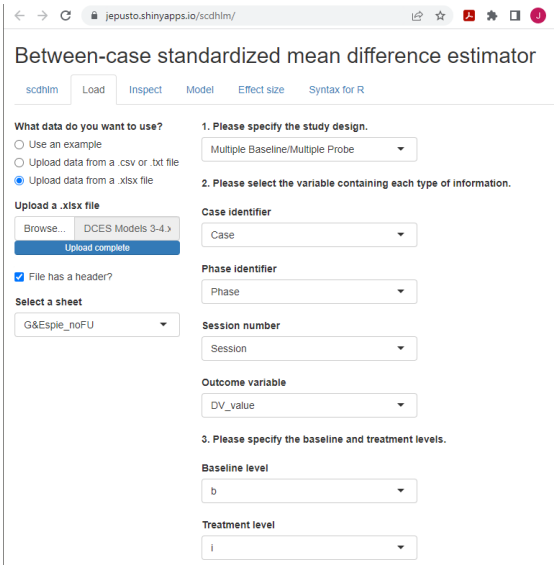
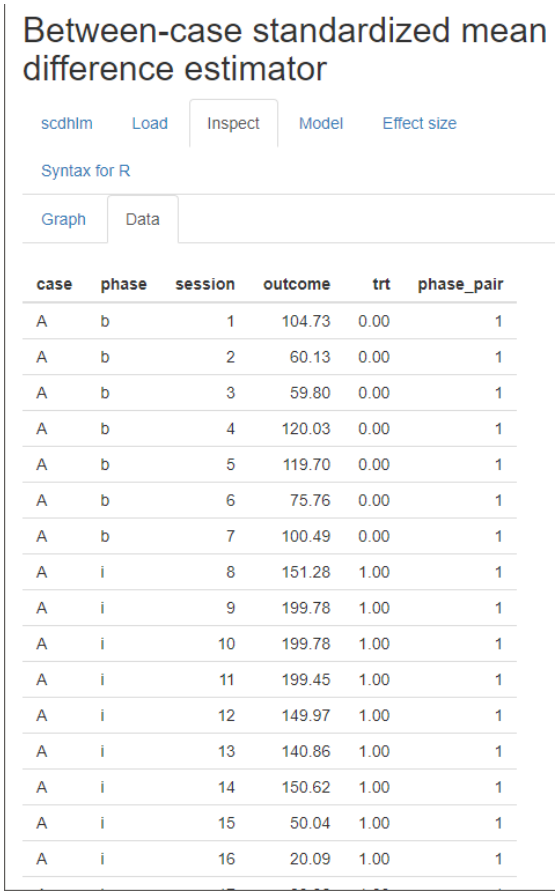


Figure 4.6: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Load Tab for Gunning & Espie (2003)

phase, and an effect that varies across cases. Specification begins with the *Baseline phase* section, where we select *level* under *Type of time trend* because we assume that there are no time trends in the baseline phases. We then choose to include *level* as a fixed effect, enabling the model to estimate the average baseline level. We also include *level* as a random effect so that the baseline level can vary from case to case.

Next, we specify the model for the *Treatment phase*. In *Type of time trend*, we choose the option *change in linear trend* because Models 3 and 4 allow for possible time trends in only the treatment phases, implying that the trend changes across phases. To specify Model 4, we include both *change in level* and *change in linear trend* as fixed effects, so that the average trajectory across cases can reflect both an immediate change in level plus a change in the trend. We also select the option to include *change* in level as a random effect to allow the shift in level (i.e., treatment effect) to vary across cases. In addition, we opt to allow the *change in linear trend* to vary from case to case by including it as a random effect. At this point, we have specified the model for the design-comparable effect size that matches Model 4. Note the app allows us to make different potential assumptions about the correlation structure of the session-level errors. Shown are the default options of autoregressive and constant variance across phases. These defaults match the model presented in Equations (**??**) and (**??**) and are used because they seem appropriate for this data set. Also note that at the bottom of the screen (see Figure **??**), the *scdhlm* app provides a graph of the data with trend lines that are based on the specified model (Figure **??**). We



case	phase	session	outcome	trt	phase_pair
A	b	1	104.73	0.00	1
A	b	2	60.13	0.00	1
A	b	3	59.80	0.00	1
A	b	4	120.03	0.00	1
A	b	5	119.70	0.00	1
A	b	6	75.76	0.00	1
A	b	7	100.49	0.00	1
A	i	8	151.28	1.00	1
A	i	9	199.78	1.00	1
A	i	10	199.78	1.00	1
A	i	11	199.45	1.00	1
A	i	12	149.97	1.00	1
A	i	13	140.86	1.00	1
A	i	14	150.62	1.00	1
A	i	15	50.04	1.00	1
A	i	16	20.09	1.00	1

Figure 4.7: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Data Tab within the Inspect Tab for Gunning & Espie (2003)

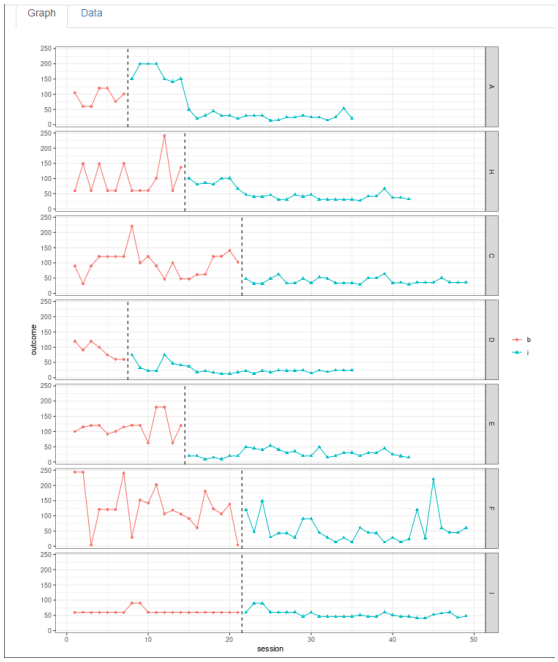


Figure 4.8: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Graph Display within the Inspect Tab for Gunning & Espie (2003)

recommend that researchers inspect this graph to ensure that the trend lines fit the data reasonably well. If they do not, it raises questions about the model choice.

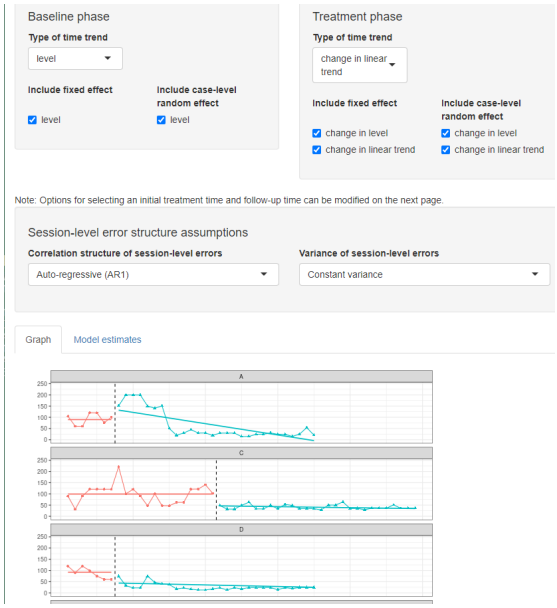


Figure 4.9: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Model 4 Specification for Gunning & Espie (2003)

For this data set, the a priori identified model provides trajectories that fit the data reasonably well, and thus we proceed to the *Effect size* tab (Figure **??**). There are two sliders in the *Hypothetical experimental parameters* section above the effect size estimates output: *Initial treatment time* and *Follow-up time*. The numbers on each slider refer to the sessions or time points in the data series. Although the app populates times automatically, researchers can manipulate them manually to obtain an effect size estimate at a desired point in time. For models with trends in the baseline phase, it matters where we set the initial time. However, for models without baseline trends, only the distance between the two slider values matters.

We imagine a relatively typical SCD study with a baseline of 5; therefore, we use the sliding scale to set the *Initial treatment time* for the **?** study to 5 to represent the last session before initiation of the treatment. Next, we estimate the treatment effect 10 observations into treatment and move the *Follow-up time* slider to the 15th observation. Given these Model 4 specifications for **?**, we find that the estimated between-case standardized mean difference 10 observations into treatment is -1.12 with a standard error (SE) of 0.34 and 95% confidence interval (CI) [-1.80, -0.45].

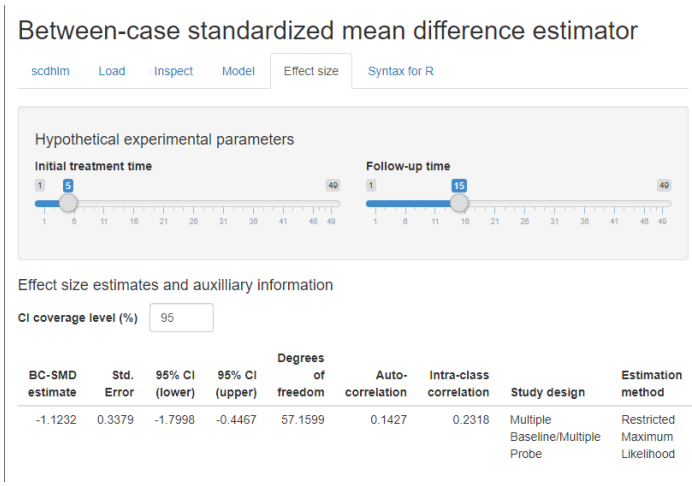


Figure 4.10: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Effect size Tab Showing Model 4 Estimate for Gunning & Espie (2003)

Now, if we lacked confidence in the a priori decision to select Model 4 and wanted to explore the Model 3 fit, we can re-run the study data without returning to the *Load* tab. Instead, we go directly to the *Model* tab to change our specification. To obtain Model 3, we remove (uncheck) *change in level* and *change in linear trend* as random effects, constraining the change in level and change in trend from baseline to treatment to be the same for each case. For Model 3, the estimated effect size is -1.10 with an SE of 0.17 and 95% CI [-1.43, -0.76]. In this case, for the **?** study, the effect size estimates for both Models are similar. However, with the more restrictive assumptions of Model 3, the CI is narrower. If we had used a model that allowed a change in level to vary randomly, but required a fixed slope change, the estimated effect size would be -1.07 with an SE of 0.28 and 95% CI [-1.63, -0.51].

The *Effect size* tab (Figure **??**) reports additional information including estimates of other model quantities, information about the model specification, and assumptions used in calculating the design-comparable effect size. The reported degrees of freedom are used by the app in making a small-sample correction to the effect size estimate, analogous to the Hedges' g correction used with group designs (**?**). Larger estimated degrees of freedom mean that the denominator of the design-comparable effect size is more precisely estimated, and that the small-sample correction is less consequential. Conversely, small degrees of freedom indicates that the denominator of the effect size is imprecisely estimated, making the small-sample correction more consequential. The reported autocorrelation is the estimate of the correlation between errors at the first level of the model for the same case that differ by one time-point (or session), based on a first-order autoregressive model. The reported intra-class correlation is an

estimate of the between-case variance of the outcome as a proportion of the total variation in the outcome (including both between-case and within-case variance) as of the selected *Follow-up time*. Larger values of the intra-class correlation indicate that more of the variation in the outcome is between participants. The remaining information in the output (*Study design*, *Estimation method*, *Baseline specification*, *Treatment specification*, *Initial treatment time*, *Follow-up time*) describe the model specification and assumptions used in the effect size calculations. The app includes it to allow for reproducibility of the calculations.

4.3.2 Example 2: Multiple Baseline Study by ?

After obtaining a design-comparable effect size for the first SCD (**?**), we repeat these steps for all other included SCD studies. In our second included SCD study (**?**), the researchers analyzed two interventions. While we could estimate a separate design-comparable effect size for each, for the purposes of illustrating the computational steps in this chapter, we have pooled the data from the six cases to estimate one design-comparable effect size. To obtain a design-comparable effect size for the **?** study, we follow the same sequence of steps: 1. Load the data. 2. Inspect the data in both tabular and graphic form. 3. Specify our selected model for the data (see Figure **??** for Model 4). 4. Estimate the design-comparable effect size.

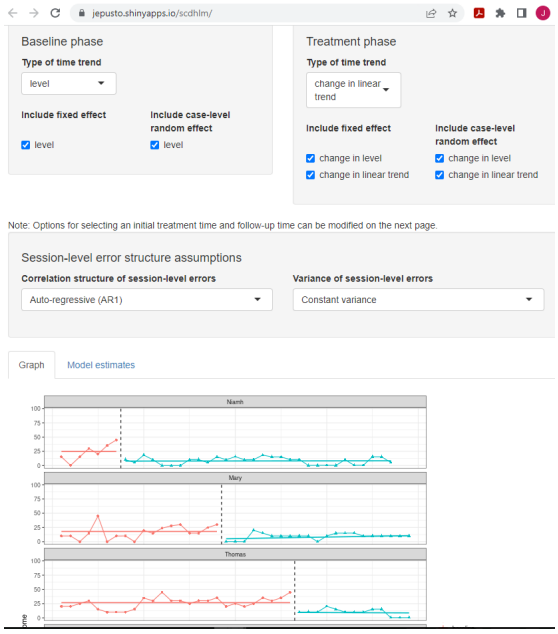


Figure 4.11: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Model 4 Specification for Delemere and Dounavi (2018)

Proceeding with Model 4, we use the Effect size tab of the *scdhlm* app to obtain a design-comparable effect size for the **?** study. As with the previous example, the effect size will depend on the time into intervention at which we estimate the effect size the focal time (indicated by the *scdhlm* app variable named *Follow-up* time). When estimating design-comparable effect sizes for SCD studies, researchers should hold the *Follow-up time* constant across studies. Therefore, we use the *Initial treatment time* and *Follow-up time* sliders at the top of the screen to obtain an estimate of effect 10 observations into treatment (i.e., we set the *Initial* and *Follow-up time* sliders to five and 15, respectively). The resulting design-comparable effect size for the **?** study is -0.70 with an SE of 0.33 and 95% CI [-1.49, 0.09].

Again, if we want to compare these results to those estimated using Model 3, we can go back to the *Model* tab to change our specification. To obtain Model 3, we remove (uncheck) *change in level* and *change in linear trend* as random effects to obtain the same change in level and change in trend from baseline to treatment across cases. We keep all other modeling options the same (e.g., *Initial treatment time* and *Follow-up time*). Figure **??** shows the Model 3 specification for the **?** study. The estimated design-comparable effect size is -0.91 with an SE of 0.28 and 95% CI [-1.52, -0.29]. Comparing the fit of the trend lines obtained from Model 4 (Figure **??**) to those from Model 3 (Figure **??**), our originally specified Model 4 appears to provide a better fit for the primary study data. Because of this and because Model 4 is more consistent with both our a priori expectations and the data from our other SCD study (i.e., **?**), it seems reasonable to proceed with the effect estimate from Model 4 (i.e., -0.70).

4.4 Estimating the Design-Comparable Effect Size for the Group Studies

After estimating the design-comparable effect size for the included SCD studies, the next step is to estimate a design-comparable effect size for the included group design study. Details on estimating standardized mean difference effect sizes from group studies are readily available from a variety of sources, including Chapter 12 of *The Handbook on Research Synthesis and Meta-Analysis* (**?**), books (e.g., **?**) and journal articles (e.g., **??**). Therefore, we do not demonstrate the step-by-step effect size estimation methods for group design studies in this methods guide. Instead, we summarize the results below.

The included group design study, **?**, reported a randomized trial comparing the efficacy of sleep interventions versus a wait-list control condition for families with children with severe learning disabilities. The researchers assigned participating families to a conventional face-to-face intervention ($n = 20$), a brief treatment delivered as a booklet ($n = 22$), or a wait-list control condition ($n = 24$). At posttest, the researchers compared participants' composite sleep disturbance score (derived from parent reports) across groups, with higher

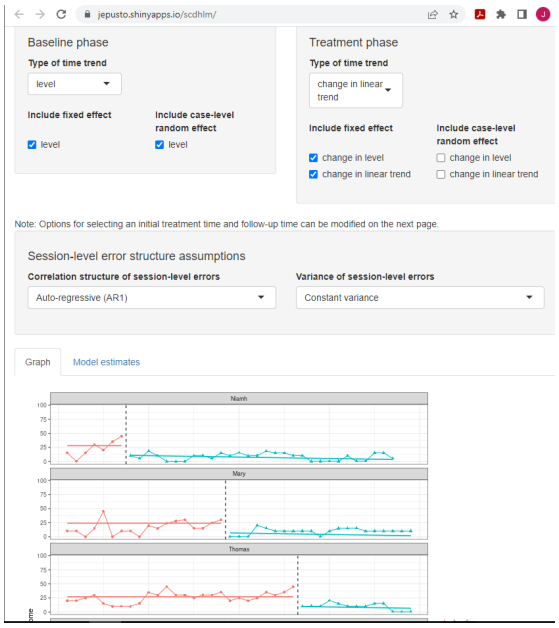


Figure 4.12: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Model 3 Specification for Delemere and Dounavi (2018)

scores corresponding to more severe sleep disturbance. For purposes of effect size calculations, we pooled the results from the face-to-face and booklet intervention conditions. Based on the post-treatment means and standard deviations (**?**, Table 1), the standardized mean difference in sleep onset at the end of intervention was -1.52, with an SE of 0.29. This effect size estimate is based on Hedges' g, which corrects for small sample size bias.

4.5 Analyzing the Effect Sizes

As a final step, we synthesize the effect sizes across both SCD and group design studies. A variety of tools and approaches are available to researchers depending on the goals of their synthesis. For example, researchers can: (a) create graphical displays that show the effect size for each study along with its CI, (b) report an overall average effect size and CI by averaging effect sizes across studies and, (c) examine the extent of effect size variation across studies, (d) explore potential moderators of the effects, and (e) examine the effect sizes for evidence of publication bias. Because the use of design-comparable effect size for the SCD studies produces effect estimates that are like the commonly used standardized mean difference effect sizes from group studies, researchers can accomplish these goals using the methods already established for group design studies. Details on these methods are readily available elsewhere (e.g., **??**). We illustrate the aver-

Table 4.1: Fixed Effect Meta-Analysis Calculations for Example Sleep Intervention Studies

Study	Effect Size Estimate (A)	Standard Error (B)	Inverse-variance We
Model 4			
Gunning & Espie (2003)	-1.12	0.34	8.65 (29.
Delemere & Dounavi (2018)	-0.70	0.33	9.18 (30.
Montgomery et al. (2004)	-1.52	0.29	11.89 (40
Fixed effect meta-analysis	-1.15	0.18	29.72 (10
Model 3			
Gunning & Espie (2003)	-1.10	0.17	34.60 (58
Delemere & Dounavi (2018)	-0.91	0.28	12.76 (21
Montgomery et al. (2004)	-1.52	0.29	11.89 (20
Fixed effect meta-analysis	-1.14	0.13	59.25 (10

aging of the effect sizes from our studies here using a fixed effect meta-analysis, consistent with the approach used in What Works Clearinghouse intervention reports (**?**).

Table **??** reports the effect size estimates, SEs, and fixed effect meta-analysis calculations for the example studies. The top panel uses the design-comparable effect size results for SCD studies based on Model 4. As a sensitivity analysis, the bottom panel presents the results based on Model 3. Note that the effect size estimate from the group design study is the same in both panels. In fixed effect meta-analysis, the overall average effect size estimate is a weighted average of the effect size estimates from the individual studies, with weights proportional to the inverse of the sampling variance (squared SE) of each effect size estimate. Further, the SE of the overall effect size is the square root of the inverse of the total weight.

Column C of Table **??** reports the inverse-variance weight assigned to each of the studies, with the percentage of the total weight listed in parentheses. In the top panel based on Model 4, the effect size estimate from the group design study receives 40% of the total weight, while the effect size estimates from the SCD studies receive 29% and 31% of the total weight, respectively. The total inverse variance weight is 29.72. The overall average effect size estimate based on Model 4 is -1.15 with an SE of 0.18[1] and an approximate 95% CI [-1.51, -0.79]. The Q-test for heterogeneity is not significant, $Q(2) = 3.50$, $p = .174$, indicating that the included effect size estimates are consistent with the possibility that all studies were estimating a common effect size parameter. Interestingly, in this example, the effect size estimate from the group design is larger in magnitude than the effect size estimates from the SCD studies.

The bottom panel of Table **??** reports the same calculations but using the design-

[1]The SE of the overall effect size is the square root of the inverse of the total weight.

comparable effect size estimates based on Model 3 for the two SCD studies. The most notable difference is that Model 3 more precisely estimates the design-comparable effect size for the **?**, resulting in more weight (58%) assigned in the fixed effect meta-analysis. For Model 3, the overall average effect size is nearly identical to that obtained for Model 4 (-1.14), but the substantially smaller Model 3 design-comparable effect size for the **?** study results in an SE of 0.13 and 95% CI [-1.40, -0.89].

In fixed effect meta-analysis, the overall average effect size estimate is a summary of the effect size estimates across the included studies, which are treated as fixed. Therefore, the SE and CI in fixed effect meta-analysis take into account the uncertainty in the process of estimating the effect sizes in each of the individual studies, but they do not account for uncertainty in the process of identifying studies for inclusion in the meta-analysis (**??**). Consequently, they do not provide a basis for generalization beyond the included studies. When conducting syntheses of larger bodies of literature—and especially of studies with heterogeneous populations, design features, or dependent effect sizes—researchers will often prefer to use random effects models (**?**) or their further extensions (**??**).

Chapter 5

Illustration of Design-Comparable Effect Sizes When Assuming Trends in Baseline and Different Trends in Treatment

Chapter 5 illustrates the computation of design-comparable effect sizes in contexts where one assumes time trends in baseline and that the treatment will lead to changes in both the level and slope of the time trends. We demonstrate the calculations using data from a multiple probe study, a multiple baseline study, and a group study. For the single-case studies, we provide step-by-step instructions for selecting design-comparable effect sizes and estimating them using the scdhlm *app. We also discuss estimating the effect size for the group study and synthesizing the effect sizes across the group and single-case studies.*

In this chapter, we describe how researchers can compute design-comparable effect sizes and synthesize results using models that assume baseline time trends and a treatment effect that has an immediate impact on the level of the outcome, along with an effect on the trend—assumptions that correspond to Models 5 and 6 from Figure **??**. For illustrative purposes, we calculate effect sizes and synthesize findings from three studies of interventions designed to improve writing skills for students with learning disabilities, including a multiple probe study, a multiple baseline study, and a group design study. The multiple probe study by

? investigated intervention effects on the writing behavior of four adolescents with writing difficulties. The multiple baseline study by **?** investigated the effects of a writing intervention for four postsecondary students with intellectual and developmental disabilities. Finally, the group study by **?** investigated the effects of a writing intervention for 61 fourth and fifth grade students designated as struggling readers (i.e., at least one grade level behind their peers).

For this hypothetical synthesis, we choose to use a design-comparable effect size to aggregate effects across these study designs. We organize the procedures into four stages: (1) selecting a design-comparable effect size for the SCD studies, (2) estimating the design-comparable effect size using the *scdhlm* application, a web-based calculator for design-comparable effect sizes (**?**), (3) estimating the effect size for the group study, and (4) synthesizing the effect sizes across the SCD and group studies. Because well-developed effect size estimation methods for group studies are illustrated elsewhere (**??**), we concentrate primarily on the first two steps.

5.1 Selecting a Design-Comparable Effect Size for the Single-Case Studies

We use the decision rules in Figure **??** to select an appropriate design-comparable effect size for the SCD studies. To do this, we consider the characteristics of the population, the context, and the outcome of interest. We make predictions about the type of trends we expect to see in the baseline and treatment phases, and we reflect on the rationale for our expectations. Based on our prior research and experience with this population, we hypothesize that the outcome variable of student writing quality may have baseline phase trends. For example, some students may show slight positive trends because of practice effects, others may show a deteriorating trend in absence of writing feedback, or students' baseline observations may portray little to no trend at all in absence of intervention. Despite expected variability in the baseline phase, we do expect a noticeable positive shift in the level of responding early in the treatment phase, followed by a gradual increase in writing quality over time. Because we are assuming the presence of trends in both baseline and treatment phases, we tentatively consider Model 5 or Model 6 (see Figure **??**) to estimate the included study effect sizes. We make the choice between Model 5 and Model 6 based on our logic model and a priori expectation that the treatment effect will not vary across cases (Model 5) or that the treatment effect could vary across cases (Model 6). Based on the different writing abilities of the participants, as well as their unique learning histories and abilities, we anticipate some across-case variation in effects. Thus, we tentatively choose Model 6 as most consistent with our understanding of this context.

Next, we visually analyze the graphs from the SCD studies to see if our initial assumptions are reasonable considering the data. Figure **??** shows the data we

extracted from **?**, and Figure **??** shows the data we extracted from **?**. In viewing the graphs, we examine the extent to which the data are reasonably consistent with the homogeneity and normality assumptions underlying the models for design-comparable effect sizes. Note that in Figures **??** and **??**, we see similar between-case variation and similar variation across phases within a case and can conclude that our homogeneity assumptions are reasonable. In addition, we do not detect the presence of outlying values or clear departures from normality. Consequently, our initial decision to use design-comparable effect sizes appears to be a reasonable choice.

During visual analysis of the primary study graphs, we now need to determine the appropriateness of choosing a model with baseline trends. Consistent with our expectations, the typical baseline pattern across cases in Figures **??** and **??** does exhibit a trend. For example, Richard and Danyell (Figure 5.1) and all cases in Figure **??** (except Denny) appear to have a decline in writing skills performance in baseline, as indicated by a downward trend in the data. However, the baseline trends for Nathan and Kendrick (Figure **??**) are less clear or appear to be increasing (e.g., Denny in Figure **??**). Because the typical pattern across the two SCD studies is consistent with our trend expectations, we proceed with an assumption of baseline trends.

Next, we conduct a similar visual analysis of the treatment phases. For most of the cases, we see an immediate shift in performance along with a change in the trend. Richard, Danyell, and Kendrick (Figure **??**) show an improvement in writing skills that generally increases over time in treatment (i.e., positive trend lines). Brett, Steve, and Denny (Figure **??**) show a similar pattern of increasing improvement in writing skills. Because the design-comparable effect size assumes a common model across cases and estimates an average effect across the cases, we find it best to select a model that is consistent with the typical and expected pattern. Thus, it appears reasonable to proceed with a model that assumes a trend in baseline and a different trend in the treatment phases (i.e., Model 5 or Model 6 from Figure **??**).

5.2 Details of the Models for Design-Comparable Effect Sizes

To differentiate between Model 5 and Model 6, we provide a formal specification of each. For both Model 5 and Model 6, we write the within-case (level-1) model as:

$$Y_{ij} = \beta_{0j} + \beta_{1j} Tx_{ij} + \beta_{2j}(Time_{ij} - B) + \beta_{3j} Tx_{ij} \times ((Time_{ij} - k_j) - (B - A)) + e_{ij}, \tag{5.1}$$

where Y_{ij} is the score on the outcome variable Y at measurement occasion i for case j, Tx_{ij} is dummy coded with a value of 0 for baseline observations and a value of 1 for the treatment observations, $Time_{ij}$ is the time-point for measurement occasion i for case j, B is a centering constant that defines the

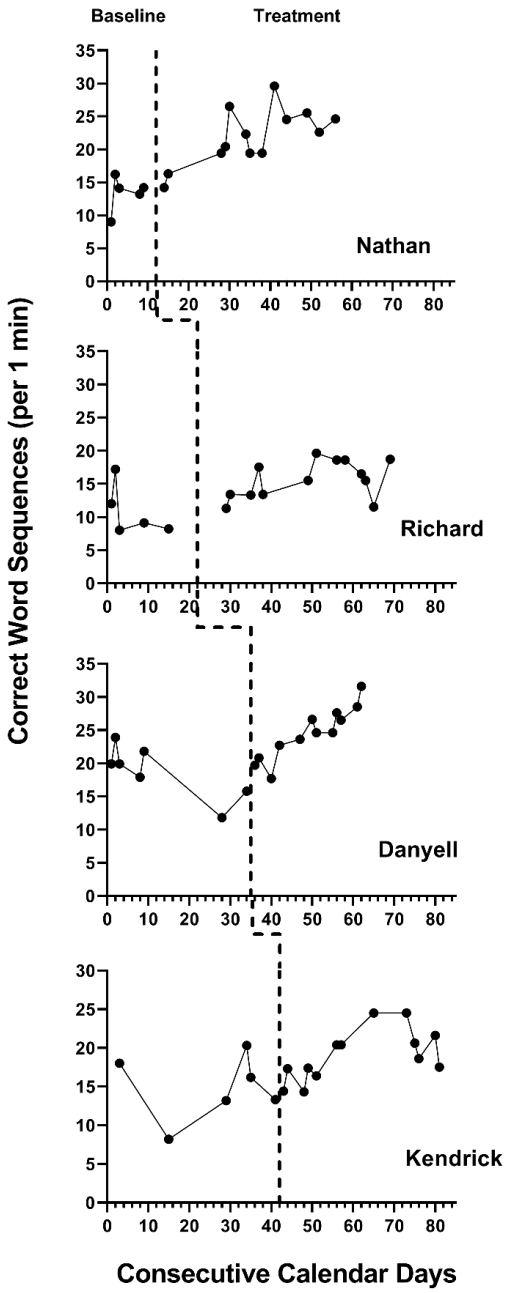


Figure 5.1: Multiple Probe Data Extracted from Datchuk (2016)

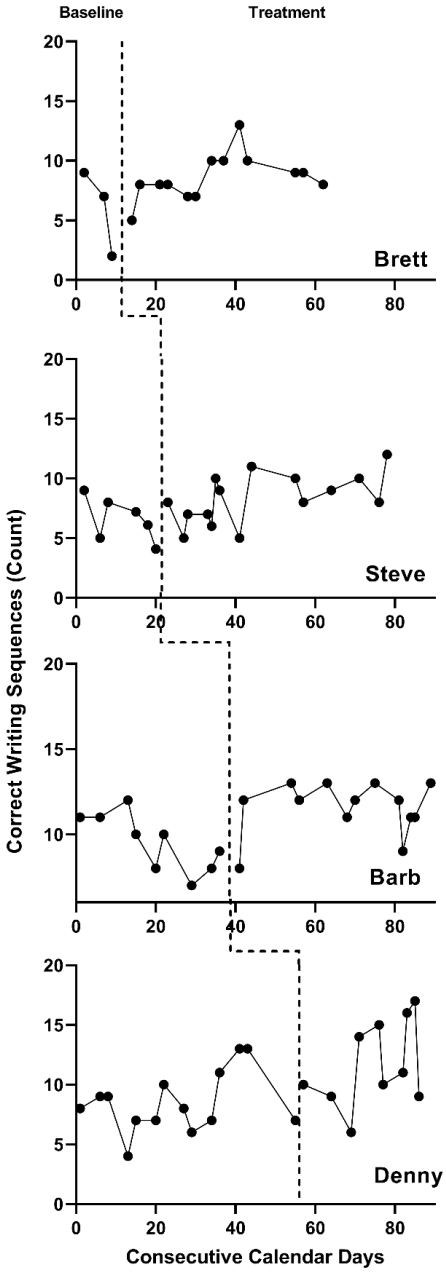


Figure 5.2: Multiple Baseline Data Extracted from Rodgers et al. (2021)

focal follow up time for defining the effect size, k_j is the last time-point before case j begins their treatment phase, and $B - A$ corresponds to the treatment duration (i.e., the number of time points the case has been in intervention at the focal time). β_{1j} indexes the raw score treatment effect for case j, which is the distance between the treatment phase trend line and the extended baseline phase trend line at the focal follow-up time. In addition, this model includes β_{0j}, which corresponds to the expected baseline value for case j at the focal follow up time B, β_{2j} is the slope of the baseline phase, and β_{3j} is the change in slope that occurs with intervention (i.e., the difference between the treatment and baseline phase slopes for case j). Finally, the error (e_{ij}) is time- and case-specific, assumed to be normally distributed, and first-order autoregressive with variance σ_e^2.

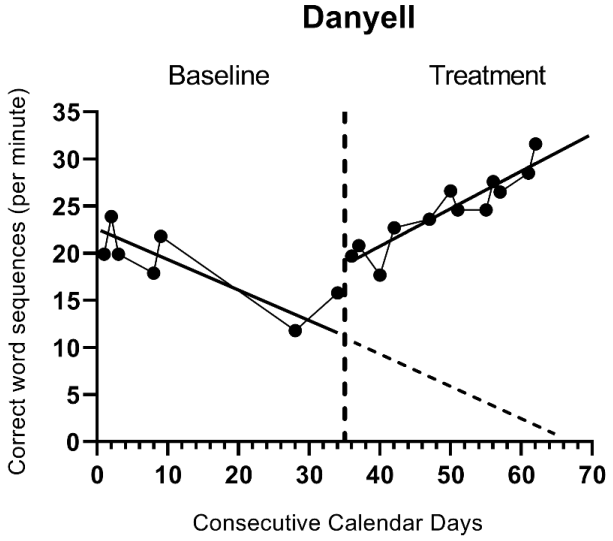


Figure 5.3: Illustration of Treatment Effect 5 Observations into Treatment for Danyell (Datchuk, 2016)

As shown in Figure **??**, the size of the raw score treatment effect depends on its measurement time (i.e., the raw score effect size depends on the focal time selected). It becomes important for researchers to determine the time at which the effect should be estimated because: (a) the raw score effect size varies over time in the treatment phase, and (b) it is estimated specifically after a specific treatment duration [i.e., when $(Time_{ij} - k_j) - (B - A) = 0$]. To choose a focal time, we find it helpful to review the SCD research literature and examine the treatment phase lengths of the extant studies. It is also helpful to consider the duration of the treatment in the group design studies, because such studies typically index the treatment effect immediately after completing the intervention.

Generally, We we do not want to avoid extrapolate extrapolating much beyond the length of the actual treatment for most cases.

For this illustration we choose to estimate the effect 12 observations into treatment because the group design study (**?**) had 12 intervention sessions. Further, the SCD studies in this area—particularly the two we are using in this chapter—tend to have at least 12 treatment observations (with the exception of Denny in Figure **??**, who has 10 treatment observations).

For Model 5, we write the between-case model as:

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{5.2}$$

$$\beta_{1j} = \gamma_{10} \tag{5.3}$$

$$\beta_{2j} = \gamma_{20} + u_{2j} \tag{5.4}$$

$$\beta_{3j} = \gamma_{30} \tag{5.5}$$

where γ_{00} is the across-case average value of the outcome at time zero (0) based on a linear projection of the average baseline trajectory to the focal follow-up time B and u_{0j} is a case-specific error, which accounts for variation between cases in their expected baseline levels at time point B. Thus, $\sigma^2_{u_0}$ corresponds to the projected between-case variance at the focal time if the cases remained in baseline. The across-case average baseline slope is γ_{20}, and the corresponding error term (u_{2j}) allows this slope to vary across cases. We assume that the errors in these equations (u_{0j} and u_{2j})) are distributed multivariate normal with covariance Σ_u. The across-case average raw score treatment effect at the focal time is γ_{10} and the across-case average change in slope with intervention is γ_{30}. We assume these effects to be the same for each case in Model 5, so we do not include error terms in Equations (**??**) and (**??**).

We define the design-comparable effect size at the focal time as the raw score treatment effect at the focal time γ_{10} divided by a standard deviation (SD) that is comparable to the SD used to standardize mean differences in the group design studies. Compared to the models discussed in previous chapters, the added complexity of Models 5 and 6 makes researchers' choice of time points more consequential. Specifically, when we assume that the baseline slopes vary, we assume that the between-case variability changes with time. Thus, the design-comparable effect size will depend on both the duration of the treatment and the projected baseline between-case variability at the follow-up time. It may also be helpful to acknowledge the assumed heterogeneity across time in the SCD studies and to consider a sensitivity analysis that would examines the extent to which the design-comparable effect size depends on the between-case variance estimation time point. The assumed heterogeneity poses a lesser threat for the synthesis if there is little variation between the design-comparable effect size and the different choices for when to estimate the between-case variance. lead to relatively little variation in the design-comparable effect size estimate.

With these caveats in mind, and using the model specification given in Equations (**??**)-(**??**), we define the design-comparable effect size as:

$$\delta = \frac{\gamma_{10}}{\sqrt{\sigma^2_{u_0} + \sigma^2_e}}, \tag{5.6}$$

where $\sigma^2_{u_0}$ corresponds to the between-case-variance in expected baseline levels at the target time B. The numerator of the effect size corresponds to the average effect of receiving treatment for duration $B - A$, where both A and B are constants specified by the meta-analyst. Therefore, one can interpret this effect size as describing the effect in a hypothetical study where all cases start intervention after time A and where outcomes are assessed at the focal follow-up time B (see **?**, for more details on this interpretation).

The specification of Model 6 is like Model 5, but we add error terms (u_{1j} and u_{3j}) to the equations to account for between-case variation in the treatment effect and between-case variation in the change in slopes with intervention. More specifically:

$$Y_{ij} = \beta_{0j} + \beta_{1j}Tx_{ij} + \beta_{2j}(Time_{ij} - B) + \beta_{3j}Tx_{ij} \times ((Time_{ij} - k_j) - (B - A)) + e_{ij} \tag{5.7}$$

$$\beta_{0j} = \gamma_{00} + u_{0j} \tag{5.8}$$

$$\beta_{1j} = \gamma_{10} + u_{1j} \tag{5.9}$$

$$\beta_{2j} = \gamma_{20} + u_{2j} \tag{5.10}$$

$$\beta_{3j} = \gamma_{30} + u_{3j} \tag{5.11}$$

The added error terms contribute to variability under the intervention condition. However, because the SD used to standardize the design-comparable effect size is based on the variability in the absence of intervention, the definition of the design-comparable effect size for the Model 6 definition matches that of for Model 5. Like Model 5, the choice of the focal time will impact the design-comparable effect size, not only because it impacts influences the raw score treatment effect γ_{10}, but also because it impacts influences the between-case variance.

For the purposes of this chapter, we tentatively select Model 6 based on our a priori considerations and illustrate the estimation of design-comparable effect sizes for these data. We also provide a contrast of Model 6 results compared to Model 5 specifications. This contrast is useful for several reasons, including that it provides an additional way for us to examine the empirical support for the model we have chosen (e.g., visual analyses of the implied trajectories for Model 6 may fit the raw data better than those for Model 5). In addition, by contrasting the results of both models, we can examine the sensitivity of our effect size estimates to the model chosen (e.g., whether we assume the effect size varies across cases may have little to no effect on the design-comparable effect size). Finally, contrasting model results allows us to illustrate a method

of selecting between models in circumstances where a priori information is not sufficient to inform the choice between specifications.

Before illustrating the estimation of the design-comparable effect size for our SCD studies, we emphasize that there are other variations of these models with slopes in both baseline and treatment phases. For example, if we remove the error term from Equation (**??**) or Equation (**??**), we constrain the model so that all cases have the same baseline slope. Constraining the slopes to be equal across cases makes the between-case variance homogeneous across time, which simplifies the definition of the design-comparable effect size. We could also consider removing the error term from Equation (**??**), which would constrain the change in slopes to be equal across the cases. Yet another possible variation is to remove the main treatment effect (i.e., remove the term $\beta_{1j}Tx_{ij}$ from Equation (**??**) or Equation (**??**)). In the latter example, by removing the main effect for treatment, we constrain the model so there is no immediate shift in level with intervention, but only a change in the slope of the trend line. For this chapter, we illustrate models that correspond to either assuming the effect of the treatment is constant across cases (Model 5) or the that the treatment effect can vary across cases in trajectories' magnitudes of level and slope change (Model 6).

5.3 Estimating the Design-Comparable Effect Size for the Single-Case Studies

5.3.1 Example 1: Multiple Probe Study by ?

We can estimate a design-comparable effect size for Models 5 or 6, as well as the other models suggested in Figure **??**, using a web-based calculator for design-comparable effect sizes (**?**), which is available at https://jepusto.shinyapps.io/scdhlm. In using this application, we recommend that researchers have a data file that aligns with the format expected by the application, as shown in Figure **??** for the SCD study by **?**. To use this app, the data file must be saved as an Excel file (.xlsx), comma delimited file (.csv), or text file (.txt). In addition, the data must include columns for the *case identifier*, *phase identifier*, *session number*, and *outcome*. We illustrate this arrangement in Figure **??** for the **?** study. For the *phase identifier*, we have used "baseline" to indicate baseline observations and "treatment" to indicate intervention observations. Researchers can also elect to use another labeling scheme if it clearly distinguishes between baseline and intervention conditions (e.g., 0 to indicate baseline observations and 1 to indicate intervention observations). Although the data format for the phase identifier variable is flexible (text or numeric), the *scdhlm* app requires use of only numerical values for the *session number* and *outcome* variables. Additionally, we recommend that users arrange their data by first case (i.e., enter all the rows of data for the first case before any of the rows of data for the second case), followed by *session number*.

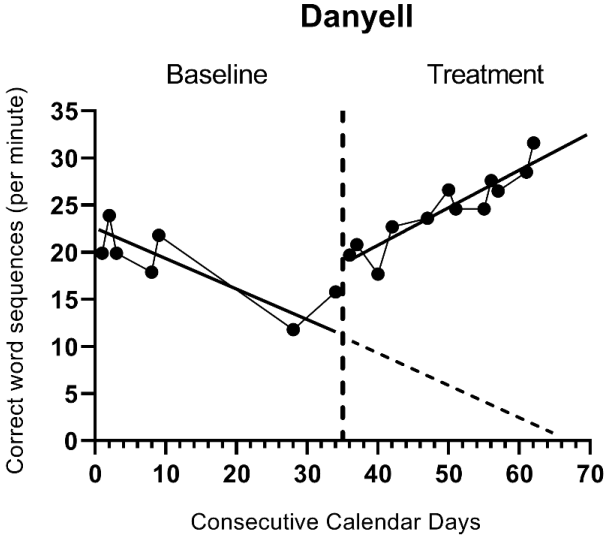


Figure 5.4: Snapshot of Spreadsheet Containing Extracted Datchuk (2016) Data

After starting the app, we use the *Load* tab to load the data file, as illustrated in Figure **??**. The data file could be a .txt or .csv file that includes one dataset or could be an .xlsx file that has either one spreadsheet (e.g., a data set for one study), or multiple spreadsheets (one spreadsheet for each of several studies). If using an .xlsx file with multiple spreadsheets, we can select the spreadsheet containing the data for the study of interest from the *Load* tab. Then, we use the drop-down menus on the right of the screen to indicate the study design (*Treatment Reversal* or *Multiple Baseline/Multiple Probe across participants*) and define which variables in the data set correspond to the case identifier, *phase identifier*, session, and *outcome* (see Figure **??**).

After loading the data, we use the Inspect tab to ensure that the raw data were imported correctly and mapped to their corresponding variable names (Figure **??**). In addition, we can use the *Inspect* tab to view a graph of the data (Figure **??**). At this point, we recommend that researchers compare these data with the graphed data from the original study to ensure the study data uploaded to the app correctly. Later, we can check these graphed data for consistency with the tentatively selected model for the design-comparable effect size.

After data inspection, we go to the Model tab to specify the model for the design-comparable effect size. Figure **??** shows the specification for Model 6 (i.e., the model that assumes trends in baseline, different trends in treatment, and an effect that varies across cases). Starting with specification of the *Baseline phase*, we select *linear trend* under *Type of time trend* because we assume the

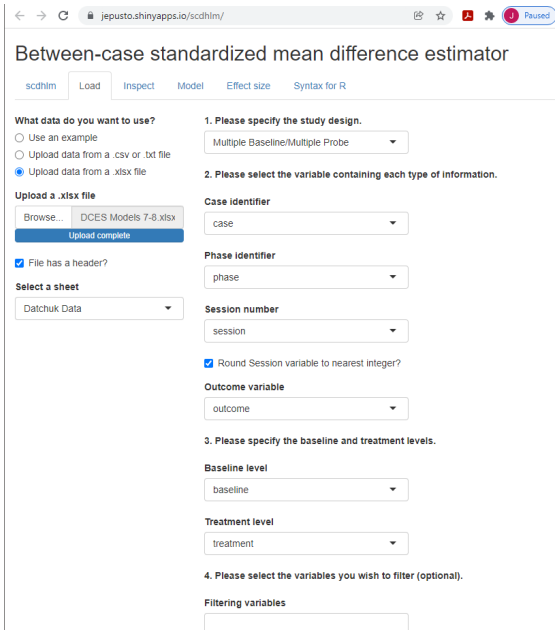
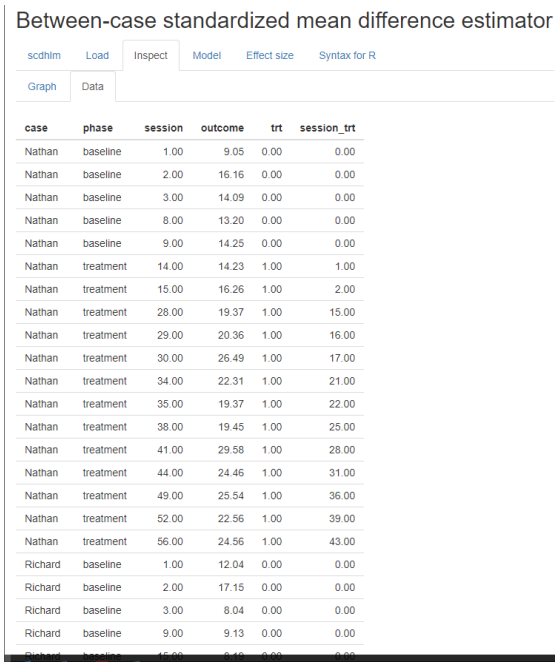


Figure 5.5: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Load Tab for Datchuk (2016)

presence of linear time trends in the baseline phases. We then include *level* and *linear trend* as fixed effects to enable model estimation for an across-case average linear trend line for the baseline phase. To allow the intercepts of the trend lines to vary from case to case, we select the option to include *level* as a random effect. We also include *linear trend* as a random effect to allow the baseline trends to vary across cases.

After *Baseline phase* model specification, we attend to the specification of the *Treatment phase*. For *Type of time trend*, we select *change in linear trend* because we assume trends differ across phases. We then elect to include *change in level* and *change in linear trend* as fixed effects to allow the across-case average trend line for the treatment phase to differ from the across-case average baseline phase trend line, in both level and slope. We also check the boxes to include *change in level* and *change in linear trend* as random effects to allow the changes in trend lines from baseline to treatment phase to vary across cases. Note the app allows us to make different potential assumptions about the correlation structure of the session-level errors. Shown are the default options of autoregressive and constant variance across phases. These defaults match the model presented in Equations (**??**) and (**??**) and were used because they seem appropriate for this data set.

For this data set, we see that our a priori identified model provides trajectories



case	phase	session	outcome	trt	session_trt
Nathan	baseline	1.00	9.05	0.00	0.00
Nathan	baseline	2.00	16.16	0.00	0.00
Nathan	baseline	3.00	14.09	0.00	0.00
Nathan	baseline	8.00	13.20	0.00	0.00
Nathan	baseline	9.00	14.25	0.00	0.00
Nathan	treatment	14.00	14.23	1.00	1.00
Nathan	treatment	15.00	16.26	1.00	2.00
Nathan	treatment	28.00	19.37	1.00	15.00
Nathan	treatment	29.00	20.36	1.00	16.00
Nathan	treatment	30.00	26.49	1.00	17.00
Nathan	treatment	34.00	22.31	1.00	21.00
Nathan	treatment	35.00	19.37	1.00	22.00
Nathan	treatment	38.00	19.45	1.00	25.00
Nathan	treatment	41.00	29.58	1.00	28.00
Nathan	treatment	44.00	24.46	1.00	31.00
Nathan	treatment	49.00	25.54	1.00	36.00
Nathan	treatment	52.00	22.56	1.00	39.00
Nathan	treatment	56.00	24.56	1.00	43.00
Richard	baseline	1.00	12.04	0.00	0.00
Richard	baseline	2.00	17.15	0.00	0.00
Richard	baseline	3.00	8.04	0.00	0.00
Richard	baseline	9.00	9.13	0.00	0.00

Figure 5.6: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Data Tab within the Inspect Tab for Datchuk (2016)

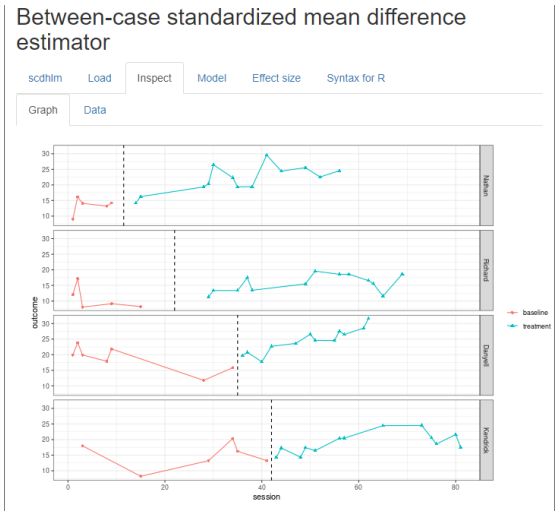


Figure 5.7: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Graph Display within the Inspect Tab for Datchuk (2016)

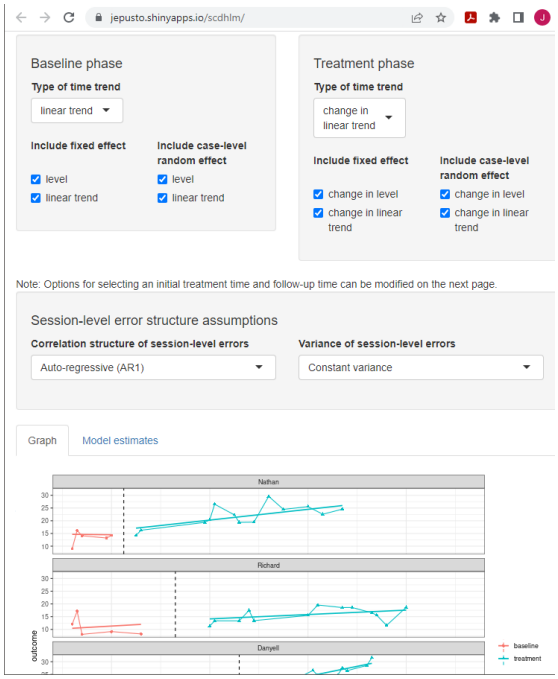


Figure 5.8: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Model Tab Showing Model 6 Specification for Datchuk (2016)

that fit the data reasonably well. Thus, we proceed to the *Effect Size* tab, as shown in Figure **??**. On the *Effect Size* tab screen, there are two sliders—*Initial treatment time* and *Follow-up time*. Moving these, we can change the size of the effect size to specify how far into treatment we want to estimate an effect. In our example, we use the default slider value of 9 for the *Initial treatment time*, which corresponds with the time point after which we would introduce treatment in a hypothetical study (were each case to begin the treatment phase at the same time). Next, because we want to estimate the treatment effect 12 observations into the treatment phase, we set the second slider (*Follow-up time*) to 21. This number corresponds to the 12th observation of a treatment phase that started immediately after observation 9. Following these specifications, the design-comparable effect size estimate for this study is 1.48, with a standard error (SE) of 0.98, and 95% confidence interval (CI) [-0.65, 3.62].

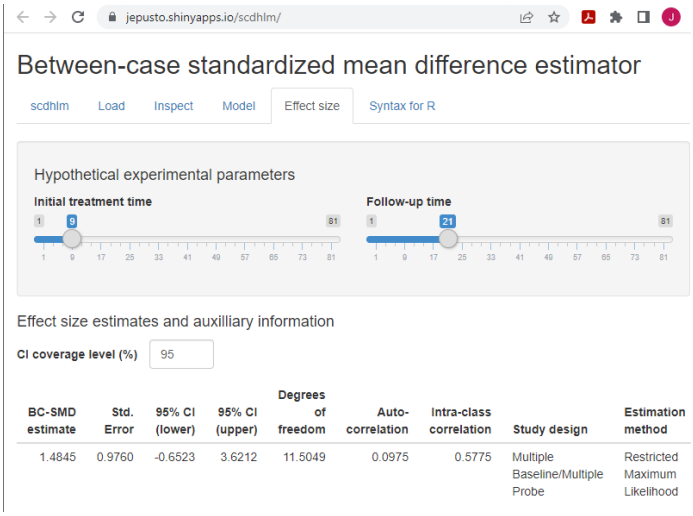


Figure 5.9: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Effect size Tab Showing Model 6 Estimate for Datchuk (2016)

The additional information reported on the *Effect size* tab (Figure **??**) includes estimates of other quantities from the model and information about the model specification and assumptions used in calculating the design-comparable effect size. The reported degrees of freedom are used in making a small-sample correction to the effect size estimate, analogous to the Hedges' g correction used with group designs (**?**). Larger estimated degrees of freedom mean that the denominator of the design-comparable effect size is more precisely estimated, and that the small-sample correction is less consequential. Conversely, small degrees of freedom indicates that the denominator of the effect size is imprecisely estimated, and that the small-sample correction is more consequential. The reported autocorrelation is the estimate of the correlation between errors at the first level of the model for the same case that differ by one time-point

(or session), based on a first-order autoregressive model. The reported intra-class correlation is an estimate of the between-case variance of the outcome as a proportion of the total variation in the outcome (including both between-case and within-case variance) as of the selected *Follow-up time*. Larger values of the intra-class correlation indicate that more of the variation in the outcome is between participants. The remaining information in the output (*Study design*, *Estimation method*, *Baseline specification*, *Treatment specification*, *Initial treatment time*, *Follow-up time*) describes the model specification and assumptions used in the effect size calculations. The app includes this information to allow for reproducibility of the calculations.

Having obtained our main design-comparable effect size estimate based on Model 6, we also want to check its sensitivity to our selected time at which we estimate the between-case variability. To do so, we move the sliders on the *Effect Size* tab so that the *Initial treatment time* slider is now set to 5, corresponding to the time of the last baseline observation in our hypothetical experiment. We change the *Follow-up time* to 17, which keeps the duration of the treatment at 12 observations. After making these changes, the design-comparable effect size is an estimated 1.34 with an SE of 0.91 and 95% CI [-.69, 3.37]. The sensitivity of the design-comparable effect size estimate to our choice of when to estimate the between-case variance demonstrates a potential challenge of using design-comparable effect sizes in contexts where it is assumed that the baseline slopes vary across cases.

If our a priori arguments for selecting Model 6 were tenuous, or if some members of our research team thought we should consider Model 5, we can do so by going back to the *Model* tab and changing our specification. Specifically, under the specification of the *Treatment phase*, we uncheck the boxes to exclude *change in level* and *change in linear trend* as random effects. By removing those two random effects, we constrain the treatment effect to be the same for all cases (i.e., the change in level and change in linear trend is the same for each case). We keep all other modeling options the same as Model 6 (e.g., time sliders). In Figure **??**, we can see that the simpler, more restrictive Model 5 specification has a slightly less desirable fit for the trend lines than those of Model 6. The estimated design-comparable effect size for Model 5 is 1.41, with an SE of 0.64, and 95% CI [-0.09, 2.92]. Ultimately, if we had remaining SCD studies to be included in our synthesis, we would hold off selecting between Models 5 and 6 until fit was examined for the data from each of the SCD studies.

5.3.2 Example 2: Multiple Baseline Study by ?

After estimating the design-comparable effect size for the first SCD study by **?**, we repeat these steps for all remaining single-case studies in our synthesis. For our second SCD study (**?**), we run through this same sequence of steps: 1. Load the data. 2. Inspect the data in both tabular and graphic form. 3. Specify our selected model for the data (i.e., Model 6 for this illustration). 4. Estimate the design-comparable effect size.

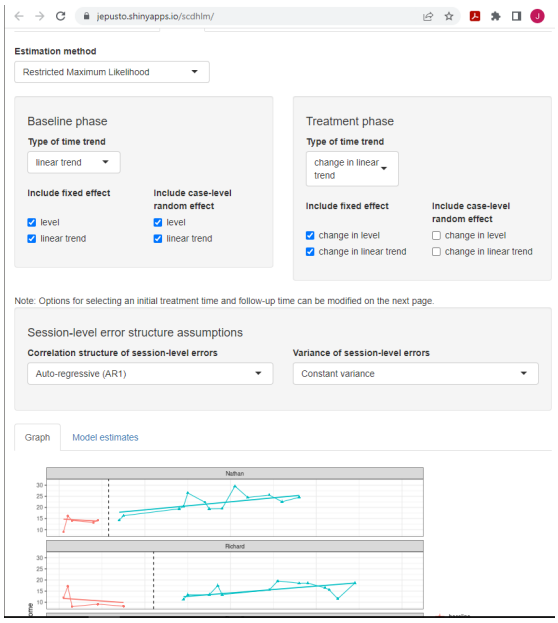


Figure 5.10: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Model 5 Specification for Datchuk (2016)

As shown in Figure **??**, the Model 6 estimated trajectories fit the data reasonably well. After completing step 2, we proceed with use of Model 6 for the **?** dataset (step 3) and then start the process of estimating the design-comparable effect size (step 4). On the *Effect Size* tab, we need to make choices about the initial treatment time and follow-up time-points. Locating the shortest baseline (Brett), we leave the *Initial treatment time* slider at the default value of 9—the last baseline observation for Brett. We set the *Follow-up time* slider to 21, so that the treatment duration for the hypothetical study is 12 and consistent with the time specification for the previous SCD study (**?**). After defining the times, the resulting design-comparable effect size estimate is 0.76 with an SE of 0.74 and 95% CI [-1.36, 2.88]. Finally, to check the sensitivity of our estimate to our specified times for between-case variability, we change the values of the *Initial treatment time* and *Follow-up time* sliders to 5 and 17, respectively. The resulting design-comparable effect size changes to 0.90 with an SE of 0.75 and 95% CI [-1.06, 2.87].

If our a priori arguments for selecting Model 6 were tenuous, or if some members of our research team thought we should consider Model 5, we can do so by going back to the *Model* tab for each of our SCD studies and changing our specification. Specifically, under the specification of the *Treatment phase*, we uncheck the boxes where we had previously included random effects for *change in level* and *change in linear trend*. By removing those two random effects, we

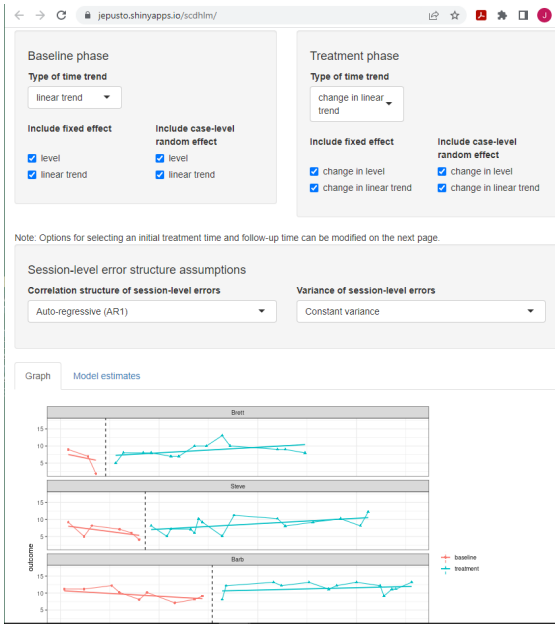


Figure 5.11: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Model 6 Specification for Rodgers (2021)

constrain the treatment effect to be the same for all cases (i.e., the change in level and change in linear trend is the same for each case). We keep all other modeling options the same. We show the Model 5 specification for the **?** study in Figure **??**. With the more restrictive model, the fit of the trend lines is similar but not quite as good compared to those in the originally specified Model 6. For Model 5, the estimated design-comparable effect size is 0.46, with an SE of 0.40 and 95% CI [-0.39, 1.31]. Across the included SCD studies, if the fit is similar across models or is better for the model selected a priori, then it is preferable for researchers to use the one that best aligns with their logic model. For this illustration, we move forward with the Model 6 estimates because this model is most consistent with our a priori expectations, and the resulting fit is similar to or slightly better than its Model 5 contrast.

5.4 Estimating the Design-Comparable Effect Size for the Group Study

After estimating the design-comparable effect size for each SCD in the synthesis, we estimate the design-comparable effect size for each included group study. Details on estimating standardized mean difference effect sizes from group studies are readily available from a variety of sources, including Chapter

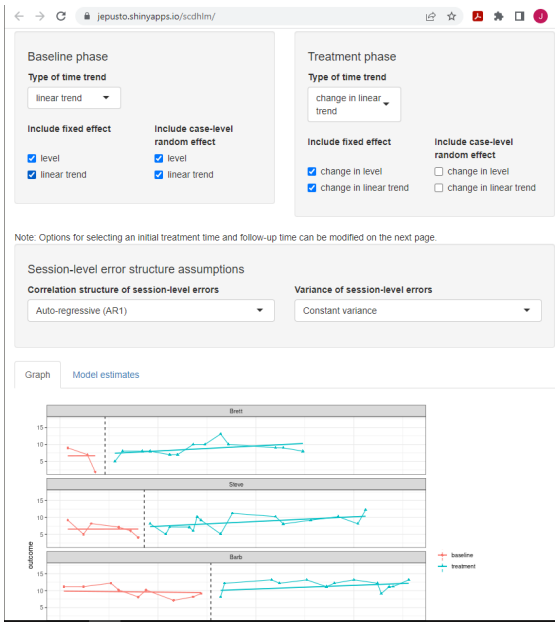


Figure 5.12: Between-Case Standardized Mean Difference Estimator (scdhlm, v. 0.6.0) Model 5 Specification for Rodgers (2021)

12 of *The Handbook on Research Synthesis and Meta-Analysis* (**?**), books (e.g., **?**) and journal articles (e.g., **??**). We therefore omit the details of the effect size calculations for the group design study in this example.

? reported a randomized-control trial examining the impact of a writing intervention for 51 fourth and fifth grade students. The researchers randomly assigned the students to an informational text writing intervention ($n = 32$) or to a control group ($n = 29$), in which students focused on mathematics writing so that the total amount of writing time was controlled. While the **?** study included various outcome data, we focus on the proximal measure of student ability to write informational text using simple description. To measure the effect, we calculate Hedges' g, which corrects for small sample size bias, using the adjusted mean difference between groups controlling for the pretest writing assessment and student characteristics. The resulting effect size estimate is 0.62, with an SE of 0.22.

5.5 Analyzing the Effect Sizes

As a final step, we synthesize the effect sizes across both the SCD and group design studies. Depending on the researchers' goals of the synthesis, a variety of options are available: (a) creating graphical displays that show the effect

size and CIs for each study, (b) averaging the effect sizes and then creating a CI for the overall average effect, (c) examining the extent of variation in effect sizes across studies, (d) exploring potential moderators of the effect, and (e) examining the effect sizes for evidence of publication bias. Because the use of design-comparable effect size for the single-case studies produces effect size estimates that are in the same metric as the commonly used standardized mean difference effect sizes from group studies, researchers can accomplish these goals using the methods developed for group studies. Details on these methods are readily available elsewhere (e.g., **??**). We illustrate the pooling of the effect sizes from our studies here using a fixed effect meta-analysis, consistent with the approach used in What Works Clearinghouse intervention reports (**?**).

Table **??** reports the effect size estimates, SEs, and fixed effect meta-analysis calculations for the example studies. The top panel uses the design-comparable effect size results for SCD studies based on Model 6. As a sensitivity analysis, the bottom panel presents the results based on Model 5. Note that the effect size estimate from the group design study is the same in both panels. In fixed effect meta-analysis, the overall average effect size estimate is a weighted average of the individual studies' effect size estimates, with weights proportional to the inverse of the sampling variance (squared SE) of each effect size estimate. Column C of Table **??** reports the inverse variance weight assigned to each of the studies, with the percentage of the total weight listed in parentheses. For the analysis based on Model 6, the effect size estimate from the group design study receives 88% of the total weight, while the effect size estimates from the single-case studies receive 4% and 8% of the total weight, respectively. The total inverse variance weight is 23.5. The overall average effect size estimate is 0.67 with an SE of 0.21[1] and an approximate 95% CI [0.26, 1.08]. The Q-test for heterogeneity is not significant, $Q(2) = 0.75$, $p = .69$, indicating that the evidence is consistent with the possibility that all the studies provide estimates of the same true effect size.

The bottom panel of Table **??** reports the same calculations using the Model 5 design-comparable effect size estimates for the two SCD studies. For both **?** and **?**, Model 5 effect sizes appear to be more precisely estimated and therefore receive more weight in the fixed effect meta-analysis (8% and 21%, respectively). The overall average effect size based on Model 5 is quite like the average effect size based on Model 6, but is somewhat more precisely estimated, with an SE of 0.18 and a 95% CI [0.29, 1.01], due to the smaller SEs in the SCD studies.

In fixed effect meta-analysis, the overall average effect size estimate is a summary of the effect size estimates across the included studies, which are treated as fixed. The SE and CI in fixed effect meta-analysis consider the uncertainty in the process of estimating the effect size estimates in each of the individual studies. However, they do not provide a basis for generalization beyond the included studies because the fixed meta-analysis does not account for uncertainty in the process of identifying studies for inclusion in the meta-analysis (**??**). When

[1]The SE of the overall effect size is the square root of the inverse of the total weight.

Table 5.1: Fixed Effect Meta-Analysis Calculations for Example Writing Intervention Studies

Study	Effect Size Estimate (A)	Standard Error (B)	Inverse-variance Weig
Model 6			
Datchuk (2016)	1.48	0.98	1.04 (4.4)
Rodgers et al. (2020)	0.76	0.74	1.82 (7.8)
Hebert et al. (2018)	0.62	0.22	20.66 (87.8
Fixed effect meta-analysis	0.67	0.21	25.53 (100)
Model 5			
Datchuk (2016)	1.41	0.46	2.44 (8.3)
Rodgers et al. (2020)	0.46	0.40	6.25 (21.3)
Hebert et al. (2018)	0.62	0.22	20.66 (70.4
Fixed effect meta-analysis	0.65	0.18	29.35 (100)

conducting syntheses of larger bodies of literature—and especially of studies with heterogeneous populations, design features, or dependent effect sizes—researchers will often prefer to use random effects models (**?**) or their further extensions (**??**).

Chapter 6

Introduction to Multilevel Modeling of Raw Participant Data

This chapter provides background on multilevel modeling of the raw data from studies included in a synthesis of SCDs. We describe when this approach is useful, then discuss various types of study outcomes (e.g., continuous, count, and percentage-based), corresponding assumptions (e.g., within-phase trends, normality), and modeling options for synthesizing findings using multilevel models of raw, individual participant data. We conclude this chapter by providing a set of decision rules for meta-analysts to use when selecting among the currently available multilevel modeling options for synthesis of SCDs.

6.1 Background

The use of multilevel models to analyze and synthesize effects from single-case studies is based on the recognition that the raw data often have a hierarchical data structure. Initially, two-level models were developed for data from single-case design (SCD) studies with multiple participants, such as multiple baseline designs, where observations are nested within cases (**???**). To conduct syntheses across multiple studies, researchers have expanded upon these methods to allow for three-level models where individual observations are nested within cases and cases are nested in studies (**????**).

The multilevel modeling approach is valuable when the outcome and time variables are operationalized in a common way across the cases and studies and when the researchers are interested in how treatment effects vary across cases or change over time. However, because this approach entails analyzing data

from multiple studies using a single multilevel model, it is limited to contexts where a common model is consistent with the logic model for the area of research and the included data. In this chapter, we describe when researchers should elect to use multilevel modeling of the raw data and the assumptions involved in using this approach for effect estimation and synthesis. We then provide a set of the most common modeling options, along with a decision-making guide for selecting among the various multilevel modeling options (e.g., when there are/not baseline trends).

6.2 When to Use Multilevel Models of the Raw Data

When researchers are synthesizing the literature in an area, the purpose or aims of the synthesis should guide their methodological choices. In some situations, researchers may want to explore the degree to which the effect changes over time. Yet, information about variation in intervention effect over time is lost when a single value is used to summarize the effect for a case (as with case-specific effect sizes) or for a full study (as with design-comparable effect sizes). Using a multilevel model of the raw data allows researchers to retain and examine this information. Researchers should also consider multilevel models if the study's aim is to document the time it takes for an intervention to have an effect or whether the effects decay over time. However, as we detail in the next section, the multilevel modeling approach assumes a common model for all cases in the included studies. Thus, multilevel modeling of the raw data is only viable when time is scaled the same way across studies and the outcomes are operationalized either (a) in the same way for each study or (b) in way that is similar enough to allow the outcome of any study to be equated, scaled, or standardized so that it matches the outcome for the other studies (e.g., converting all latency outcomes to minutes). See **?** for more details on standardizing and the efficacy of doing so.

When the outcome operationalizations are so disparate that the outcomes follow different probability distributions in different studies, it may be unreasonable to standardize the raw data and thus infeasible to use multilevel modeling. For example, consider two different study outcomes: "A" is the elapsed time from the beginning of a routine until the first disruptive behavior (latency), and "B" is the number of occurrences of the disruptive behavior during the routine (count). Outcome A might follow a normal distribution and Outcome B might follow a Poisson distribution. In a situation like this, the statistical models for each outcome are different, and we would not advise the use of a multilevel modeling approach. For this guide, we will focus on situations where the outcome is the same across cases and studies.

6.3 What We Assume with Multilevel Models of the Raw Data

There are several assumptions underlying the multilevel modeling approach to the synthesis of single-case data. The overarching assumption is that a specific multilevel model describes the process by which the raw data were generated. If we have observations nested in cases, and those cases are nested within studies, we assume a specific three-level model. The first level represents a model for within-case observations. At this level, we make assumptions about the presence or absence of trends within each phase and about the deviations of individual observations from the trend lines (e.g., normally or Poisson distributed). Different ways of modeling the distributions allow for different choices in the effect size metric (e.g., standardized mean difference versus log response ratio). The second level of the model accounts for variation in the trend lines and effect sizes between cases within a study (e.g., some cases may have a lower baseline average level of responding, or a larger shift in responding with intervention). Within the level-2 model, there are also distributional assumptions made, such as assuming the baseline levels for cases are normally distributed around some across-case average baseline level. Finally, the third level of the model accounts for variation between studies in the average baseline trend lines and effect sizes. Again, the level-3 model is specified with the typical assumption that the parameters of these average study trajectories are normally distributed around the parameters of the overall across-study average trajectory. In the following sections, we more fully define and elaborate on the underlying assumptions for multilevel modeling of raw single-case data.

6.3.1 Within-Case Model Assumptions

The within-case model accounts for the variation in the outcome measurements over time within a case. We must model the distribution of the observations around the trend lines, specify the trend line for each phase, and choose the effect size metric on which to quantify changes due to intervention (e.g., changes in level and/or trend between baseline and intervention phases). When specifying the distribution of observations around the trend line, we consider several characteristics: distribution shape (e.g., normal versus some non-normal distribution), whether deviations from the trend line are independently distributed or serially dependent, and whether the variance is homogeneous or heterogeneous across the study (e.g., when the baseline phase variance is larger than that of the treatment phase). When specifying the trend lines, we consider the presence of trends and whether they are linear or follow some non-linear form.

When making decisions about modeling the shape of the distribution of observations around the trend line, normality is the simplest and most common assumption (e.g., **???**), and leads to estimation via linear mixed models (LMMs). In some contexts, assuming a normal distribution aligns well with the single-case data. However, there are circumstances where it may not be appropriate

to make such an assumption. For example, when an SCD study outcome is a count variable with a low mean for one of the phases (e.g., frequency of a target behavior having a treatment phase mean at or near zero), a non-normal distribution like a Poisson, quasi-Poisson[1], or negative binomial may be more appropriate (e.g., **???**).

In other circumstances, the study outcome may be a percentage (e.g., the percentage of academically engaged time) with a phase mean that is close to the minimum (0%) or maximum (100%). When a study phase mean for a percentage variable is near the minimum or maximum value possible, the resulting non-normal distribution might be better approximated with a binomial, quasi-binomial[2], or beta-binomial distribution (**??**). Models involving non-normal probability distributions are known as generalized linear mixed models (GLMMs).

With GLMMs for non-normal distributions, it is common to use other effect size metrics that describe change in percentage or multiplicative terms. Outcomes measured as counts and proportions have bounded ranges: counts must be non-negative and proportions must be between zero and one. For such outcomes, additive changes may not generalize well because of floor or ceiling effects. For instance, a case with an average baseline latency of 20 seconds cannot logically experience a 25 second decrease in latency. In contrast, percentage or multiplicative changes are less constrained when outcome scales are bounded. Following the previous example, a case with an average baseline latency of 20 seconds can experience a 10% decrease or a 99% decrease in latency or could experience an increase of 2.2 times the baseline level. In GLMMs for count outcomes, it is common to specify the model in terms of the natural logarithm of the expected outcome (using a log link function), so intervention effects are quantified as log response ratios. In GLMMs for proportion outcomes, it is common to specify the model in terms of the log of the odds of the expected outcome (using a logistic link function), so that intervention effects are quantified as log odds ratios.

As an example, consider a multiple baseline study by **?** that examined the impact of a mathematics software intervention on the percentage of intervals where students were off task. Figure **??** presents the **?** study data. During baseline phases, observations appear somewhat normally distributed around the baseline means, with a pooled skewness value near 0 ($sk = 0.15$) and a pooled kurtosis value near 0 (ku = -0.77). However, non-normality is present in the treatment phases because the outcome occurs near the floor (0%) for much of the phase ($sk = 1.86$; $ku = 4.99$).

[1]Strictly speaking, the quasi-Poisson is not a true probability distribution but rather only an approximation. However, it functions in effectively the same way as probability distributions such as the Poisson or negative binomial when developing a statistical model for count outcomes.

[2]Similar to the quasi-Poisson, the quasi-binomial is not a true probability distribution but rather only an approximation.

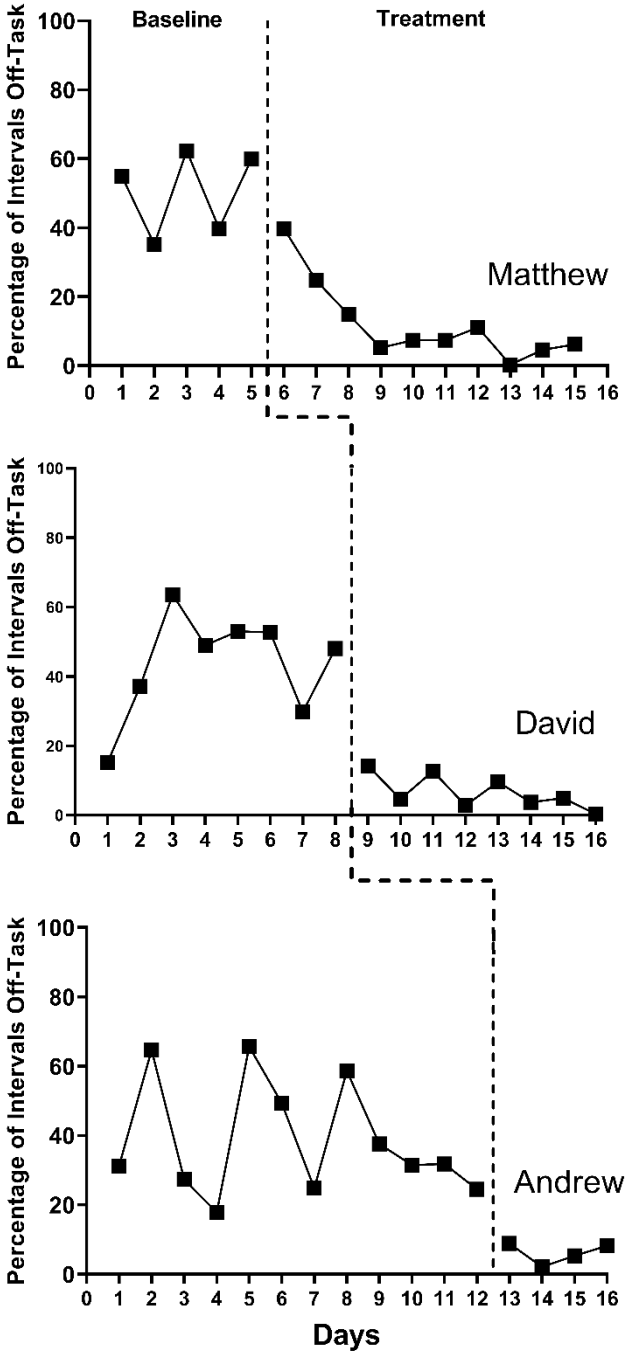


Figure 6.1: Multiple Baseline Design Across Three Participants (Ota & DuPaul, 2002)

There is some research suggesting that the simpler multilevel models based on the normality assumption (i.e., LMMs) can produce unbiased average effect estimates and accurate inferences, even when the data are not strictly normal (**??**). However, at some point the departures can become so severe that methods accounting for non-normality (i.e., GLMMs) appear preferable. We expect that future research will better define the point at which researchers should move to these more complex models with non-normal probability distributions. We also anticipate that future research will further define the sample sizes (i.e., series lengths, number of cases, number of studies) needed for estimation of the more complex GLMMs.

The decision about an appropriate distributional model for the outcomes is closely related to the metric on which effect sizes are measured within the model. Typical LMMs have parameters that describe additive changes in the outcome. As a result, the parameters that describe intervention effects share the same units as the outcome scale. For instance, if the outcome variable is a measure of latency in seconds, then an LMM will describe effect sizes in terms of additive shifts in latency, such as an immediate decrease of 20 seconds or (if the model includes linear trends in the treatment phase) a reduction of 4 seconds per intervention session.

In addition to making decisions about an appropriate shape for the distribution and an appropriate effect size metric, meta-analysts must decide whether the deviations from the trend line are independently distributed or serially dependent. If researchers assume serial dependency, they next decide how to model the serial dependence. The most commonly assumed model for serial dependence in SCD studies is a first-order autoregressive model. There is a long history of debate about the prevalence of serial dependence in single-case data (see **??**). Research suggests that the impact of modeling and not modeling serial dependence depends on whether the statistical model is a single-level model (i.e., a regression for a single-case), a two-level model (e.g., a multilevel model for a single multiple baseline study having four cases), or a three-level model (e.g., a multilevel model of raw data extracted from 10 multiple baseline studies). For single-level models where serial dependence is present, a failure to model it does not bias the regression coefficient providing the point estimate of the treatment effect. However, not modeling serial dependence in a regression model can bias the standard error (SE), leading to substantial problems when making inferences about the treatment effect (**??**). Two-level models not accounting for serial dependence have less severe inferential problems than single-level models, but some problems still persist (**?**). However, for three-level models, the choice about modeling potential serial dependence appears to have minimal, if any, impact on the treatment effect estimates or inferences (**??**). Furthermore, for two- and three-level models, inferential techniques such as robust variance estimation can be applied to provide some protection against mis-specified assumptions regarding error dependence. Because we are focusing on the application of multilevel models for synthesizing across studies (three-level models) in this methods guide, we will use the simpler models that assume independence in our

illustrations.

Next, when specifying the model for the distribution of observations around the trend line, the researchers must consider the degree to which the variability in those observations across phases is consistent (homogeneous) or different (heterogeneous). In GLMMs of count data, the variance is assumed to depend on the outcome level. However, for models assuming a normal distribution, it is typically assumed that the variance does not depend on the outcome level. In Figure **??**, one might question not only the normality assumption (as previously noted), but also the difference in variance between the two phases (i.e., less variance in the treatment phases than the baseline phases). When using models that assume a normal distribution of the outcomes, researchers can specify either that the variance is equal across phases or that the variance differs between the phases. The simpler homogeneous variance assumption may be a reasonable choice. Simulation research has demonstrated that when the Kenward Roger method was used for inference, the simpler model provided unbiased treatment effect estimates and accurate confidence intervals even when the treatment phase variance was up to four times that of the baseline variance (**?**). We anticipate that future research will help to provide more concrete guidance as to when changes in variance are substantial enough to warrant use of more complex models that assume heterogeneity.

Finally, the within-case model specifies the form of the growth trajectory for each phase. The simplest trajectory is one without trend, but researchers can specify linear or non-linear trends within the multilevel modeling framework (e.g., **???**). However, because the resulting effect estimates can vary substantially based on the modeling of trends, we encourage researchers to think carefully about the participants, outcomes, and contexts of the studies they are synthesizing. We also encourage researchers to visually analyze the data from each included SCD study. If visual inspection of the data reveals patterns generally consistent with expectations (e.g., the logic model and visual analysis both suggest baseline and treatment phase trends), then the researcher should specify the multilevel model to be consistent with the logic model and the raw data. Alternatively, when there is ambiguity about the appropriate model for the growth trajectory for the baseline and/or treatment phase, we suggest researchers estimate and contrast alternative plausible models to document the sensitivity of the effect estimate and inferences relative to the specification of the growth trajectory.

6.3.2 Case Similarity Assumptions

At level-2 of the multilevel model, we make assumptions about case similarity within a study. Generally, if we assume that the outcome for one case follows a normal distribution, we assume that the outcome for all cases within the study follows a normal distribution. Similarly, if we assume that the level-1 errors for one case are independent and homogenous, we assume that the errors for all cases are independent and homogeneous; and if we assume that trends are absent within a case, we assume a within-case model without trends is appropriate for

all cases.

In addition to assuming that the within-case model is appropriate for generalization across all cases, researchers must make assumptions related to the within-case model parameters. Researchers must decide which parameters in the within-case model are the same for all cases and which parameters vary randomly across cases. Typically, the within-case variance parameters are assumed to be the same for each case (e.g., the session-to-session outcome variation for one case equals the session-to-session variation in the outcome for any other case). In contrast, we typically assume that the regression coefficients (e.g., the change in level and slope parameters that index the treatment effect) vary from case to case and that these regression coefficients are normally distributed across the cases. Indeed, if the aims of the synthesis include investigating heterogeneity of treatment effects across participants, then using a model in which the regression coefficients vary across cases may be critical for addressing research questions of interest. Finally, for the regression coefficients that vary, we must consider whether they also covary. For example, if cases with lower baseline levels tend to show smaller treatment effects, there would be covariation between the coefficients. It is possible to specify multilevel models that assume covariation and models that do not.

We encourage researchers to consider the plausibility of the case similarity assumptions prior to jumping into a multilevel modeling application. At this time, methodological research is limited regarding the degree to which the case similarity assumptions can be violated, and thus it is not clear when more complex models and estimation approaches are preferable. **?** provide a method for relaxing the assumption that the within-case variance is constant across cases within a study. They also show that differences in the within-case variance can often be tolerated by models that assume homogeneity. For the illustrations in this methods guide, we assume that within-case variances are equal across cases, and that the regression coefficients that vary follow normal distributions. We anticipate that future research will continue to develop alternatives and guidance for contexts where these assumptions are not tenable.

6.3.3 Study Similarity Assumptions

At the third level of the multilevel model, we make assumptions about the similarity of the studies regarding the parameters at both of the first and second levels of the model. In terms of the model for observations nested within cases, we need to make assumptions about the similarity of distributions and variances. Typically, multilevel models and GLMMs assume that all outcomes from all studies follow a single distributional family (e.g., normal, negative binomial, beta-binomial). Furthermore, for linear mixed models, we typically assume that the session-to-session variation in the outcome is constant not only across the cases within each study, but also across all cases in all included studies. Analogously, for GLMMs involving distributions with dispersion parameters (such as the quasi-Poisson or negative binomial), we typically assume that the

dispersion is constant across all cases in all included studies.

Regarding the regression coefficients of the case-specific model, if we assume the baseline level varies randomly across cases within one study, we assume that the baseline levels vary randomly across cases in all studies. Similarly, if we assume that the treatment effect varies randomly across cases in one study, we assume that the treatment effect varies randomly across cases in all studies, and if covariation is assumed among the random effects in one study, covariation is assumed for all studies.

In addition to assuming that the general between-case model is appropriate for generalization across all studies, researchers must make assumptions related to the between-case model parameters (i.e., which parameters are the same for all studies and which parameters vary randomly across studies). Typically, the between-case variance parameters are assumed to be the same for each study. Consequently, there is an assumption of homogeneity across studies in the variation (and covariation) between cases. In contrast, we typically assume that the regression coefficients (e.g., the across case average change in level and slope parameters) vary from study to study and that these coefficients are normally distributed across the studies.

As with the previous sets of assumptions, we encourage researchers to consider the plausibility of the study similarity assumptions prior to applying a multilevel model. For example, if the participant sampling methods for one study are markedly different from the sampling methods of another study (e.g., homogeneous versus heterogeneous purposive sampling), the assumption of equal between-case variance has likely been violated. Similarly, if studies use heterogeneous approaches to measuring outcomes, such as widely varying observation session lengths, the assumption of common within-case variance may be implausible. Hopefully, future research will provide guidance about the degree to which multilevel models can tolerate some violation of these assumptions.

6.4 Comparison to Other Synthesis Approaches

Compared to the other broad approaches to synthesis of single-case studies, the assumptions involved in the raw data synthesis approach have some commonalities and some unique features. Both the raw data synthesis approach and design-comparable effect size syntheses are based on multilevel modeling, as the latter entails specifying a separate multilevel model for the data from each study to estimate a study-level summary effect size on a scale-free metric. The resulting design-comparable effect size can then be compared to the effect size from a between-group design. Thus, both approaches require making assumptions about case similarity within a study. Further, we have recommended using a common model specification when estimating design-comparable effect sizes, similar to how the raw data synthesis approach requires specifying a model that captures the form of time trends across all included studies. However, synthesis

of design-comparable effect sizes does not involve study similarity assumptions (such as homogeneity of outcome variance) to the same extent as the raw data synthesis approach. Of course, along with additional assumptions, the raw data synthesis model also provides richer descriptions of the distribution of and variation in effects—not only variation across studies, but also variation across cases within studies and variation over time.

In contrast to the design-comparable effect size approach and the raw data synthesis approach, syntheses of case-specific effect size indices do not entail as extensive a set of assumptions about case similarity. Rather, the case-specific effect size approach avoids making case similarity assumptions about regression coefficients or session-to-session outcome variation, and instead focuses on estimating a summary effect size index for each case. Because no case similarity assumptions are made, it is more difficult to accommodate complex data features such as time trends and autocorrelation than it is with multilevel modeling. Thus, synthesis using case-specific effect size indices loses some flexibility but gains a degree of simplicity by avoiding the need to make so many assumptions.

6.5 Multilevel Modeling Options for Synthesizing Single-Case Research

We summarize in this section the decisions to be made when choosing a specific multilevel model for synthesizing single-case research. For the studies included in the synthesis, we suggest that meta-analysts first attend to the operationalization of the outcome variable(s). For multilevel modeling of the raw data across studies, recall that outcomes must be operationalized in the same way, or in a manner similar enough so that all included study outcomes can be transformed or standardized so they consistently use the same unit of measurement. In considering the outcome, it is also important to determine if assuming a normal distribution is reasonable. If assuming normality is reasonable, the meta-analyst can turn to LMMs for estimation. However, if it is not reasonable to assume normality (as with some count-based variables), then the meta-analyst should consider GLMMs. Additionally, when using GLMM for non-normal data, they must consider the most appropriate probability distribution for modeling the outcome (e.g., Poison, quasi-Poisson, or negative binomial).

Regardless of whether the meta-analyst is assuming normality and working with a LMM or assuming some non-normal distribution and working with a GLMM, the next major decision to make is whether to model trends. We suggest researchers base their decision on: (a) the logic model underlying the research to be synthesized, considering what is known about the participants, outcome, intervention, and setting; and (b) visual analysis of the data from each of the studies being synthesized.

Regardless of whether trends are included, there are several further considerations they must make to fully define the variances and covariances that will

be estimated. For example, they will need to specify the deviations of observations from the trend line as either autocorrelated or independent. They will also need to specify whether the variance in the deviations from the trend line remains constant or changes over the course of the study. Finally, beyond the within-case assumptions of variability, the meta-analyst must determine if it is reasonable to assume homogeneity of within-case variance across all cases in all included studies, and whether between-case variance terms are homogeneous across studies.

If the meta-analyst specifies the model to include trends, their list of considerations is slightly longer. Not only do they consider the assumptions about the variances and covariances, as done with a model assuming no trends, but the meta-analyst must now address questions about the trends and measurement of time. More specifically, they must make decisions about whether there are trends in either or both baseline and treatment phases, whether the intervention affects the trend, and whether the trends are linear or nonlinear. Additionally, they must make assumptions about the time variable. For example, in one study context it may be assumed that slopes are attributed to maturation over time and the time variable is measured in calendar days. Conversely, in another study context, the researchers may attribute the within-case treatment phase slope to time exposed to the intervention, and the time variable may be measured on a scale defined as "sessions." Because meta-analysts must use the same multilevel model when conducting a synthesis of all included studies, they need to be consistent in how they approach the definition of the time variable. Finally, meta-analysts must determine when the effect should be indexed in the treatment phase, leading to decisions about how to center the time variable in the multilevel model. When centering or rescaling a time variable, meta-analysts effectively move the zero on the time scale so that it corresponds to the time at which they want to estimate the treatment effect. For example, a researcher who wants to estimate the immediate intervention effect will center time so that zero corresponds to the first treatment phase observation, whereas a researcher who wants to estimate the effect five sessions into intervention would center time so that zero corresponds to the fifth treatment phase observation.

In Figure **??**, we arrange these considerations into a series of decision rules that researchers can use to select between LMMs and GLMMs, and models with and without trends. Recent methodological research has started to develop more concrete guidance on the most appropriate modeling and estimation choices in contexts where non-normality is assumed and GLMMs are more appropriate (**?**). However, for the purposes of this guide, we illustrate the application of multilevel models for study contexts where it is reasonable to assume normality and thus LMMs are appropriate. We first consider the application of LMMs when expecting no trends (Chapter 7), and then address the more complex situation when anticipating trends (Chapter 8).

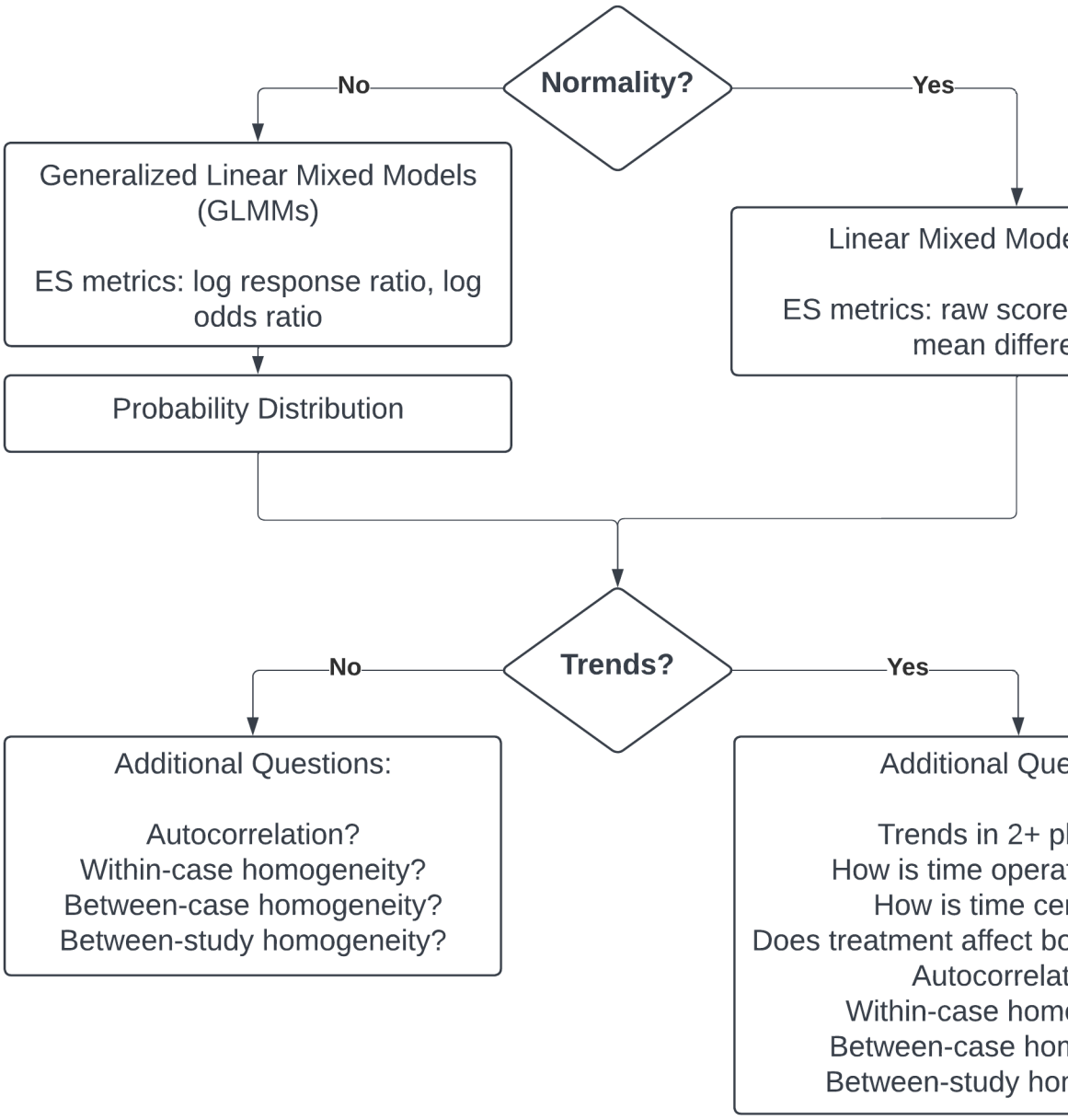


Figure 6.2: Flow Chart for the Selection of Multilevel Modeling Approach

Chapter 7

Illustration of Multilevel Modeling When No Trends Are Assumed

This chapter illustrates the application of multilevel models in contexts where one does not expect trends in either baseline or treatment phases. We provide instruction for using the MultiSCED app to specify and estimate the multilevel model to synthesize results from a set of single-case studies. We offer step-by-step instructions using data from four multiple baseline design studies examining methods to improve use of behavior-specific praise in classroom contexts.

In this chapter, we illustrate the application of multilevel modeling in contexts where one does not expect time trends to occur for a set of single-case design (SCD) studies included in a synthesis. Multilevel models assume that the analysis focuses on the same outcome for each case in each study. In addition to this assumption, the multilevel model explored in this chapter is specified based on the prediction that the dependent variable for each case has a stable level throughout the baseline and treatment phases and that introduction of treatment may result in an immediate shift in the level of the dependent variable.

We present an example scenario where we want to synthesize evidence from several SCD studies that examine intervention effects on the improvement of behavior specific praise (BSP) used by educational staff (e.g., teachers or aides). It is of note that while all studies are implemented by educational staff in English, they took place in different countries (e.g., U.S. and Ghana) and included different intervention ingredients. Due to the paucity of research investigating the effects of BSP implemented by educators (teachers or teacher assistants) in school settings, and the need to succinctly model this process, we limited our illustration to four multiple baseline (MB) design studies, described in fur-

ther detail below. In one study, **?** examined the impact of one-on-one training coupled with emailed visual performance feedback on four first-year general education classroom teachers' use of BSP with their students. In a second study, **?** investigated whether prompts emailed daily to second and fifth grade classroom teachers would promote their use of BSP with students. The third study by **?** examined the effects of a multicomponent (self-monitoring, performance feedback, goal setting, modeling, and action planning) teacher-led training program on the improvement of BSP used by four paraprofessionals with elementary students identified with specific learning disabilities or other health impairment. Finally, **?** studied how self-monitoring coupled with performance feedback affected the use of BSP by four classroom staff (who were also caregivers) assigned to work with students with autism spectrum disorder (ages 7-17 years).

Synthesizing the data from these four studies involves estimating the average treatment effect on the scale of the outcome variable by modeling the outcome data for all cases in all studies. While our goal is feasible because each study examined the same outcome, the studies measured the BSP outcomes on different scales. More specifically, the included studies reported BSP frequencies using different time intervals (e.g., frequency per 10 minutes versus 15 minutes). To make the outcome comparable across the studies, we converted all frequencies to report a rate of BSP per 10 minutes. When the included studies had comparable outcomes, we chose to synthesize the results using the multilevel modeling approach (see the decision rules in Figure **??**). In this chapter, we break our illustration into two stages: (a) selecting a multilevel model for the studies, and (b) estimating the multilevel model using the MultiSCED app, a web-based calculator for conducting multilevel models of single-case studies (**?**).

7.1 Selecting a Multilevel Model for the Single-Case Studies

Using the decision rules in Figure **??** (see Chapter 6), we first select the general multilevel modeling approach. We must decide whether we can treat the outcome as a continuous variable that is normally distributed around the trend lines. With an outcome rate of BSP per 10 minutes, it would be reasonable to consider a generalized linear mixed model (GLMM) that does not assume normality. However, GLMMs are more complex than linear mixed models (LMMs) that assume normality, and there is limited evidence of GLMM utility for synthesizing SCD data. For this methods guide, we will examine our expectations about the outcome for the contexts studied. In addition, we will visually analyze the graphed raw outcome data for each case over time to gauge the severity of any violation to the normality assumption. If not too severe, we will proceed with using an LMM, which can produce unbiased estimates of treatment effects quantified as mean differences on the outcome scale, along with appropriate effect inferences in a variety of contexts where the outcome is not normally distributed [**?**; Joo_Ferron_2019]. In the future, we anticipate that method-

ological research will better clarify the relative merits of GLMMs versus LMMs and provide clearer guidance for researchers to select the best option for their models.

After selecting our general multilevel modeling approach (i.e., LMM), we then use our prior knowledge and experience with the dependent variable to state our expectations for data patterns within and across phases and cases. Here, we anticipate that levels of BSP will be lower than desirable in baseline. However, we predict that some skills already exist within participants' repertoires, so they will not occur at floor levels (i.e., we do not expect participants' baselines to consist of consecutive values of zero). Conversely, we predict the frequencies to be further from zero and to more closely approximate a normal distribution after introducing the intervention.

We present the graphs for all four MB studies in Figure **??** (**?**), Figure **??** (**?**), Figure **??** (**?**), and Figure **??** (**?**). The graphs for the first three studies (Figures **??** - **??**) align with our expectations; none of the participants have notable floor effects. Except for two outlying values for Teacher 1 at the beginning of the baseline phase in Figure **??**, the outcome distributions across cases appear relatively normal. However, in Figure **??**, the baseline rate of BSP for two participants is zero, and a third participant has many baseline values of zero. Upon further inspection of the **?** study, we hypothesize that the intervention setting of Ghana and adult participants being caregivers (not professional educators) could account for lower levels of BSP in baseline. To better understand the degree to which the normality assumption was violated, we pooled the deviations from the phase means for each case in each study and estimated the skew and kurtosis of the distribution of the deviations. Resulting summary indices suggest moderate departures from normality in baseline ($sk = 1.49$; $ku = 6.35$), and approximately normal distributions during intervention ($sk = 0.31$; $ku = 1.22$). Because the distributions are not severely non-normal and prior evidence suggests that LMMs for SCD data can handle some non-normality in the outcomes (**??**), we deem it appropriate to proceed with LMM for the purpose of illustrating multilevel modeling using MultiSCED.

Next, we must decide whether to include trends in our model. We expect the BSP outcome for each case in each study to be relatively stable during baseline and not likely to systematically trend up or down in the absence of intervention. Visual inspections of the study graphs (see Figures **??**-**??**) are largely consistent with our expectation. We discussed the baseline of Teacher 1 in Figure **??**, and concluded the first two observations were more likely outliers than evidence of a systematic linear trend. Based on our understanding of the BSP outcome and visual inspection of the graphs which show no trend for most cases, it appears reasonable to use a multilevel model without baseline trends.

Next, we turn to the treatment phases. We expect that the intervention will lead to an immediate shift in the rate of BSP and that participants will maintain this higher level throughout the treatment phase. In Figures **??** and **??**, respectively, visual inspection of the treatment phases for **?** and **?** reveals

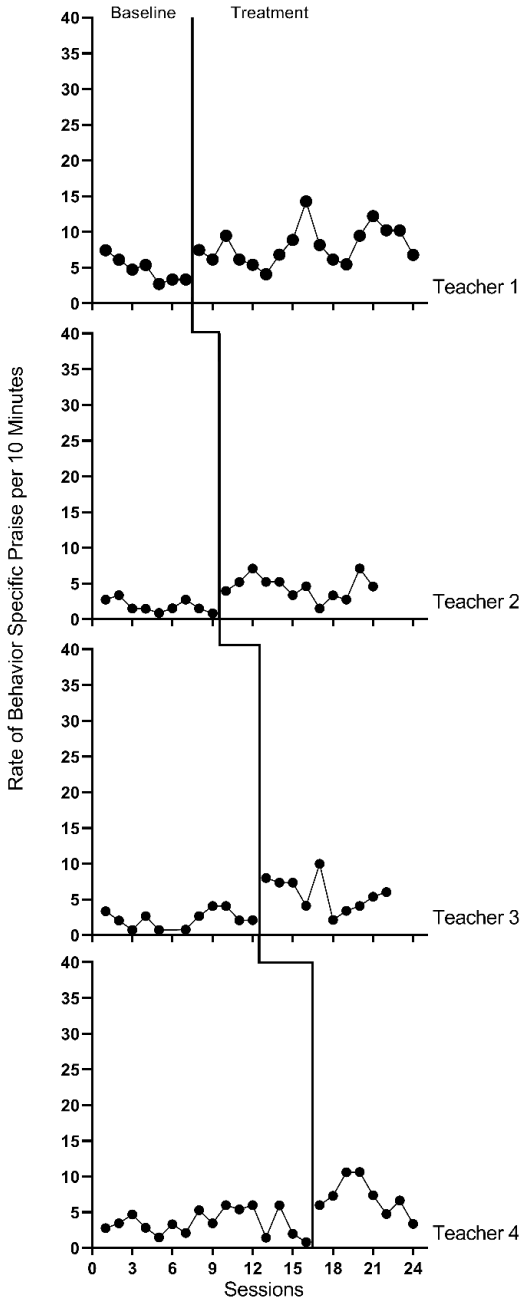


Figure 7.1: Effect of Intervention on the Rate of Behavior Specific Praise (Gage et al., 2018)

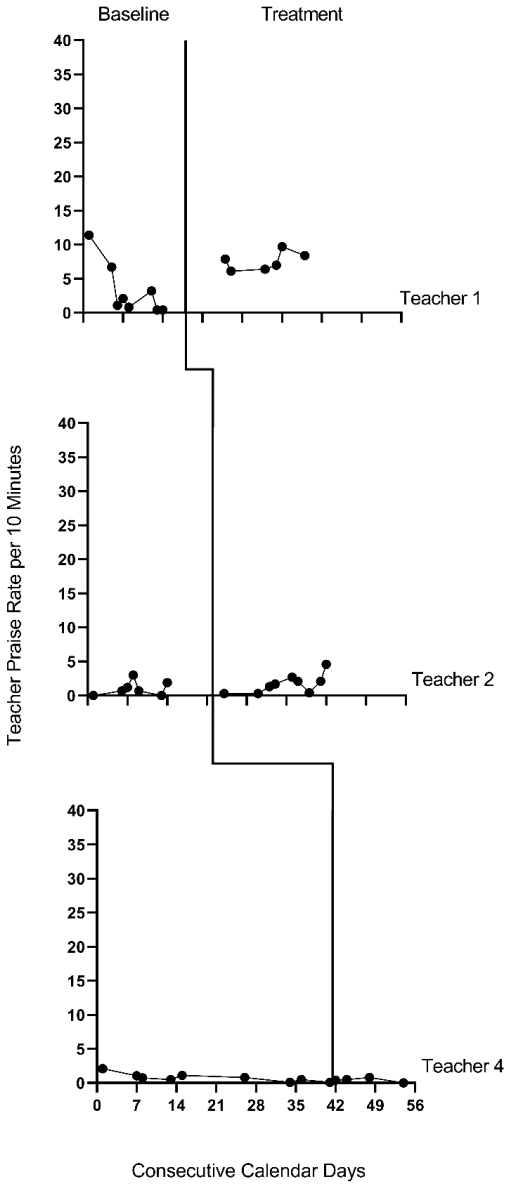


Figure 7.2: Effect of Intervention on the Rate of Behavior Specific Praise (Collier-Meek et al., 2017)

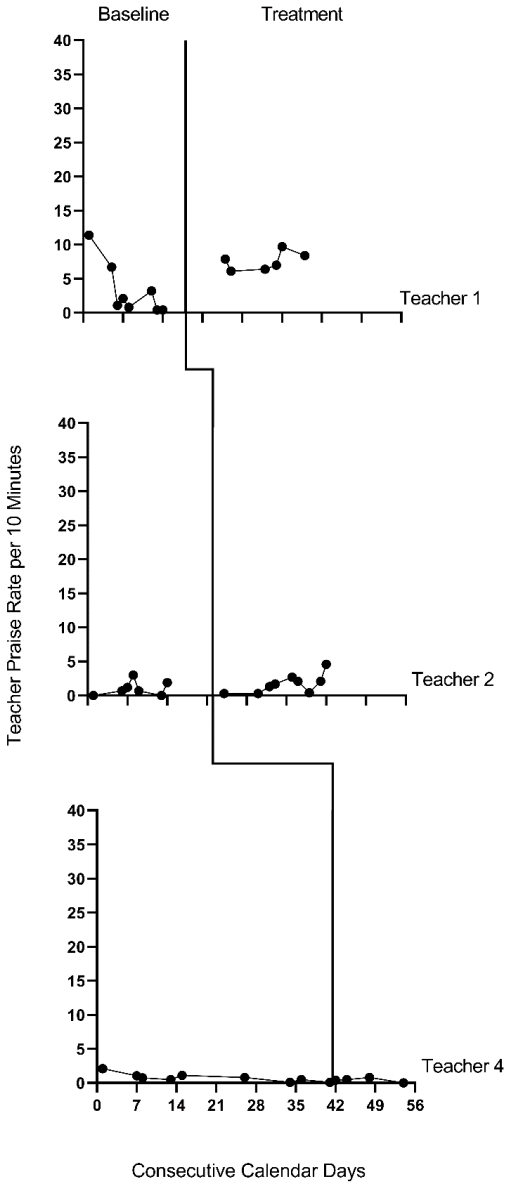


Figure 7.3: Effect of Intervention on the Rate of Behavior Specific Praise (Sallese & Vannest, 2022)

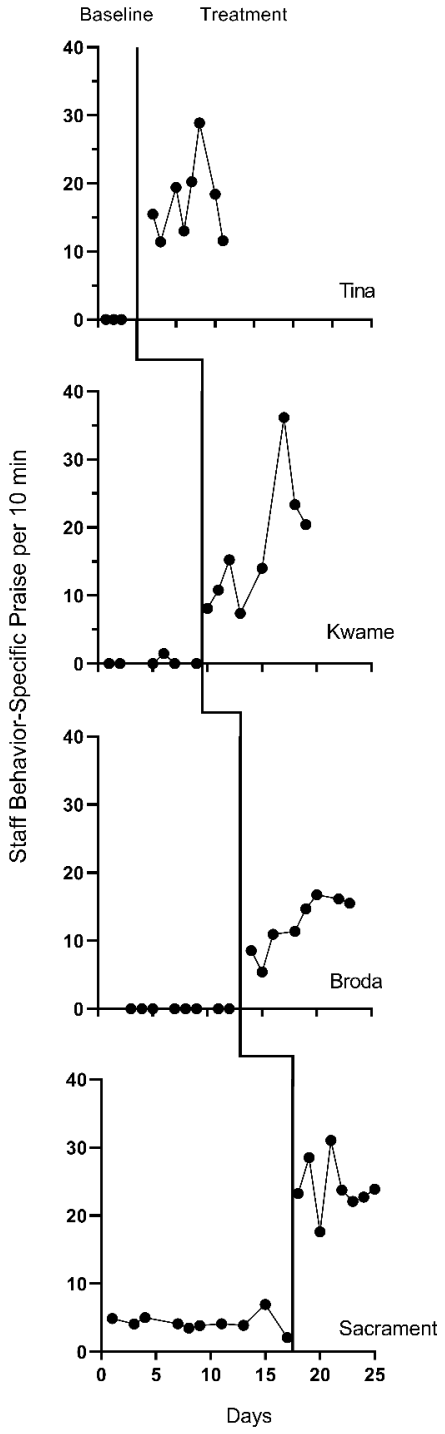


Figure 7.4: Effect of Intervention on the Rate of Behavior Specific Praise (Knochel et al., 2021)

relatively stable responding, as anticipated. However, in the **?** study (Figure **??**), upward trends appear for some of the participants as the rate of BSP increases over time in intervention (e.g., Mia). In Figure **??** (**?**), we see that the rate of BSP for two participants (Kwame and Broda) increases with time spent in intervention (unlike the other participants who have relatively stable rates throughout treatment). Because multilevel modeling requires us to assume a common model (e.g., trends for all or no cases), we proceed with estimating a model that assumes no treatment phase trends. For the three cases that appear to have treatment phase trends, their contribution to the across-case average effect estimate will be based on their across-time average effect, not on their initial effect or final effect. In addition, by not modeling the trends for these few individuals, we anticipate this decision will impact their residuals (deviations of their observations from their trend lines) in a manner that potentially increases the within-case error variance and degree of autocorrelation. However, as previously noted, these models are relatively robust to small or moderate violations of the (co)variance assumptions.

After selecting a model that assumes no trends, we turn next to an examination of the within-case variance using visual analysis of the graphed data and prior knowledge of the BSP outcome. In Figures **??**-**??**, we note increased variation in participants' treatment phase outcome data, as expected for an outcome variable reported as a rate. Thus, we anticipate increases in the rate of BSP will lead to an increase in the variance of BSP. For example, the change in variance is pronounced for the **?** participants (Figure **??**). Their baseline levels are relatively low, and treatment phase levels are relatively high. To define the between-phase variance change more precisely, we pool the deviations across cases for each phase. The standard deviation (SD) for the baseline phase deviations is 1.52. The SD of treatment phase deviations is 3.17. While we could estimate a heterogeneous variance model, multilevel models that assume homogeneity can tolerate differences in variance of this magnitude (**?**).

In summary, our visual analyses of primary study data identified multiple potential violations to the assumptions of the basic LMM for SCD studies, like evidence of baseline phase non-normality, across-phase heterogeneity, and a few cases that present with treatment phase trends. Thus, the LMM we use in this chapter is surely simpler than the process that generated the real data under analysis. However, the multilevel model estimate of the across-case average treatment effect and the associated inferences (e.g., confidence interval, significance tests) are relatively robust to violations of the normality and homogeneity assumptions. In addition, our purpose is to increase accessibility by illustrating the use of the MultiSCED app, and this app is limited to LMMs and homogeneous variance assumptions. Thus, we proceed to illustrate the estimation of the no-trend LMM.

7.2 Details of the No-Trend Multilevel Model

The multilevel data structure has repeated observations that are nested within cases and cases that are nested within studies, and thus there is a model for the variation of observations within a case (level-1), a model for the variation between cases within a study (level-2), and a model for variation between studies (level-3). The formal specification of the no-trend within-case model (level-1) is:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} Tx_{ijk} + e_{ijk}, \tag{7.1}$$

where Y_ijk is the outcome for variable Y at measurement occasion i for case j of study k and Tx_{ijk} is dummy coded with a value of 0 for baseline phase observations and a value of 1 for treatment phase observations. The mean baseline level for case j of study k is β_{0j} (see Figure **??** for an example visually depicting β_{0jk} and β_{1jk}). The raw score treatment effect for case j of study k is indexed by β_{1jk}, which is the difference between the baseline and treatment phase outcome means. The error term, e_{ijk}, is time-, case-, and study-specific. The error variance, σ_e^2, is typically assumed to be independent and normally distributed; however, it is feasible for researchers to estimate a different variance structure for e_{ijk}, such as first-order autoregressive or heterogeneous across phases (**??**). Typically, the error variance σ_e^2 is treated as constant across all cases in all included studies.[1]

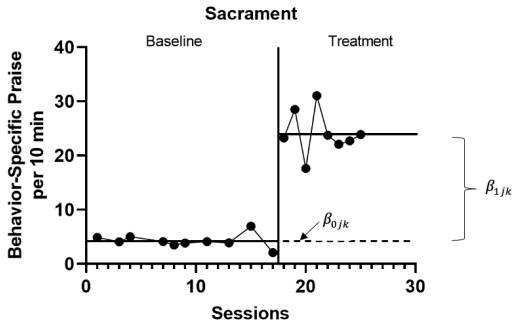


Figure 7.5: Illustration of Treatment Effect for Sacrament (Knochel et al., 2021)

The between-case model (level-2) is:

$$\beta_{0jk} = \gamma_{00k} + u_{0jk} \tag{7.2}$$

$$\beta_{1jk} = \gamma_{10k} + u_{1jk} \tag{7.3}$$

[1] In principle, this assumption could be relaxed, allowing for the error variance to differ from case to case or from study to study (**?**). However, the MultiSCED app does not allow such models to be fit, and some evidence indicates that average treatment effect size estimates (i.e., change in level) and related inferences are relatively robust to this sort of violation of the homogeneity assumption (**?**).

where γ_{00k} is the across-case average baseline mean for study k, and γ_{10k} is the across-case average treatment effect for study k. The case-specific errors, u_{0jk} and u_{1jk}, correspond to the deviations of case j of study k from the across-case average baseline level for study k and the across-case average treatment effect for study k. In the MultiSCED app, the case-specific errors (u_{0jk} and u_{1jk}) are assumed multivariate normal with covariance $\Sigma_u = \begin{bmatrix} \sigma^2_{u_0} & \\ \sigma_{u_1 u_0} & \sigma^2_{u_1} \end{bmatrix}$, and this covariance matrix is assumed to be the same across all included studies. In some applications, an unstructured covariance matrix will be difficult to estimate with the limited number of cases available. A common alternative, which could be estimated using more general multilevel modeling software, is a diagonal covariance structure where the covariance is assumed to be zero $\Sigma_u = \begin{bmatrix} \sigma^2_{u_0} & \\ 0 & \sigma^2_{u_1} \end{bmatrix}$.

The between-study model (level-3) is:

$$\gamma_{00k} = \theta_{000} + v_{00k} \tag{7.4}$$

$$\gamma_{10k} = \theta_{100} + v_{10k} \tag{7.5}$$

where θ_{000} is the across-study average baseline mean and θ_{100} is the across-study average treatment effect. The study-specific errors, v_{00k} and v_{10k}, correspond to the deviations of study k from the across-study average baseline level and the across-study average treatment effect. Within the MultiSCED app, the study-specific errors (v_{00k} and v_{10k}) are assumed to be multivariate normal with covariance $\Sigma_v = \begin{bmatrix} \sigma^2_{v_{00}} & \\ \sigma_{v_{10} v_{00}} & \sigma^2_{v_{10}} \end{bmatrix}$. In some contexts, an unstructured covariance matrix will be difficult to estimate with the limited number of studies available, so meta-analysts might prefer a simpler covariance structure (e.g., diagonal covariance structure), which could be estimated using more general multilevel modeling software.

For the purposes of the methods guide, we dedicate the remainder of this chapter to illustrating the estimation of the multilevel model for the BSP studies using the MultiSCED app (**?**). This app limits users to LMMs and the basic covariance structures described in this section. Using this relatively simple model, we contrast the obtained results with the results of a more complex model that assumes heterogeneous variances across phases and autocorrelation. In doing so, we provide a way to examine the sensitivity of estimates and inferences to alternative covariance structure specifications. Researchers can estimate alternative covariance structures outside of the MultiSCED app using statistical software programs such as R or SAS (see the Appendix to this chapter).

7.3 Estimating a Multilevel Model for the Behavior Specific Praise Studies

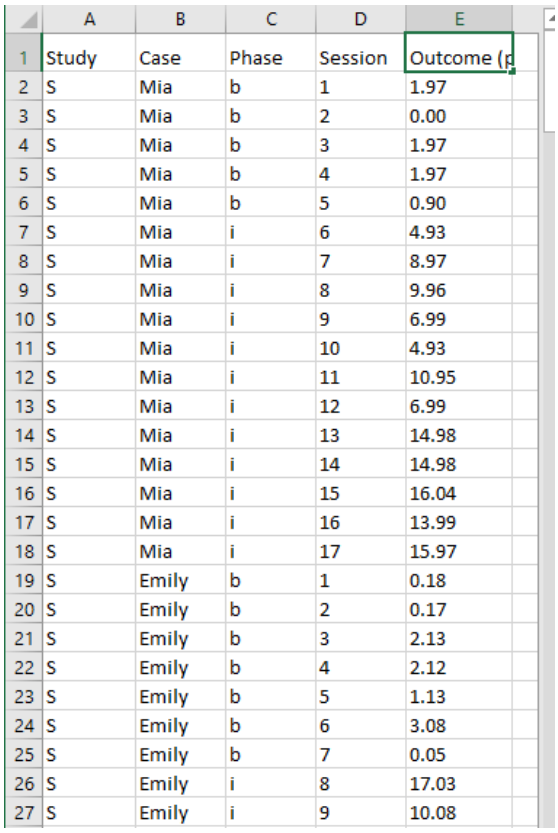
Using the web-based MultiSCED calculator (**?**) for the multilevel modeling of SCD data, we can estimate the across-case average treatment effect, along with a standard error and confidence interval. The MultiSCED app is available at http://34.251.13.245/MultiSCED/. To use this app, researchers must save their dataset as a text file (.txt) or a comma-delimited file (.csv). We recommend that researchers first enter their data in Excel, and then save the Excel dataset as a tab-delimited text file.

Figure **??** shows a screenshot of a portion of the original Excel spreadsheet containing the data extracted from our four BSP multiple baseline design studies. The spreadsheet contains five columns of data corresponding to the following five variables commonly used in multilevel modeling applications: (a) study identifier (i.e., Study), (b) case identifier (i.e., Case), (c) phase identifier (i.e., Phase), (d) point in time for which the outcome was measured (i.e., Session), and (e) outcome [i.e., Outcome (per 10 min)]. Because our model assumes no trends and no autocorrelation structure, we technically do not need the session variable, but it is acceptable to leave it in the dataset. In contrast, we do require the study identifier, case identifier, phase identifier, and outcome for specifying the model. With a more complex application including moderators of interest, additional columns should be included that contain values of potential moderating variables.

To use the MultiSCED app, the session and outcome variables must be numerical values. However, the variables for study, case, and phase identifiers may be entered as alphanumeric (word labels) or numeric values. The app requires that we organize the spreadsheet rows using a long data format, where the value of the outcome at each time point for a specific case in a specific study is represented in a unique row of the dataset, with the rows sorted by study, case, and time point. For each study (e.g., S in Figure **??**), we first arrange cases by baseline length (in order of shortest to longest baseline length). Then, we enter all observations for one case at a time, in order of time. After entering all data for the case with the shortest baseline, we enter data for the case with the second-shortest baseline, and so on for all cases within a particular study. We repeat the process for the remaining studies, with data from each successive study entered below the data from the previous study.

After organizing the data within the Excel spreadsheet, we then save the Excel spreadsheet as a tab delimited text file (.txt). To do the same, one can follow the steps below. For Windows:

1. From the *File* menu, select *Save As*.
2. Select *Browse* to select a folder where the spreadsheet will be saved.
3. To save the spreadsheet, enter a file name in the corresponding field.
4. Then use the *Save as Type* menu to select the option to save the file type



	A	B	C	D	E
1	Study	Case	Phase	Session	Outcome (p
2	S	Mia	b	1	1.97
3	S	Mia	b	2	0.00
4	S	Mia	b	3	1.97
5	S	Mia	b	4	1.97
6	S	Mia	b	5	0.90
7	S	Mia	i	6	4.93
8	S	Mia	i	7	8.97
9	S	Mia	i	8	9.96
10	S	Mia	i	9	6.99
11	S	Mia	i	10	4.93
12	S	Mia	i	11	10.95
13	S	Mia	i	12	6.99
14	S	Mia	i	13	14.98
15	S	Mia	i	14	14.98
16	S	Mia	i	15	16.04
17	S	Mia	i	16	13.99
18	S	Mia	i	17	15.97
19	S	Emily	b	1	0.18
20	S	Emily	b	2	0.17
21	S	Emily	b	3	2.13
22	S	Emily	b	4	2.12
23	S	Emily	b	5	1.13
24	S	Emily	b	6	3.08
25	S	Emily	b	7	0.05
26	S	Emily	i	8	17.03
27	S	Emily	i	9	10.08

Figure 7.6: Screenshot of the Original Excel Spreadsheet Containing the BSP Data

as Text (Tab delimited) (*.txt).
5. Finally, click *Save*.

If there are multiple worksheets within the original Excel spreadsheet file, a pop-up message will appear to request permission to save only the active sheet. If this occurs, you can select OK to save the data file. We strongly recommend that researchers open the newly created tab delimited text file to ensure that the data are stored as expected. We present a screenshot of a portion of the resulting text file in Figure **??**.

For MacOS:

1. From the *File* menu, select *Save As*.
2. Enter a file name in the *Save As* field.
3. Select a location from the dropdown menu, *Where*, to save your spreadsheet.
4. Then use the *File Format* menu to select the option to save your file type as Tab delimited Text (.txt).
5. Finally, click *Save*.

It is important to note here that an error message will appear if attempting to save a spreadsheet with multiple worksheets. A pop-up will inform the user that the workbook cannot be saved in the selected file format. If encountering this error, researchers should open a newly created file with one worksheet for the data needed in their analysis.

After converting and saving our dataset as a tab delimited text file, we are prepared to use the MultiSCED app (http://34.251.13.245/MultiSCED/). Figure @**??**fig:MultiSCED-homepage) is a screenshot of the Home page for the MultiSCED app. At the top of the website, in the blue page header, there are several links to other pages on the website. In this header, the link *Home* is highlighted with a darker background than the other links on the menu bar, indicating that we are on the *Home* page. To access the MultiSCED application, we must navigate to the *Input* section of the site, by clicking on the *Input* link.

A screenshot of the *Input* tab is shown in Figure **??**. After navigating to the *Input* page, we find a sidebar menu with three additional links: *Data file*, *Variables*, and *Data summary*. We first select the option on the left to choose the *Data file*. While *Data file* is highlighted in dark blue (as in Figure **??**), we can upload our dataset (.txt or .csv) by clicking the *Browse* button, selecting the folder where the data file is saved, and clicking *Open*. After doing so, two additional menus appear in the *Upload box* beneath the *Browse* button: *Separator character* and *Decimal character*. The default option for the *Separator character* is *tab*. We keep the default setting because we already saved the dataset as a tab delimited text file. However, if your file is saved as a .csv file, you can select *comma* under the *Separator character* dropdown menu. Regarding the *Decimal character* menu, the option for decimal representation defaults to dot, which is the standard choice in the U.S. However, for researchers located outside of the U.S. that use commas to mark decimals (e.g., EU countries), the

MLM Extracted Data 3 (BSP).txt - Notepad

File Edit Format View Help

Study	Case	Phase	Session	Outcome (per 10 min)
S	Mia	b	1	1.97
S	Mia	b	2	0.00
S	Mia	b	3	1.97
S	Mia	b	4	1.97
S	Mia	b	5	0.90
S	Mia	i	6	4.93
S	Mia	i	7	8.97
S	Mia	i	8	9.96
S	Mia	i	9	6.99
S	Mia	i	10	4.93
S	Mia	i	11	10.95
S	Mia	i	12	6.99
S	Mia	i	13	14.98
S	Mia	i	14	14.98
S	Mia	i	15	16.04
S	Mia	i	16	13.99
S	Mia	i	17	15.97
S	Emily	b	1	0.18
S	Emily	b	2	0.17
S	Emily	b	3	2.13
S	Emily	b	4	2.12
S	Emily	b	5	1.13
S	Emily	b	6	3.08
S	Emily	b	7	0.05
S	Emily	i	8	17.03
S	Emily	i	9	10.08
S	Emily	i	10	16.03
S	Emily	i	11	14.96
S	Emily	i	12	16.99
S	Emily	i	13	22.87
S	Emily	i	14	12.98
S	Emily	i	15	21.95
S	Emily	i	16	20.96
S	Emily	i	17	19.90
S	Emily	i	18	18.91
S	Emily	i	19	21.84

Figure 7.7: Screenshot of Saved Tab Delimited Text File

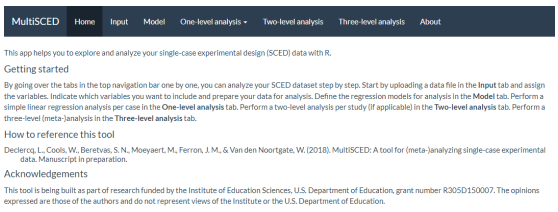


Figure 7.8: Screenshot of the Website Navigation Menu and Home Page of the MultiSCED App

setting will need to be changed from the default of dot to the comma option. We note here that researchers can use sample data to practice using the app. To access sample data, researchers should upload any file (which will not be used) and then click on the *Use testdata* checkbox. Sample data information will be automatically input into the *Variables* section which follows. After completing these actions, we transition to the *Variables* section of the *Input* page using the sidebar link on the left side of the screen.

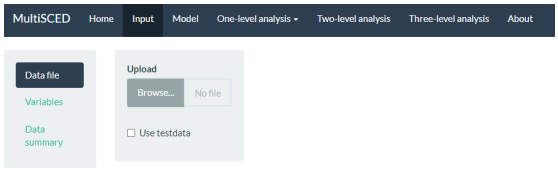


Figure 7.9: Screenshot of the Input Page of the MultiSCED App

As shown in Figure **??**, the *Variables* section appears on the left of the screen under *Data file*. When highlighted in dark blue, the *Variables* page has two primary sections: *Base variables* and *Moderator variables*. Using the drop-down menus in the *Base variables* section, we select our outcome variable from the *Response* menu (i.e., *Outcome per 10 min*), case identifier from the *Case* menu, study identifier from the *Study* menu, and phase identifier from the *Phase* variable. We also use the *Phase control group* drop-down menu to specify the value we used for baseline observations (i.e., *b*) and specify our session variable from options in the *Time* drop-down menu (i.e., Session). Because we have no moderators and no trends in this illustration, we leave the moderator and time centering options unselected.

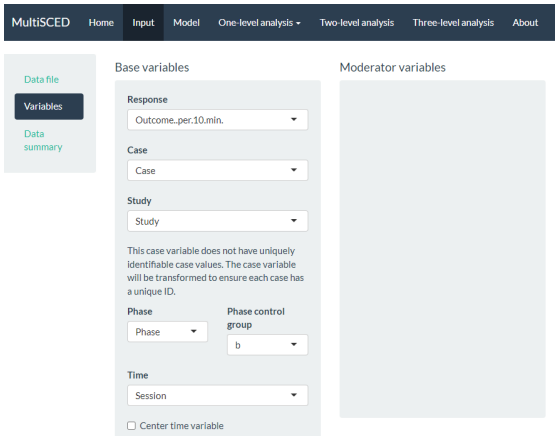


Figure 7.10: Screenshot of the Variables Section on the Input Page of the MultiSCED App

After specifying our variables, we navigate to the *Data summary* section, as

shown in Figure **??**. At the top of this section, the app summarizes the data for each of the three levels in our model: the number of studies in our dataset (i.e., 4), the number of cases nested within these four studies (i.e., 15), and the total number of observations of the BSP outcome (i.e., 297) that are nested within the cases nested within the studies. These values correspond with what we know about our dataset.

Next, we see blue checkboxes next to each variable under the column headers *Studies* (i.e., S, G, C, K) and *Cases* (e.g., S_Mia), These checkmarks indicate that the app automatically included our entire dataset in the analysis. This default is appropriate for the purpose of this methods guide, as we intend to demonstrate the process for synthesizing data across studies and simultaneously fitting a model to all cases and all studies. As an aside, the MultiSCED app allows for statistical models of a single case or for two-level models of cases nested within a single study. Thus, it provides the *(De)select all* option (see Figure **??**) to allow researchers to include only those studies and cases of interest in their analysis.

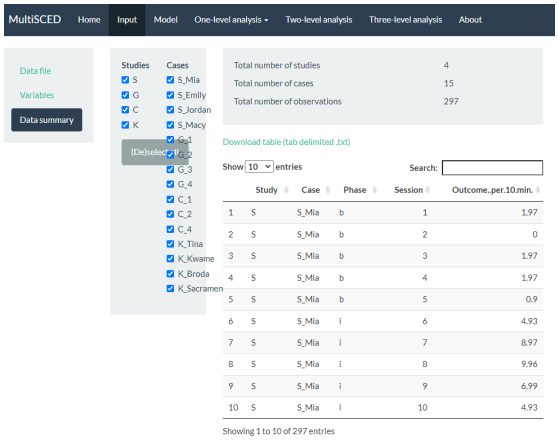


Figure 7.11: Screenshot of the Data summary Section on the Input Page of the MultiSCED App

After examining the *Data summary*, we are ready to specify the multilevel model. To do so, we navigate to the *Model* page by clicking the link in the navigation header bar at the top of the screen. There are five sections to the *Model* page (see Figure **??** that represents the model post-specification).

Under the *Fixed effects* subsection, we want to keep the default *intercept* and *Phase* options selected. We do not check the boxes next to the *Session* and *Phase X Session* because our model does not include trends. Under the *Random effects* section, there is nothing checked by default. To define the model to match Equations (**??**)-(**??**), we need to check each of the *intercept* and *Phase* options under the *Case level* and *Study level* subheadings. By checking the *intercept* box

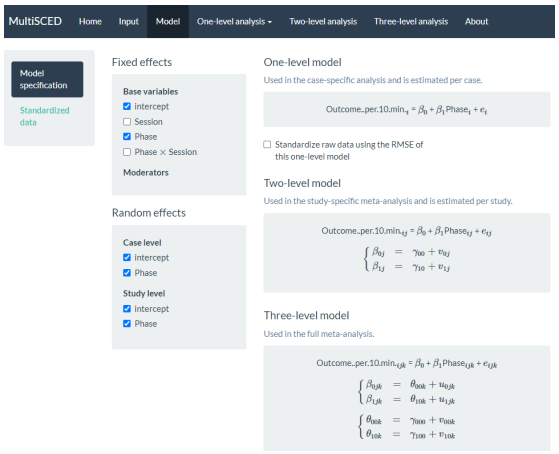


Figure 7.12: Screenshot of the Model Specification Section of the Model Page of the MultiSCED App

under *Case level*, we allow the baseline level to vary across cases within a study, and by checking the *Phase* box under *Case level*, we allow the treatment effect to vary across cases within a study. Similarly, under the *Study level* sub header, we allow the average study baseline level to vary across studies by checking the box next to *intercept*, and by checking the box next to *Phase*, we allow the average treatment effect for a study to vary across the studies.

As we opt to include these *Case level* and *Study level* variables, our model specification automatically updates and appears on the right side of the screen. Under the *One-level model* equation, note that we did not opt to standardize the raw data using the root mean square error (RMSE) because the outcome scale across our studies was the same (see Figure **??** for the full model specification that matches the model defined by Equations (**??**)-(**??**)).

In some situations, it may be difficult to estimate so many variances obtained from data extracted from a set of SCD studies. By removing some of the random effects (i.e., eliminating some of the checks under Random effects), researchers can analyze simpler models. Logic models for SCD data that expect baseline levels and treatment effects to vary across cases and studies align well with models that include the full set of variance components. Inclusion of a full set of variance components also aligns well with what researchers have typically used in methodological studies focused on evaluating the appropriateness of three-level models for single-case data (**???**). Therefore, we encourage researchers to start by including those variance estimates that correspond with their defined model, and only simplify the variance structure to remove the error term(s) if estimation problems are encountered that suggest that some of the variance components may be 0 or close to zero. When exploring such modifications, researchers should use their best judgement to specify a model that is reasonably

consistent with their logic model and with the data they are analyzing.

After specifying our model, we finally navigate to the multilevel model results via the *Three-level analysis* link in the website header navigation menu at the top of the screen (as shown in Figure **??**). On the *Three-level analysis* page, the model we estimated appears on the left, alongside two tabbed sections, *Table* and *Plot*. The app defaults to the *Table* tab, where we find the model results under *Fixed effects* and *Random effects* sub headers. *Fixed effects* contains the average value for the baseline level (i.e., *Intercept*). For the *Intercept* variable shown in Figure **??**, the average baseline rate of BSP per 10 minutes is 2.54, with an SE of 0.60. The estimated average treatment effect across all cases and studies is 7.95, with an SE of 3.41. We can interpret this result to mean that on average, the participants engaged in approximately 8 more BSPs per 10 minutes following intervention. However, the p-value for the treatment effect is .102, so we cannot be confident that the true mean effect of treatment differs from zero (i.e., we cannot reject the null hypothesis of no treatment effect).

We next examine the results under the *Random effects* sub header, shown at the bottom of the screenshot in Figure **??**. The MultiSCED app reports the *Random effects* variance components as SDs. Unlike the fixed effects, which have been shown to be well estimated even with the small sample sizes typical of single-case research, the variance components are typically not well estimated. Rather, we expect that they were estimated with considerable bias due to the limited number of studies and cases in this example (**??**). Furthermore, our examination of the data prior to modeling also suggested that we likely oversimplified our model by specifying an independent and homogeneous variance structure, which leads to even more bias in the estimated variance components (**?**). Consequently, while we report these parameter estimates from our model as a conceptual illustration for the purpose of this methods guide, we do not suggest that users of the app give serious weight to the interpretation of the *Random effects* variance components unless they have a larger number of studies and greater confidence in the accuracy of the model specification.

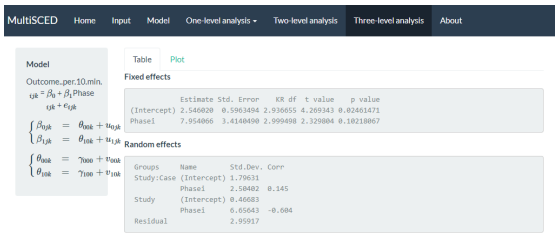


Figure 7.13: Screenshot of the Three-level analysis of the MultiSCED App

Per the results reported under the *Random effects* sub header (shown above), the estimated SD of the observations around the case-specific phase means (i.e., *Residual*) is 2.96. We did not obtain separate estimates for baseline and treat-

Table 7.1: Estimates Obtained Across Model Variations

Model	Effect Estimate	SE	95% CI	
			LL	U
Original specification	7.95	3.41	-2.92	18
Within-case heterogeneity + autocorrelation	7.94	3.40	-2.93	18
Within-case heterogeneity + autocorrelation + diagonal covariance matrices	7.93	3.38	-2.87	18

ment phases because we assumed a homogenous variance model. Next, examining the variability between cases nested within studies (i.e., *Study: Case*), we see that the estimated SD of the case-specific baseline levels is 1.80, the estimated SD of the case-specific treatment effects is 2.50, and the case-specific treatment effects are estimated to be correlated to the case-specific baseline levels by .15. The model results reporting the variability between studies (i.e., *Study*) estimated the SD of the study-specific average baseline levels as 0.47, the estimated study-specific average treatment effects SD is 6.66, and the estimated correlation between the study-specific average baseline levels and the study-specific average treatment effects is -.60.

Although the MultiSCED app provides the overall results, its ability to estimate more complex models is limited. We need to use other tools if we want to examine sensitivity of our results under more complex models, such as models that allow for heterogeneity in the within-case variance between phases or models that allow for autocorrelation. We believe this question may be of interest for meta-analysts, so we moved outside of the MultiSCED app. Using SAS, we estimated a more complex model that allowed for both potential differences in within-case variance between phases and autocorrelation of successive observations. We also estimated a model where we assumed heterogeneity across phases and autocorrelation, but using a diagonal (versus unstructured) covariance matrix at level-2 and level-3. Table **??** reports estimates of the overall average effects based on each of our three model specifications. Although the latter two model details and their estimation are outside the scope of this methods guide, we provide the syntax in the Appendix for those interested.

Generally, we reach the same conclusion—the point estimate of the average treatment effect is similar across models, and we cannot rule out the possibility that the overall average effect is zero. Research shows treatment effect estimates and inferences from LMMs of SCD data are relatively robust to violations of distributional assumptions, particularly when they incorporate Kenward-Roger-based small sample size adjustments (**???**). Therefore, we anticipated similarity in results across modeling options. In some contexts, however, the assumption violations will be more severe than what we encountered here. Relevant directions for future research include: (a) determining the point at which violations are severe enough and sample sizes large enough to warrant movement from LMMs to GLMMs, (b) identifying the most appropriate methods for specify-

ing and estimating GLMMs, and (c) producing easy-to-use software to make GLMMs more readily accessible to single-case researchers.

7.4 Appendix

7.4.1 SAS Code

Original specification: a three-level model assuming independent, normally distributed session-level errors with homogeneous variance across phases and cases and unstructured covariance matrices at level-2 and level-3.

```
proc mixed covtest;
class case study phase;
model outcome = tx / s ddfm=kr cl;
random intercept tx / sub=study type=un;
random intercept tx / sub=case(study) type=un;
run;
```

Three-level model assuming heterogeneity across phases, a first-order autoregressive covariance structure within the case, and unstructured covariance matrices at level-2 and level-3.

```
proc mixed covtest;
class case study phase;
model outcome = tx / s ddfm=kr cl;
random intercept tx / sub=study type=un;
random intercept tx / sub=case(study) type=un;
repeated / sub=case(study) group=phase type=ar(1);
run;
```

Three-level model assuming heterogeneity across phases, a first-order autoregressive covariance structure within the case, and diagonal covariance matrices at level-2 and level-3.

```
proc mixed covtest;
class case study phase;
model outcome = tx / s ddfm=kr cl;
random intercept tx / sub=study;
random intercept tx / sub=case(study);
repeated / sub=case(study) group=phase type=ar(1);
run;
```

7.4.2 R Code

Original specification: a three-level model assuming independent, normally distributed session-level errors with homogeneous variance across phases and cases and unstructured covariance matrices at level-2 and level-3.

```
model_homog_unstructured <-
  lme(
    fixed = outcome ~ phase,
    random = ~ phase | study / case,
    data = BSP_dat,
    method = "REML"
  )

summary(model_homog_unstructured)
intervals(model_homog_unstructured, which = "fixed")
```

Three-level model assuming heterogeneity across phases, a first-order autoregressive covariance structure within the case, and unstructured covariance matrices at level-2 and level-3.

```
model_heter_unstructured <-
  lme(
    fixed = outcome ~ phase,
    random = ~ phase | study / case,
    correlation = corAR1(0, ~ 1 | study / case),
    weights = varIdent(form = ~ 1 | phase),
    data = BSP_dat,
    method = "REML"
  )

summary(model_heter_unstructured)
intervals(model_heter_unstructured, which = "fixed")
```

Three-level model assuming heterogeneity across phases, a first-order autoregressive covariance structure within the case, and diagonal covariance matrices at level-2 and level-3.

```
model_heter_diag <-
  lme(
    fixed = outcome ~ phase,
    random = list(
      study = pdDiag(~ 1 + phase),
      case = pdDiag(~ 1 + phase)
    ),
    correlation = corAR1(0, ~ 1 | study/case),
    weights = varIdent(form = ~ 1 | phase),
    data = BSP_dat,
    method = "REML"
  )

summary(model_diag)
intervals(model_diag, which = "fixed")
```

Chapter 8

Illustration of Multilevel Modeling When No Trends Are Assumed

This chapter demonstrates the application of multilevel models that include time trends in the baseline and/or treatment phases. For illustrative purposes, we use data from four single-case studies examining interventions for improving student writing. We provide step-by-step instructions to demonstrate organization of the data, necessary coding of variables, appropriate multilevel model selection, and model specification and estimation using the MultiSCED app.

In this chapter, we illustrate the application of multilevel modeling in contexts where one expects single-case design (SCD) study data to include time trends. Just as in the previous chapter, the approach presented in this chapter assumes that all included SCD studies measured the dependent variable in the same way, or in ways that can be standardized so that the outcomes and measures of time are operationalized using a common scale. Unique to this chapter, we demonstrate the specification and analysis process for a model in which the dependent variable is expected to systematically increase or decrease during a phase, and the intervention is expected to impact the outcome's level and rate of change over time. As a running example, we imagine a scenario where our goal is to synthesize evidence from SCDs studies that examine intervention effects on writing outcomes—specifically on students' rate of correct word sequences written per minute.

Writing interventions are often necessary when students' writing skill lags behind their peers. Due to natural school-based instructional activities, we anticipate that some students may show slight improvement during baseline. However, only when the struggling writers begin intervention do we expect an im-

mediate increase in the production of correct word sequences, and an increase in their rate of improvement over time. To illustrate the steps for applying multilevel models in contexts where trends are expected, we use a set of four multiple probe and multiple baseline SCDs previously included in a larger review of intervention effects on student writing outcomes (**?**). In a real, full-scale meta-analysis project, we would synthesize all studies meeting the inclusion criteria. However, our illustrations include only four studies from the larger review to succinctly illustrate the details of the modeling process that includes visually analyzing the data of each primary study for consistency with our expectations and with the assumptions of the selected multilevel model.

In two multiple probe studies, **?** and **?**, researchers examined the impact of a rule-based direct instruction intervention on the expressive writing of high schoolers with learning disabilities. In our third study, **?** investigated the effects of a self-graphing intervention using a multiple baseline design across three fourth graders with high-incidence disabilities. In the fourth study, **?** examined the effects of a planning and drafting intervention on writing using a multiple baseline design across six elementary students with attention difficulties. While this final study included two participants with only two baseline observations, we include it here for the purposes of our illustration. However, for other purposes, researchers might choose to exclude the **?** study because its design does not meet the *What Works Clearinghouse Standards* (**?**) for multiple baseline designs.

In this chapter illustration, our goal is to model the outcome data from all cases in all studies to estimate the average effect of writing interventions on the rate of writing correct word sequences. This objective is feasible because each of the studies measures the number of correct word sequences written by students. To analyze these outcomes on a common scale, we convert all data to correct word sequences per minute.

Because our model includes trends, we must make two important decisions about the most appropriate method for operationally defining a common time variable. First, to determine the unit of time measurement, we consider how the outcome we are trying to measure changes over time. For instance, changes could occur linearly across calendar days, across school days, or across intervention sessions (which might occur more or less frequently than once per school day). If we anticipate changes occurring during baseline due to maturation, the calendar day unit of measurement might be appropriate. If we anticipate baseline change to occur related to regular teaching and learning activities in the classroom, it may be best to measure the outcome using the unit of school days. Alternatively, if we expect changes to occur with each additional intervention session, then perhaps session number is the ideal choice for the units of time.

Next, when primary studies use different units to measure time, meta-analysts must convert raw data to use a consistent unit for time in the analysis. In three of our four included studies, the authors reported "sessions" as the only unit of time (the exception was **?**, who measured the outcome across consecutive school days). Although less ideal than a more precise time specification (e.g.,

days), our operationalization of time to model trends is limited the use of "sessions" because no other information was available within the **?** or Walker (**?**; **?**) manuscripts. While we acknowledge the potential biases in estimating effects with a less precise definition of time, we find it reasonable to use "sessions" as the unit of time because we expect writing to improve with practice, and writing practice is largely expected to occur within the observation sessions. We recommend meta-analysts in a similar situation provide rationales for their selection of a common time variable so these limitations can be taken into consideration when reviewing the effect estimation results.

After operationalizing our outcome measure and time variables, we proceed with the selection of a multilevel modeling approach in effect estimation because we are interested in how intervention impacts both the level and slope of students' writing over time (see the decision rules in Figure 1.1). Thus, we divide this chapter illustration into two stages: (a) selecting a multilevel model for the studies and (b) estimating the multilevel model using the MultiSCED app, a web-based application for conducting multilevel models of SCD studies (**?**).

8.1 Selecting a Multilevel Model for the Single-Case Studies

Using the decision rules in Figure **??**, we first select among the general multilevel modeling approaches. We must decide whether we should treat the outcome as a continuous variable that is normally distributed around the trend lines. The common study outcome is a rate of correct writing sequences per minute, which is based on a count. Thus, it would be reasonable to consider a generalized linear mixed model (GLMM) that does not assume normality. However, GLMMs are more complex than linear mixed models (LMMs) and there is limited evidence of GLMM utility for synthesizing SCDs. For the purposes of this methods guide, we will examine our expectations about the outcome for the contexts studied. In addition, we will visually analyze the graphed outcome data over time per case to gauge the potential severity of any violation to the normality assumption. If not too severe, we will proceed with LMMs because there is evidence of their ability to produce unbiased estimates of treatment effects quantified as average mean differences or mean changes in linear slope on the scale of the outcome, along with appropriate inferences, even in contexts where outcomes are not normally distributed (**??**). In the future, we anticipate that further methodological research will better clarify the relative merits of GLMMs versus LMMs and provide clearer guidance for researchers to select among the options.

In this illustration, the outcome is students' rate of correct word sequences written per minute. Given our prior knowledge and experience with this dependent variable, we anticipate that outcome levels will be lower than desirable in baseline, but not at the floor. That is, we do not expect participants' baselines to solely contain observations with values of zero. Following intervention, we pre-

dict the rates of correctly written word sequences to be further from zero and to more closely approximate a normal distribution.

We present the included study graphs in Figure **??** (**?**), Figure **??** (**?**), Figure **??** (**?**), and Figure **??** (**?**). For each of these studies, we visually inspect the graphs to rule out obvious non-normality (i.e., constant values within a phase, or trend lines so close to the floor or ceiling that variation on one side of the trend line is notably different than on the other side). Visual analysis of the graphed data in Figures **??**-**??** suggest variation around the trend lines that are not obviously non-normal. To better understand the potential degree of a normality assumption violation, we pooled the deviations from the ordinary least squares estimates of within-phase trend lines for each case included in the studies and estimated the skew and kurtosis of the distribution of the deviations. Resulting summary indices suggest moderate departures from normality in baseline ($sk = 1.56$; $ku = 7.43$), and a closer-to-normal distribution in the treatment phase ($sk = 0.11$; $ku = 4.79$). Because the distributions are not severely non-normal and research suggests that LMMs for SCD data can handle some non-normality in the outcomes (**??**), we proceed with use of an LMM, which also allows us to illustrate the MultiSCED app.

Next, we must decide whether to include trends in our model. The outcome in each study is the rate of writing correct word sequences. Using prior knowledge of this academic outcome and the context of the studies, it is plausible that some participants show a slight degree of improvement in the writing outcome during the baseline phase because students engage in writing activities in a variety of academic contexts. However, we do not anticipate a general positive linear trend to be present across all cases in all studies. Unlike our expectations for baseline data, we are more confident in our assumption that the interventions will result not only in an immediate and noticeable increase in students' correctly written word sequences, but also in a change in the rate of growth. Therefore, we plan to model trends in the treatment phases.

Visual inspections of the graphs in Figures **??**-**??** are partially consistent with our expectations for baseline and treatment phase trends. Overall, we do not see a consistent pattern in trends for participants' baseline writing rates. Turning to the treatment phases, visual inspection of participants' graphs across studies reveals a general positive linear trend in the data as expected. For most cases, the treatment phase trends appear steeper than any trends present in the baseline phase, but the pattern of positive trends across the cases is not consistent. There is also variability in intervention duration (i.e., treatment phase length). However, based on our understanding of the study outcome (i.e., writing) and context (i.e., school setting), and consistent trends during intervention across most cases (especially in Figures **??** and **??**), it appears reasonable to use a multilevel model with trends (at least in the treatment phases).

We turn next to an examination of the within-case variance using visual analysis of study graphs. Figures **??**-**??** show similar variation across baseline and treatment phases and a follow-up statistical analysis of the deviations from trend

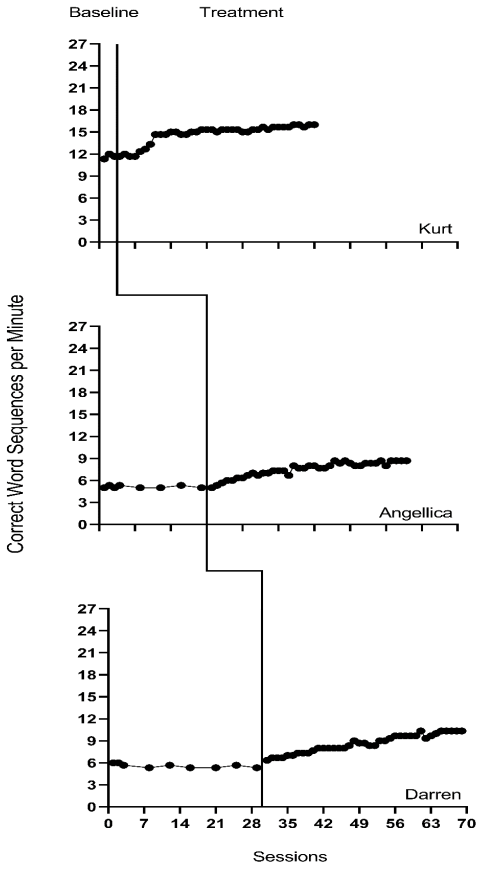


Figure 8.1: Effect of Intervention on Correct Word Sequences (Walker et al., 2005)

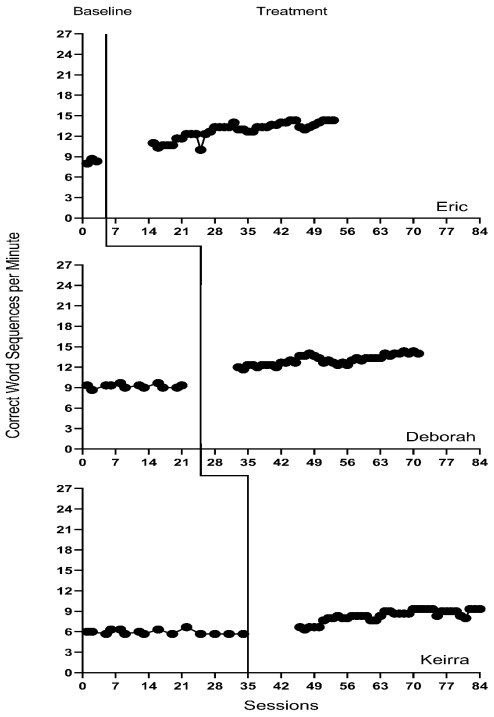


Figure 8.2: Effect of Intervention on Correct Word Sequences (Walker et al., 2007)

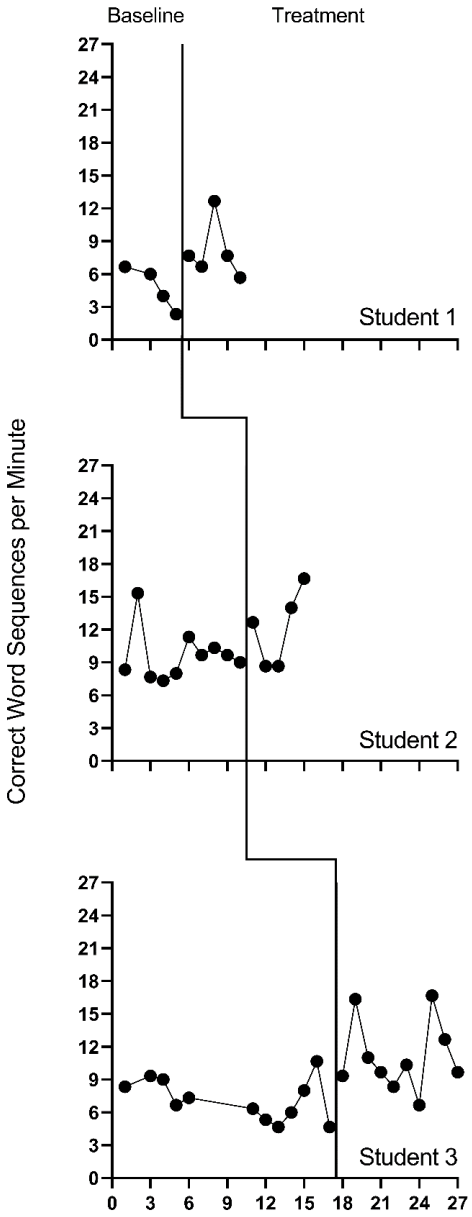


Figure 8.3: Effect of Intervention on Correct Word Sequences (Stotz et al., 2008)

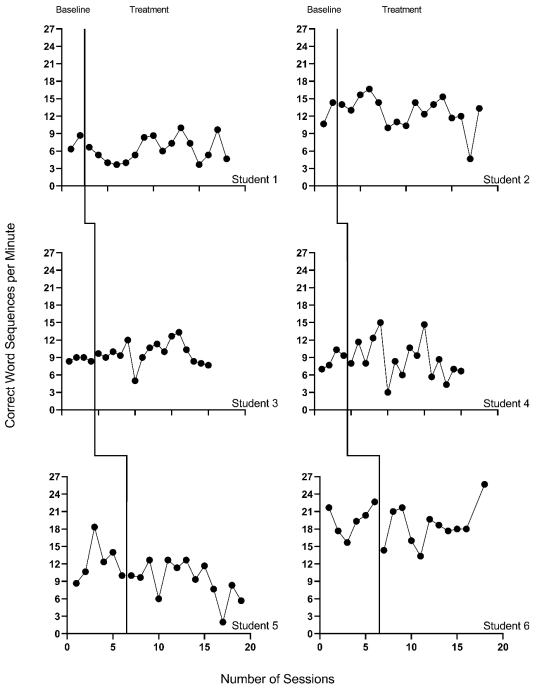


Figure 8.4: Effect of Intervention on Correct Word Sequences (Lewandowski, 2011)

lines shows that the standard deviation (SD) of these deviations was 1.35 in baseline and 1.53 in intervention. Although the within-case variability is similar across phases, it appears to differ across studies, with more variability in the latter two studies that examined younger children. Adding confidence to this conclusion, descriptive statistics for the Walker et al. (**?**, -**?**) studies indicate deviations from the trend lines of 0.46 and 0.50 SDs respectively, whereas the SDs from the trend lines for the studies by **?** and **?** are 2.32 and 2.46 respectively. This obvious violation of the homogeneity assumption appears to be at least partially responsible for the high kurtosis values seen in our examination of normality. Examining normality within each study, the maximum kurtosis value is only 1.5 (for **?**). While it is feasible to estimate separate covariance matrices for the studies with older and younger participants (**??**), these more complex heterogeneous variance models are not accessible via the MultiSCED app. In addition, the average treatment effect size estimates (changes in level and slope) and related inferences are relatively robust to this sort of violation of the homogeneity assumption (**?**), so proceeding with the simpler homogeneous variance models seems reasonable.

In summary, our visual analyses of the primary study data resulted in several conflicts between our expectations, the study results, and the assumptions associated with a multilevel model with trends. First, the inconsistencies raise questions about whether we need to include baseline trends in our model specification. Second, there are multiple violations to the assumptions of the basic LMM for SCD studies, like evidence of non-normality (particularly in the baseline phases) and heterogeneity of variance across cases. The model we use in this chapter is surely simpler than the process that actually generated the data under analysis. With a relatively small number of studies, one may want to further simplify their model specification by removing baseline trends. However, for purposes of illustrating multilevel modeling of trends using the MultiSCED app, we proceed with the more elaborate model specification. We also note that the multilevel model estimates of the across-case average treatment effects (i.e., change in level, change in trend), as well as the inferences associated with those effects (e.g., confidence intervals, significance tests), are relatively robust to violations of the normality and homogeneity assumptions.

8.2 Details of the Multilevel Model with Trends

The multilevel data structure has repeated observations that are nested within cases and cases that are nested within studies, and thus there is a model for the variation of observations within a case (level-1), a model for the variation between cases within a study (level-2), and a model for variation between studies (level-3). The formal specification of the within-case (level-1) model with trends in both the baseline phase and the treatment phase is:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} Tx_{ijk} + \beta_{2jk} Time_{ijk} + \beta_{3jk} Tx_{ijk} \times Time_{ijk} + e_{ijk}, \quad (8.1)$$

where Y_{ijk} is the outcome variable Y at measurement occasion i for case j of study k. Tx_{ijk} is dummy coded with a value of 0 for baseline phase observations and a value of 1 for treatment phase observations. $Time_{ijk}$ is the value of the time variable on occasion i for case j of study k. The error term, e_{ijk}, is time-, case-, and study-specific, and assumed independently and normally distributed with variance σ_e^2. Outside of the MultiSCED app, it is feasible for researchers to estimate a different (co)variance structure for e_{ijk}, such as first-order autoregressive or heterogeneous across phases or cases (**???**).

Second, we must decide how to center the $Time_{ijk}$ variable (i.e., at what point in time is the value of $Time_{ijk}$ equal to zero). Our choice of the focal time to index the effect has consequences for how we interpret the regression coefficients of the model. The most common choice is to code $Time_{ijk}$ such that zero (0) corresponds to the first treatment phase observation for case j for study k. Because this centering choice is appropriate for this example, and because the MultiSCED app will do this centering of time for us, we do not need to enter the centered time variable into our dataset. If we wanted to enter this centered time variable during our data organization phase, we would code the treatment phase observations sequentially starting from 0 (i.e., 1 for the second treatment phase observation, 2 for the third, and so on). For baseline phase observations, coding begins by working back from 0. In other words, we would code the last baseline phase observation as -1, the second-to-last baseline observation as -2, and so on.

Then, after deciding to center $Time_{ijk}$ such that 0 corresponds to the first treatment observation, we can interpret β_{0jk} as the expected baseline value for case j of study k if the baseline is extended to the first treatment observation. β_{1jk} is then interpreted as the immediate effect of the intervention (i.e., the difference between the treatment phase trend line and the extended baseline trend line at time zero). Figure **??** visually depicts these regression coefficients. The baseline slope is β_{2jk}, and the change in slope between baseline and treatment phases is β_{3jk}.

The between-case (level-2) model is:

$$\beta_{0jk} = \gamma_{00k} + u_{0jk} \tag{8.2}$$

$$\beta_{1jk} = \gamma_{10k} + u_{1jk} \tag{8.3}$$

$$\beta_{2jk} = \gamma_{20k} + u_{2jk} \tag{8.4}$$

$$\beta_{3jk} = \gamma_{30k} + u_{3jk} \tag{8.5}$$

where γ_{00k} is the expected across-case average baseline value at the time of the first treatment observation of study k, γ_{10k} is the across-case average immediate treatment effect of study k, γ_{20k} is the across-case average baseline slope, and γ_{30k} is the across-case average change in slope between baseline and treatment phases. The case-specific errors (u_{0jk}, u_{1jk}, u_{2jk}, and u_{3jk}) correspond to the regression coefficient deviations of individual case j of study k from the

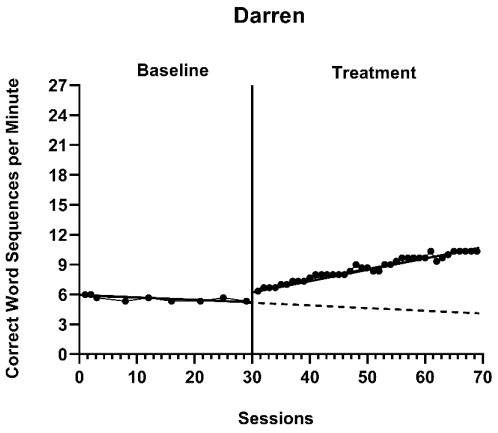


Figure 8.5: Illustration of Treatment Effect for Darren (Walker et al., 2005)

across-case average values for study k. These case-specific errors are assumed multivariate normal with covariance $\Sigma_u = \begin{bmatrix} \sigma^2_{u_0} & & & \\ \sigma_{u_1 u_0} & \sigma^2_{u_1} & & \\ \sigma_{u_2 u_0} & \sigma_{u_2 u_1} & \sigma^2_{u_2} & \\ \sigma_{u_3 u_0} & \sigma_{u_3 u_1} & \sigma_{u_3 u_2} & \sigma^2_{u_3} \end{bmatrix}$. In some applications, an unstructured matrix will be difficult to estimate with the limited number of cases available. Common alternatives include estimating a diagonal covariance structure or removing errors from Equations (**??**) and (**??**) so that some of the slopes do not vary randomly across cases.

The between-study (level-3) model is:

$$\gamma_{00k} = \theta_{000} + v_{00k} \tag{8.6}$$

$$\gamma_{10k} = \theta_{100} + v_{10k} \tag{8.7}$$

$$\gamma_{20k} = \theta_{200} + v_{20k} \tag{8.8}$$

$$\gamma_{30k} = \theta_{300} + v_{30k} \tag{8.9}$$

where θ_{000} is the across-study average expected baseline value at the time of the first treatment observation, θ_{100} is the across-study average immediate treatment effect, θ_{200} is the-across study average baseline slope, and θ_{300} is the across-study average change in slope between baseline and treatment phases. The study-specific errors (v_{00k}, v_{10k}, v_{20k}, and v_{30k}) correspond to the deviations of the across-case average coefficient values of study k from the overall values averaged across all studies. These study-specific errors are assumed multivariate normal with covariance $\Sigma_v = \begin{bmatrix} \sigma^2_{v_0} & & & \\ \sigma_{v_1 v_0} & \sigma^2_{v_1} & & \\ \sigma_{v_2 v_0} & \sigma_{v_2 v_1} & \sigma^2_{v_2} & \\ \sigma_{v_3 v_0} & \sigma_{v_3 v_1} & \sigma_{v_3 v_2} & \sigma^2_{v_3} \end{bmatrix}$. In many

applications, an unstructured matrix will be difficult to estimate with the limited number of studies available. In our example, the data include only four independent studies, which would not be considered adequate for estimating an unstructured covariance matrix. Common alternatives include estimating a diagonal covariance structure or removing errors from some of Equations (**??**) and (**??**) (and potentially (**??**)) so these regression coefficients do not vary randomly across the studies.

For the purposes of the methods guide, we illustrate in detail the estimation of the multilevel model for the writing outcome of correct word sequences per minute using the MultiSCED app (**?**). This app limits users to LMMs and the covariance structures described in this section. To allow for the exploration of how model specification details influence estimates and inferences, we contrast the full model results (from Equations (**??**) to (**??**)) with models having simpler covariance structures (e.g., models without as many coefficients varying randomly across studies). See Chapter 7 Appendix for examples of SAS and R code used to estimate models with diagonal covariance structures.

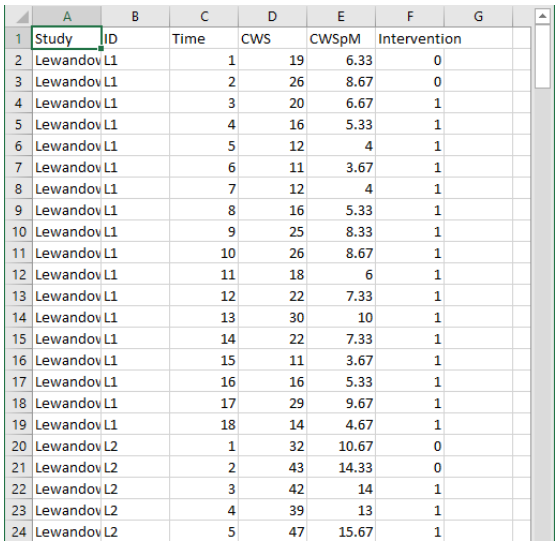
8.3 Estimating the Multilevel Model for the Included Writing Intervention Studies

Using the web-based MultiSCED calculator for the multilevel modeling of single-case data, we can estimate the across-case average treatment effect and variance components describing the degree of heterogeneity in treatment effects. The MultiSCED app is available at http://34.251.13.245/MultiSCED/. To use this app, researchers must save their dataset as a text file (.txt) or a comma-delimited file (.csv). We recommend that researchers first enter their data in Excel, and then save the Excel dataset as a tab-delimited text file.

Figure **??** shows a screenshot of a portion of the original Excel spreadsheet containing the data extracted from the set of four writing intervention studies, with correct word sequence writing outcomes converted as needed to compare rates on a common per-minute scale. This spreadsheet contains six columns of data corresponding to the following six variables: (a) study identifier (i.e., Study), (b) case identifier (i.e., ID), (c) point in time for which the outcome was recorded (i.e., Time), (d) frequency outcome (number of correct word sequences written; i.e., CWS), (e) rate outcome (rate of correct word sequences written per minute, CWSpM), and (f) phase identifier (i.e., intervention). Note that the time variable is numbered sequentially from first observation to last observation and has not yet been centered relative to the start of the treatment phase.

To use the MultiSCED app, the session and outcome variables must be numerical values. However, the variables for study, case, and phase identifiers may be entered as alphanumeric (word labels) or numeric values. Because the app requires it, we organized the data using a long format with one row per time point per case per study, where we input the value of the outcome at each time

point for a specific case below the cell containing the value at the previous time point. For each study (e.g., Lewandowski in Figure **??**), we first arranged cases by baseline length. In order of shortest to longest baseline length, we entered all observations for one case at a time, in order of time. After we entered all data for the case with the shortest baseline, we entered data for the case with the second-shortest baseline, and so on. After entering the data for all cases within a particular study, we repeated the process with the remaining studies; we entered the data from each successive study below the data from the previous study.



	A	B	C	D	E	F	G
1	Study	ID	Time	CWS	CWSpM	Intervention	
2	Lewandov	L1	1	19	6.33	0	
3	Lewandov	L1	2	26	8.67	0	
4	Lewandov	L1	3	20	6.67	1	
5	Lewandov	L1	4	16	5.33	1	
6	Lewandov	L1	5	12	4	1	
7	Lewandov	L1	6	11	3.67	1	
8	Lewandov	L1	7	12	4	1	
9	Lewandov	L1	8	16	5.33	1	
10	Lewandov	L1	9	25	8.33	1	
11	Lewandov	L1	10	26	8.67	1	
12	Lewandov	L1	11	18	6	1	
13	Lewandov	L1	12	22	7.33	1	
14	Lewandov	L1	13	30	10	1	
15	Lewandov	L1	14	22	7.33	1	
16	Lewandov	L1	15	11	3.67	1	
17	Lewandov	L1	16	16	5.33	1	
18	Lewandov	L1	17	29	9.67	1	
19	Lewandov	L1	18	14	4.67	1	
20	Lewandov	L2	1	32	10.67	0	
21	Lewandov	L2	2	43	14.33	0	
22	Lewandov	L2	3	42	14	1	
23	Lewandov	L2	4	39	13	1	
24	Lewandov	L2	5	47	15.67	1	

Figure 8.6: Screenshot of Extracted Study Data Within the Original Excel Spreadsheet

After organizing the data within the Excel spreadsheet, we then save the Excel spreadsheet as a tab delimited text file (.txt). We present a screenshot of a portion of the resulting text file in Figure **??**. To prepare the data set for the MultiSCED app, one can follow the steps below.

For Windows:

1. From the *File* menu, select *Save As*.
2. Select *Browse* to select a folder where the spreadsheet will be saved.
3. To save the spreadsheet, enter a file name in the corresponding field.
4. Then use the *Save as* Type menu to select the option to save the file type as Text (Tab delimited) (*.txt).
5. Finally, click *Save*.

If there are multiple worksheets within the original Excel spreadsheet file, a pop-up message will appear to request permission to save only the active sheet. If this occurs, select OK to save the data file. We strongly encourage researchers

to open the newly created tab delimited text file to ensure that the data appear as expected.

For MacOS:

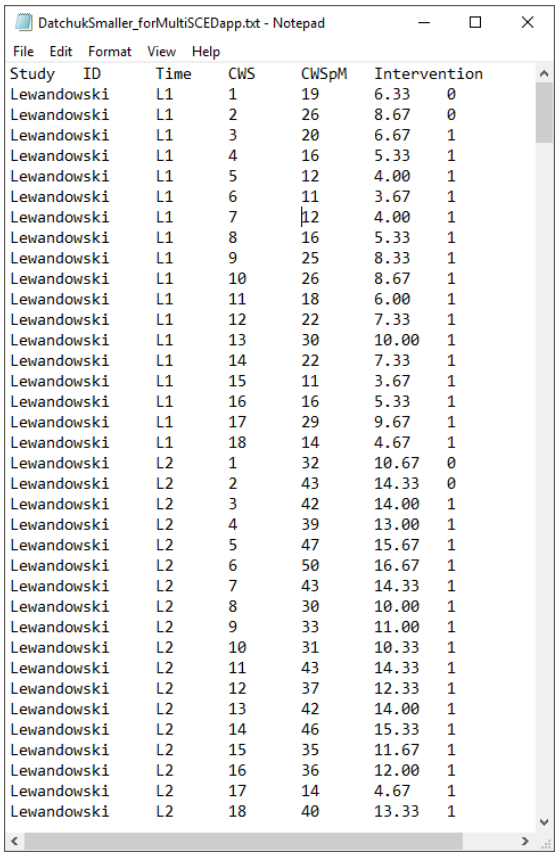
1. From the *File* menu, select *Save As*.
2. Enter a file name in the *Save As* field.
3. Select a location from the dropdown menu, *Where*, to save your spreadsheet.
4. Then use the *File Format* menu to select the option to save your file type as Tab delimited Text (.txt).
5. Finally, click *Save*.

It is important to note here that an error message will appear if attempting to save a spreadsheet with multiple worksheets. A pop-up will inform you that the workbook cannot be saved in the selected file format. If encountering this error, researchers should open a newly created file with one worksheet for the data needed in their analysis.

After converting and saving the dataset as a tab delimited text file, we are prepared to use the MultiSCED app (http://34.251.13.245/MultiSCED/). Figure 8.8 is a screenshot of the Home page for the MultiSCED app. At the top of the page, one will see that the *Home* link on the blue menu bar has a darker background than the other website page links. To access the app, we must move from the *Home* page to the *Input* page by clicking on the Input link in the website navigation menu bar.

A screenshot of the *Input* page is shown in Figure **??**. After navigating to the *Input* page, we find a sidebar menu with three additional links: *Data file*, *Variables*, and *Data summary*. We first select the sidebar option to choose the *Data file* section. When it is highlighted in dark blue (as in Figure **??**), we can upload our dataset (.txt or .csv) by clicking the *Browse* button, selecting the folder where the data file is saved, and clicking *Open*. When we did this, two additional menus appeared in the *Upload* box beneath the *Browse* button: *Separator character* and *Decimal character*. The default option for the *Separator character* is tab. We keep the default setting because we already saved the dataset as a tab delimited text file. However, if your file is saved as a .csv file, you can select *comma* under the *Separator character* dropdown menu. Regarding the *Decimal character* menu, the option for decimal representation defaults to dot, which is the standard choice in the U.S. However, for researchers located outside of the U.S. that use commas to mark decimals (e.g., EU countries), the setting will need to be changed from the default of dot to the comma option. We note here that researchers can use sample data to practice using the app. To access sample data, researchers should upload any file (which will not be used) and then click on the *Use testdata* checkbox. Sample data information will be automatically input into the *Variables* section which follows. After completing these actions, we transition to the *Variables* section on the left side of the screen.

As shown in Figure **??** of the MultiSCED app *Input* page, the *Variables* section



DatchukSmaller_forMultiSCEDapp.txt - Notepad

File Edit Format View Help

Study	ID	Time	CWS	CWSpM	Intervention
Lewandowski	L1	1	19	6.33	0
Lewandowski	L1	2	26	8.67	0
Lewandowski	L1	3	20	6.67	1
Lewandowski	L1	4	16	5.33	1
Lewandowski	L1	5	12	4.00	1
Lewandowski	L1	6	11	3.67	1
Lewandowski	L1	7	12	4.00	1
Lewandowski	L1	8	16	5.33	1
Lewandowski	L1	9	25	8.33	1
Lewandowski	L1	10	26	8.67	1
Lewandowski	L1	11	18	6.00	1
Lewandowski	L1	12	22	7.33	1
Lewandowski	L1	13	30	10.00	1
Lewandowski	L1	14	22	7.33	1
Lewandowski	L1	15	11	3.67	1
Lewandowski	L1	16	16	5.33	1
Lewandowski	L1	17	29	9.67	1
Lewandowski	L1	18	14	4.67	1
Lewandowski	L2	1	32	10.67	0
Lewandowski	L2	2	43	14.33	0
Lewandowski	L2	3	42	14.00	1
Lewandowski	L2	4	39	13.00	1
Lewandowski	L2	5	47	15.67	1
Lewandowski	L2	6	50	16.67	1
Lewandowski	L2	7	43	14.33	1
Lewandowski	L2	8	30	10.00	1
Lewandowski	L2	9	33	11.00	1
Lewandowski	L2	10	31	10.33	1
Lewandowski	L2	11	43	14.33	1
Lewandowski	L2	12	37	12.33	1
Lewandowski	L2	13	42	14.00	1
Lewandowski	L2	14	46	15.33	1
Lewandowski	L2	15	35	11.67	1
Lewandowski	L2	16	36	12.00	1
Lewandowski	L2	17	14	4.67	1
Lewandowski	L2	18	40	13.33	1

Figure 8.7: Screenshot of Saved Tab Delimited Text File

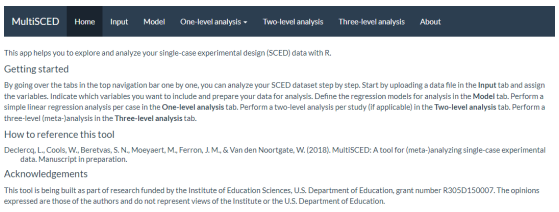


Figure 8.8: Screenshot of the Website Navigation Menu and Home Page of the MultiSCED App

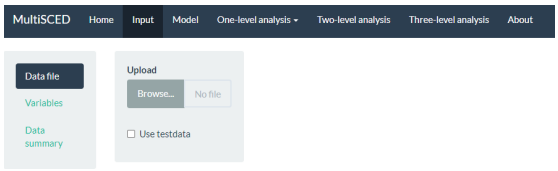


Figure 8.9: Screenshot of the Input Page of the MultiSCED App

appears as the second option in the left sidebar under *Data file*. When highlighted in dark blue, the *Variables* page has two primary sections: *Base variables* and *Moderator variables*. Using the drop-down menus in the *Base variables* section, we select our outcome variable (i.e., CWSpM) from the *Response* menu, case identifier from the *Case* menu, study identifier from the *Study* menu, and phase identifier from the *Phase* menu. We also use the *Phase control group* drop-down menu to define our dummy coded values for baseline observations and specify our time variable from the options in the *Time* drop-down menu. Because we are modeling trends, we also check the box next to *Center time variable*. The app will center time for each case so that zero corresponds to the first intervention observation allowing the coefficient for *Intervention* to index the immediate effect of the intervention. Figure **??** shows the *Variables* tab, along with the appropriate menu selections for this model.

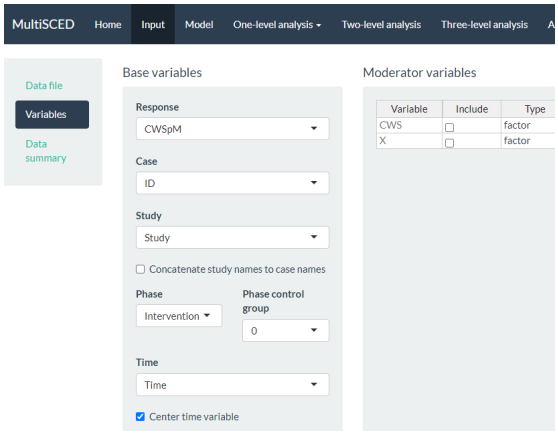


Figure 8.10: Screenshot of the Variables Section Within the Input Page of the MultiSCED App

After specifying our variables, we navigate to the *Data summary* section of the app. We show a screenshot for the *Data summary* tab in Figure **??**. At the top of the page, the app summarizes the data for each of the three levels in our model: the number of studies included (i.e., 4), the number of cases nested within these four studies (i.e., 15 cases), and the total number of observations nested within the cases that are nested within the studies (i.e., 442). These

values all correspond with what we know about our dataset.

Under the column headers *Studies* and *Cases*, the app defaults to include all studies and cases in the analysis, as indicated by the check marks (see Figure **??**). This is appropriate for our application, because for this methods guide, we intend to illustrate the synthesis of data across all cases and studies using a three-level analysis. As an aside, the MultiSCED app allows for statistical models for a single case or two-level models of the cases nested within a single study. Thus, it provides the option to click *(De)select all* to allow researchers to only include the data from studies and cases on which they want to focus their analysis.

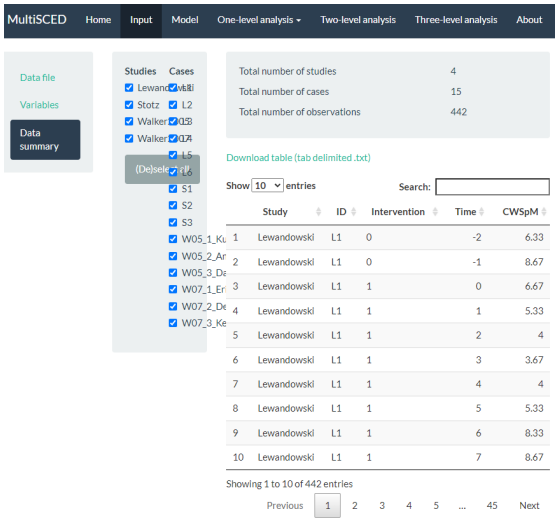


Figure 8.11: Screenshot of the Data summary Section within the Input Page of the MultiSCED App

After examining the *Data summary*, we need to specify our multilevel model. To do so, we navigate to the *Model* page (see Figure **??**). There are five sections to the *Model* page: *Fixed effects* and *Random effects* on the left, and *One-level model*, *Two-level model*, and *Three-level model* on the right side of the screen. In the *Fixed effects* section, we find a list of four *Base variables* that we can include in our model. To specify a model without trends, we leave the *Fixed effects* section untouched, as the app includes *intercept* and *Intervention* by default (as noted by the checkmark next to each). Because we want to estimate a baseline slope and a change in slope that occurs with intervention, we manually check the boxes next to *Time* and *Intervention X Time*. When we check these boxes, the app automatically adds more options [i.e., $(Time)^2$ and *Intervention X* $(Time)^2$]. However, because we anticipated linear trends and ruled out non-linear trends through visual analysis, we do not include (check) either of these additional boxes.

Under the *Random effects* section, the app leaves all options unchecked. Therefore, we manually select the effects necessary to match our three-level model specifications in Equations (**??**) to (**??**). Allowing all regression coefficients to vary randomly across cases, we check all boxes under the *Case level* sub header. Then, we check each of the boxes under the *Study level* sub header to let the across-case study averages vary randomly for each of the coefficients across studies. As we opt to include these *Case level* and *Study level* variables, the app automatically updates our model specification on the page. In Figure **??**, we present a screenshot of the resulting *Model* page, where the three-level model shown (bottom right) matches our model defined by Equations (**??**) to (**??**). Under the equations, the app provides R syntax for fitting the model.

At this point, we have specified a very complex model and do not anticipate good estimation of the (co)variance matrices, particularly for the study-level errors. Our model includes four study-level random effects, but the data include only four unique studies. Although our process thus far (e.g., including random effects on all four terms, thereby letting all coefficients vary freely) aligns with what has typically been done in prior methodological studies focused on evaluating the appropriateness of three-level models for single-case data, much of this research uses diagonal as opposed to unstructured covariance matrices (e.g., **???**). If meta-analysts encounter estimation problems that suggest some of the variances may be 0 or correlations are at the boundaries of 1 or -1, they will need to take further action, such as simplifying the variance structure (i.e., removing error terms with little to no variability) and changing the covariance structure specification.

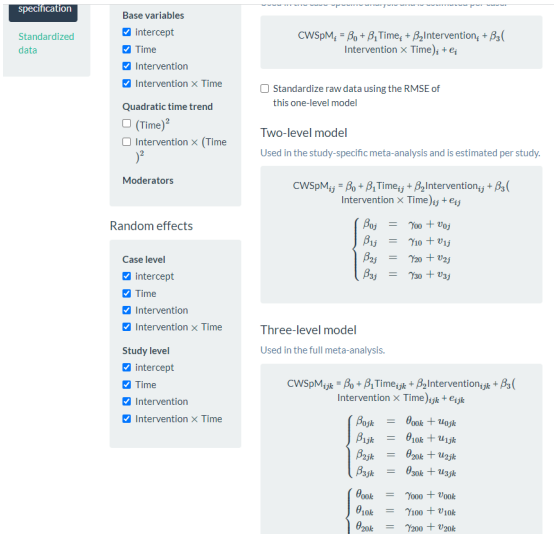


Figure 8.12: Screenshot of the Model Page of the MultiSCED App

After specifying the model, we use the header at the top of the website to

navigate to our multilevel model results by clicking the *Three-level analysis* link. In transitioning to the *Three-level analysis* page, we may immediately note a blank page (i.e., no results) or view an error message. However, we usually encounter a small progress tracking bar in the far-right bottom of the page that conveys that estimation is in progress. It takes the MultiSCED app a few seconds to complete our model estimation and present the multilevel model results (see Figure **??**).

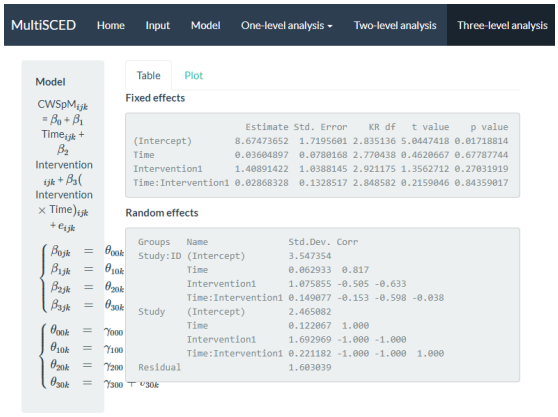


Figure 8.13: Screenshot of the Three-level analysis Page of the MultiSCED App

The multilevel model estimation results indicate that if baseline continued, the estimated across-case and across-study average correct word sequences written is 8.67 per minute at the time of the first treatment observation. The immediate treatment effect estimate was an increase of 1.41 correct writing sequences (CWS) per minute (i.e., the difference between the across-case and across-study average treatment phase trajectory and projected baseline trajectory at the time of the first treatment observation). The standard error (SE) of the estimate was 1.04, indicating substantial uncertainty in the size of the average immediate treatment effect. However, this was expected given the small number of studies we used to illustrate the MultiSCED app procedures, as more studies would have resulted in a more precise estimate (e.g., see **?**). The estimated across-case and across-study average baseline slope was 0.036 (SE = 0.078), and estimated change in slope with intervention was 0.029 CWS per minute per intervention session (SE = 0.133). Although the small, positive slope estimate in baseline and the increase in slope with intervention align with our expectations, the sampling error in both estimates is too large to conclude that either of the corresponding population parameters differ from zero.

We next examine the results under the *Random effects* sub header. The MultiSCED app reports the *Random effects* variance components as standard deviations (SDs). Unlike fixed effects, which have been shown to be well estimated even with the small sample sizes typical of single-case research, variance compo-

nents are typically not well estimated. Rather, we expect they were estimated with considerable bias due to the limited number of studies and cases (**??**). In addition, our examination of the data prior to modeling suggested that we likely oversimplified the model by specifying an independent and homogeneous variance structure, which leads to even more bias in the estimated variance components (**?**). Consequently, although we report these parameter estimates from our model as a conceptual illustration for the purpose of this guide, we do not suggest that users of the app give serious weight to the interpretation of the *Random effects* variance components unless they have more cases and studies. Per the results reported under the *Random effects* sub header, the estimated SD of the observations around the case-specific trend lines is 1.60 (i.e., *Residual*).

Next, examining the between-case variability within studies (i.e., *Study: ID*), we see that the estimated SD of the case-specific intercepts is 3.55, the estimated case-specific immediate treatment effects SD is 1.08, the estimated SD of the case-specific baseline slopes is 0.063, and the estimated SD of the case-specific changes in slope with intervention is 0.15. Also note that the correlations among these random effects are plausible correlation values. In contrast, when we examine the correlation matrix for the study random effects (i.e., *Study*), we see estimated correlations at the boundaries (1 and -1) for all random effects at the study level. This signals that the covariance structure is too complex at the study level for the app to estimate correlations between our included four variables across only four studies. Thus, we will re-estimate the model while simplifying the number of study-level random effects and look to see the impact of specifying a simpler covariance structure on the estimates of the treatment effects.

To simplify the covariance structure for the study-level random effects, we go back to the *Model* page and deselect some of the *Random effects* options listed under the *Study level subheading*. As shown in Figure **??**, we uncheck *Time*, *Intervention*, and *Intervention X Time*. Although conceptually any of these coefficients could vary from one study to the next, we only have four studies in our example, which is not sufficient to estimate so many variance components.

We show the results from this simpler re-specified model in Figure **??**. With this model, the across-case and across-study average immediate effect is 1.03 (SE = 0.64) and the estimated effect of the intervention on the slope is -0.03 (SE = 0.06). The point estimates for the effects have changed, but the conclusion of a potential non-effect (i.e., the population effect parameters may be zero) remains the same. Although we eliminated the correlations of 1 and -1 in the study-level random effect estimates, the estimated variance for the study-level intercepts is 0, suggesting insufficient between-study variability necessary to obtain a non-zero estimate.

After comparing both models, we consider the degree to which we may obtain better treatment effect estimates from one model or another. For example, could we achieve better estimates if using a model that allows all effects to vary randomly across cases and studies, or a model with a covariance structure that

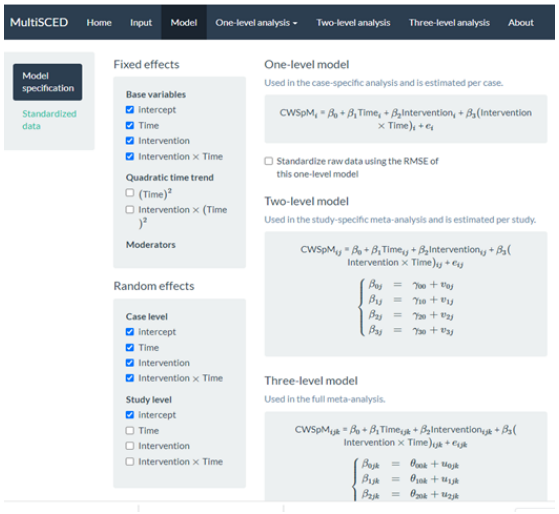


Figure 8.14: Screenshot of the Re-Specified Model with Fewer Study-Level Random Effects (Model Page)

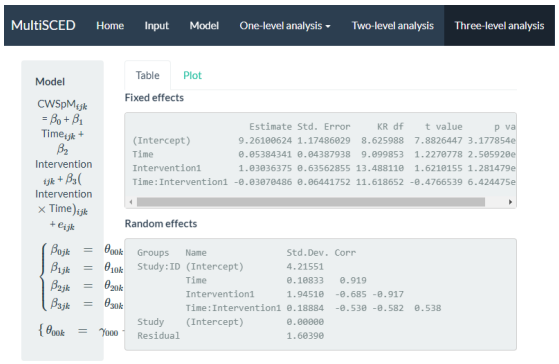


Figure 8.15: Screenshot of the Three-level analysis Page for the Re-Specified Model with Fewer Study Level Random Effects

is easier to estimate? Similarly, is it better to model heterogeneity in the error variance between cases and phases, and is it best to model autocorrelation among the case-level errors? Researchers have shown that the misspecification of the covariance structure as either too simple or too complex at any of the three levels introduces little to no bias in the average treatment effect estimates, leading to relatively accurate inferences about the average treatment effect (**??????**). However, this research examines a limited range of conditions. We hope that future research will serve to further clarify the optimal approach for estimating three-level LMMs for SCD study data. In addition, SCD studies will sometimes have outcomes that present more severe violations of the normality and homogeneity assumptions than we encountered here. Useful directions for future research include: (a) determining at what point the violations are severe enough and sample sizes large enough to prefer the use of GLMMs, (b) providing further guidance on the most appropriate ways to specify and estimate GLMMs (**?**), and (c) producing easy-to-use software to make GLMMs more readily accessible to SCD researchers.

Chapter 9

Introduction to Case-Specific Effect Sizes

This chapter provides background on case-specific effect sizes and describes when this approach is useful in the synthesis of single-case research. We discuss different case-specific effect size metrics, their underlying assumptions, and when it is appropriate to use them. We conclude this chapter by providing a set of decision rules for meta-analysts to use when selecting among the various available case-specific effect sizes.

9.1 Background

Case-specific effect sizes in single-case design (SCD) research provide an effect estimate for each participant on a common metric. In the context of a research synthesis that includes multiple studies, case-specific effect sizes are useful for summarizing the overall average effectiveness of an intervention, investigating the degree of heterogeneity of effects across participants, and identifying for whom and under what conditions the intervention is most effective. In meta-analyses of SCD studies, researchers can use several different metrics to summarize the intervention effect for a given case. One approach quantifies the effect based on the degree of nonoverlap between observations in the treatment phase and baseline phase (**??????**). A second approach standardizes a summary of the change between baseline and treatment phases using standardized mean differences (**??**) or standardized regression coefficients (**??**). A third approach compares the baseline and treatment levels through response ratios or percentage change indices (**?**). Finally, a fourth approach indexes the effect relative to goal attainment by finding the percent of zero data (**?**) or the percentage of goal obtained (**?**). Researchers can choose among these metrics to meet the needs of their single-case synthesis context, understanding that no single approach is

universally appropriate for all SCDs.

Each of the case-specific effect size estimation methods have limitations that can hinder their utility and, if not purposefully selected, lead to effect estimates misaligned with visual analysis. Consequently, researchers have developed methods for choosing among the different effect quantifications based on study purpose(s) and characteristics of the data (**?**). Further, researchers can conduct sensitivity analyses and report results from multiple indices to allow examination of the similarity in conclusions across indices (**?**).

In this chapter, we detail the available case-specific effect size options and their underlying assumptions. We provide guidance on selecting one or more metric types for a synthesis, as well as recommendations for selecting among the specific effect indices within a metric type.

9.2 When to Use Case-Specific Effect Sizes

When choosing an effect size for SCD study data, researchers should consider their aims or reason for using one or multiple effect sizes. There are two common motivations for using case-specific effect sizes to synthesize SCD effects. First, in some contexts, researchers want to quantify the treatment effect for each individual participant to summarize the average effect and examine variation in treatment effectiveness across participants, along with predictive factors that may help to identify for whom the intervention is most effective. When one purpose of a synthesis is to examine individual-level effects, case-specific effect sizes are more aligned with such a goal than are study-level effect sizes (i.e., design-comparable effect sizes that average the effect across participants). Second, in contexts where a variety of different approaches are used to measure outcomes across the cases and studies to be synthesized, researchers need to find a common metric with which to index the effects. If the same outcome is available for all cases for all studies, researchers could consider a multilevel model. However, multilevel modeling is not viable when outcomes are so disparate across cases that transformation of the raw data to a common scale is not feasible. Consider, for example, a situation where researchers are interested in examining individual-level effects of an intervention targeting the reduction of disruptive behavior during home routines. One included study measures the dependent variable as elapsed time from the beginning of a routine until the first disruptive behavior (i.e., latency). Another included study reports the effects of the intervention on reducing the number of disruptive behavior events during the routine (count per observation). In circumstances such as these, where researchers are interested in studying the variation in effects across individuals and the outcomes differ substantially across these individuals, case-specific effect sizes are preferable because they provide multiple options for indexing effects using a common metric.

9.3 Assumptions and Limitations of Case-Specific Effect Sizes

When computing case-specific effect sizes, assumptions and limitations encountered depend on both metric type used (nonoverlap, standardized, response-ratios, or goal attainment) and the specific index used within that type (e.g., NAP versus Tau-U). Thus, we divide this chapter into subsections for each metric type and detail the assumptions and limitations associated with selected indices of each type. We conclude with a discussion of approaches for synthesizing the effect size estimates. Researchers can calculate most of the illustrated effect size metrics using a web application called the Single-Case Effect Size Calculator [https://jepusto.shinyapps.io/SCD-effect-sizes//; **?**] or using the SingleCaseES R package (**?**). We provide a detailed demonstration of these tools in Chapter 10.

9.4 Assumptions and Limitations of Nonoverlap Indices

One of the first nonoverlap indices developed, percent of nonoverlapping data (PND; **?**), computes the percentage of treatment phase observations that do not overlap with baseline phase observations. Unfortunately, the index has several technical drawbacks: (a) it is sensitive to outliers, (b) it has no known SE, and (c) its expected value depends on the number of observations in the baseline phase (**???**). Two newer nonoverlap indices largely address these technical problems and are more widely recommended for use: nonoverlap of all pairs (NAP; **?**) and Tau (**?**).

To compute NAP, each baseline observation is compared to each treatment phase observation. For each of these comparisons, researchers must determine whether the treatment observation is more favorable (indicating a positive result), less favorable (indicating a negative result), or identical to the baseline phase observation (indicating a null result). After obtaining a rating for each paired baseline and treatment phase observation, comparisons are summed by their category to give the total number of positive comparisons (P), negative comparisons (N), and tied comparisons (T). With these totals, NAP is the percentage of the comparisons that are positive, $NAP = \frac{P+.5\ T}{P+N+T} \times 100$. The maximum value of NAP is 100, a value that indicates all comparisons had a positive result, suggesting a positive treatment effect. If half the comparisons were positive and half were negative, the value of NAP would be 50, suggesting a null treatment effect. The minimum value of NAP is 0, when all comparisons are negative, indicating a detrimental or iatrogenic intervention effect. However, NAP is not restricted to this scale, and outcomes can be reported as a proportion using a scale of 0-1[1].

[1]The Single Case Effect Size Calculator (**?**) app used in Chapter 10 to model case-specific

Tau is closely related to NAP and is also based on the comparison of each baseline observation to each treatment observation. However, unlike the NAP scale of 0-100 usually reported in studies, the Tau scale is -1 to 1, where 1 indicates a maximal positive effect, 0 indicates no effect, and -1 indicates a maximal negative effect. Tau can be computed from the sum of the positive, negative, and tied comparisons or can be computed directly from NAP; $Tau = \frac{P-N}{P+N+T} = (2 \times \frac{NAP}{100}) - 1$. Because of the similarities between NAP and Tau, neither is statistically superior over the other. Where one can use NAP, they one could also use Tau, and vice versa. Thus, choice of NAP or Tau is a matter of the researcher's preferred scale (i.e., an index that runs from 0 to 100, or an index that runs from -1 to 1). NAP and Tau are preferable to PND for several reasons. Outliers have less influence on NAP and Tau estimates, resulting in more stable sampling distributions (**?**). Unlike PND, the expected values of NAP and Tau are not dependent on the baseline length (**?**). Furthermore, one can compute SEs for both NAP and Tau given the assumptions that the observations are mutually independent and homogeneously distributed within each condition. However, it is reasonable to believe that these assumptions may be violated, at least to some extent. Single-case data can be autocorrelated (**??**), violating the assumption of mutual independence. Single-case data series can also involve time trends, leading to heterogeneous distributions within each phase. Yet, having SEs under relatively restrictive assumptions is preferable to not having SEs at all. Moreover, techniques such as robust variance estimation can be deployed in the synthesis of the effect sizes that can tolerate inaccuracies in estimation of the SEs for individual effect size estimates (**?**).

A limitation of both NAP and Tau, as well as for the other nonoverlap indices, is the presence of a maximum value that restricts the sensitivity of this index to large effects. Consider the multiple baseline study by **?** shown in Figure **??**. In this study, the researchers tested a culturally focused training to increase staff use of behavior-specific praise with students. Visual analysis suggests that the treatment effect varies across dyads, with Dyad 4's effect being the largest and the remaining dyads having variations in variability, level, and overall magnitude of effect (i.e., differences in level across phases). However, the value of NAP for all three cases is 100 and the value of Tau is 1, because treatment phase observations do not overlap with any baseline phase observation. The ceiling effect for the nonoverlap indices leads to less variation in the effect size estimates than there would be in an examination of true effects.

When choosing to use NAP or Tau, one implicitly assumes that these metrics will reflect meaningful variation in the magnitude of the effects. If the synthesis purpose is simply to confirm or rule out the presence of an effect, it may not be critical to distinguish between large and very large effects. However, if the research goal is to estimate the degree to which the size of the effect varies with some participant characteristic, NAP and Tau insensitivities to effect size at the high end of the scale can be problematic. Ceiling effects like those in

effect size estimation methods reports NAP on a 0-1 scale.

the **?** example can limit meta-analytic efforts to examine the degree to which participant characteristics explain intervention effectiveness.

Another limitation of interpreting both NAP and Tau results as effect sizes is that one must assume there is no baseline trend and that the difference between the baseline and treatment observations results from a functional relation between the intervention and outcome, rather than some extraneous variable. Figure **??** shows the data from Denny, one of the participants in the **?** writing intervention study. Denny's graph portrays a trend in baseline observations that suggests that some other factor has an influence on the behavior (e.g., maturation). While the value of NAP for this illustration is 80 (SE = 10; $CI_{95} = 56,\ 92$) and the value of Tau is 0.60 (SE = 0.19; $CI_{95} = 0.13,\ 0.84$), these positive values cannot be attributed to the treatment alone. In some synthesis contexts, studies with unstable baselines may not meet the quality inclusion criteria; however, ultimately it is up to the researchers to determine the inclusion criterion based on their specific research questions. If included, researchers should consider alternative effect sizes that have baseline trend adjustments.

Researchers' desire for a metric that can adjust for baseline trends led to the development of Tau-U (**?**). Tau-U adds a correction for monotonic trend to the numerator of the Tau computation. Unfortunately, this correction introduces multiple problems that make it challenging to recommend Tau-U for use in a research synthesis. As described in detail by **?**, problems with Tau-U include: (a) estimates are not bounded between -1 and 1, (b) adjustments cannot be meaningfully represented on a graph, (c) the size of the trend adjustment depends on the number of baseline observations and thus the expected value of Tau-U depends on phase lengths, and (c) the trend adjustment changes the sampling distribution of the statistic so that the Tau SEs are not appropriate for Tau-U. Furthermore, separate SEs for Tau-U have not yet been derived.

To correct some of the technical problems with Tau-U, **?** developed the baseline-corrected Tau (**?**) that attempts to resolve the first three of the aforementioned problems. While baseline-corrected Tau (Tau_{BC}) is preferred over Tau-U, appropriate SEs for this index have not been derived, which leads to some challenges in using it in research syntheses. The implementation of Tau_{BC} proposed by **?** also has two further problems. First, **?** proposed correcting baseline trends only when a pretest of the trend using a nonparametric rank correlation test is statistically significant. The power of this pretest depends on the number of observations in the baseline phase, and as a result, the size of the trend adjustment depends to an extent on the number of baseline observations. This issue can be avoided by applying the trend correction uniformly, regardless of the statistical significance of the pretest. Second, **?** proposed to calculate Tau_{BC} using Kendall's tau-b statistic, measuring the rank correlation between the trend-corrected dependent variable (y) and a binary indicator for the treatment phase (x). Because tau-b makes an adjustment for ties, which occur necessarily because x is binary, the tau-b statistic will not generally correspond to the proportion of comparisons that represent improvement after adjusting

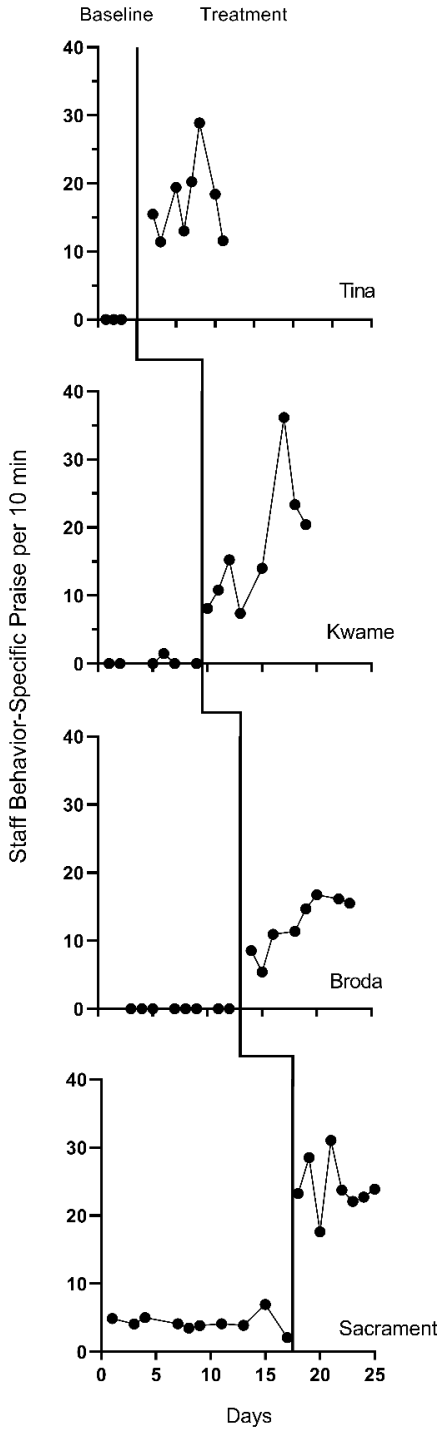


Figure 9.1: Multiple Baseline Design Across Three Participants (Knochel et al., 2021)

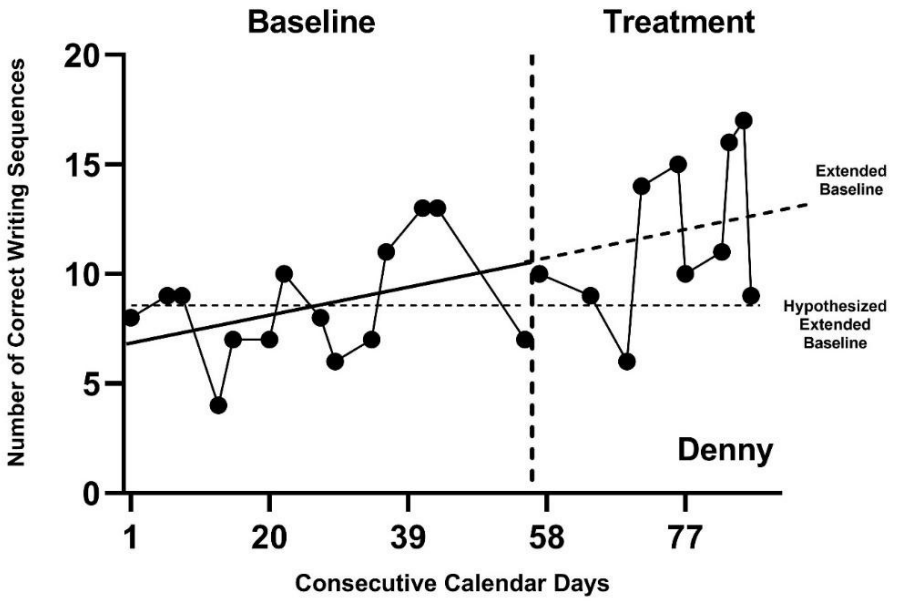


Figure 9.2: Effect for Denny (Rodgers et al., 2020)

for baseline trend, and the differences can be quite large in practice. This issue can be avoided by calculating Tau_{BC} directly from the formula, rather than via the tau-b statistic[2].

9.5 Assumptions and Limitations of Standardized Mean Differences

The within-case standardized mean difference [SMD_{W}; **?**; **?**] index describes the magnitude of an intervention effect in terms of the mean difference between treatment and baseline observations, standardized by dividing the mean difference by an estimate of the within-case standard deviation (SD). Researchers can further incorporate a small-sample correction to reduce the bias of the SMD_{W} estimate. For a given amount of variation in the dependent variable, the SMD_{W} is larger when the raw score mean difference is larger; for a given raw score mean difference, the SMD_{W} is larger when the within-phase variation in the dependent variable is smaller. The within-case SD can be calculated assuming homogenous or heterogeneous between-phase variance in observations. If researchers assume

[2]The Single Case Effect Size Calculator (**?**) allows the user to choose whether to use a pre-test for significant trend. By default, it will implement the trend adjustment regardless of the statistical significance of the baseline trend.

that the variance in observations differs across phases, only the baseline SD is used in the estimation of the effect size; if researchers assume that the variance in the observations is homogeneous across phases, the within-case SD is calculated by pooling the variance across the baseline and treatment phases.

The set of assumptions underlying the SMDW SE calculations, including those in the Single-Case Effect Size Calculator, are that observations in any given SCD study are mutually independent and homogeneously distributed (at least throughout the baseline phase). SCD study data may also violate the homogeneity of variance assumption, particularly when outcomes are reported as frequencies or percentages (**?**) where the degree of variation tends to be related to the mean level. Single-case data can be autocorrelated as opposed to independent (**??**). In theory, an autocorrelation parameter could be added to the statistical model to loosen the assumption of independence. However, adjusting the SMD_W estimator and the SEs for autocorrelation is challenging; autocorrelation parameters are difficult to estimate in short series, and the bias in these estimates leads to biases both in the effect size estimator and in its SE. Yet, to date, there is no consensus among researchers about the tolerability of independence and homogeneity assumption violations in calculating effects and SEs using standardized mean differences. If researchers need an approach that can tolerate some inaccuracies in the SE estimates in the synthesis of effect sizes, robust variance estimation may be preferable (**?**), as we illustrate in Chapter 10.

An additional limitation of the SMDW is that there are circumstances in which it cannot be calculated. Consider the case of Stephen (Gevarter & Horan, 2019) in Figure **??** below. Stephen's baseline values are all equal to zero. In situations where baseline values for a case are consistently the same value (e.g., all 0s or all 2s), the baseline SD is equal to 0 and it is not possible to divide the raw score mean difference by a baseline SD of 0. Furthermore, effect size estimates using this method can conflict with visual analysis when the baseline SD is very small because the resulting SMD_W may be very large, although the SCD graphs do not portray such. When this occurs, these very large values may show up in a synthesis as outlying effect sizes, creating challenges for summarizing the effects using meta-analysis.

Another limitation of SMD_W is that to interpret it as an effect size, one must assume that there is no baseline trend and that the difference between baseline and treatment observations is due to the intervention and not some other factor. In Figure **??** for the case of Denny, observations in baseline increase with time, suggesting an improvement in absence of the intervention. Where there is a baseline trend like this, the lack of experimental control suggests that some other factor (e.g., maturation) may have an influence on the behavior observed. It would be reasonable to assume that the upward trend in Denny's graph is due to some extraneous factor. Although the value of SMD_W for this graph is 1.23 with a SE of 0.54 and 95% CI [0.18, 2.28], this standardized mean difference cannot be attributed to the treatment alone. To interpret SMDW as the effect

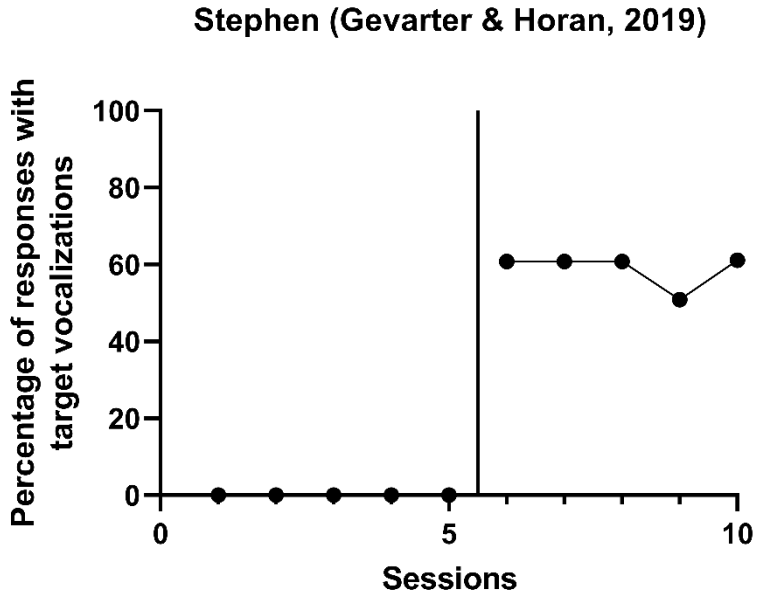


Figure 9.3: Effects of a Speech-Generating Device Intervention on Child Vocalizations (Gevarter & Horan, 2019)

of treatment, baseline stability is also assumed (i.e., no baseline trend). In some synthesis contexts, studies with unstable baselines may not meet the quality inclusion criteria; however, in other contexts, researchers may decide and have justification for the inclusion of such studies. If included, there is a need to consider alternative effect sizes that have baseline trend adjustments.

Researchers have developed adjustments to the SMD_W for baseline trend by estimating regression models where a coefficient in the model indexes the difference between the treatment and baseline phase trend lines at a specific focal time. Then, the raw score regression coefficient is standardized using an estimate of the within-case variability (**??**). Adjusting for trends, though, requires additional assumptions. One must assume a functional form for the trend (e.g., linear), and then assume that in the absence of intervention, the baseline trend would continue throughout the treatment phase. It can be challenging to argue that the unknown and unmeasured factor(s) leading to baseline instability will continue in a way that leads to the same trend. **?** provide an excellent discussion of the problems that can arise when trying to extrapolate baselines in SCD studies, as well as some potential solutions. In addition to the problems associated with trend extrapolation, the problems of trying to estimate and adjust for autocorrelation are exacerbated in models with trends (**??**). Further research is needed to provide better guidance on how to standardize mean differences when there may be both trends and autocorrelation.

9.6 Assumptions and Limitations of Percentage Change Indices and Log Response Ratios

The percentage change index, response ratios, and log response ratios are closely related case-specific effect sizes. Of these, the percentage change index (**?**) is considered the easiest to interpret. To obtain a percentage change effect size, one divides the raw score mean difference between treatment and baseline phases by the baseline mean and then multiplies the result by 100%. For example, if the baseline mean level of problem behavior is 20 and the mean treatment phase level is 5, one divides the raw score mean difference (15) by the baseline mean (20), and then multiplies the result by 100% to obtain a percentage reduction of problem behavior (i.e., percentage change) of 75%. Although easy to interpret, the sampling distribution of the percentage change index makes it challenging to use in a meta-analysis. Alternatively, the log response ratio (LRR) uses the same information (i.e., baseline and treatment phase means), but combines them into an index with improved distributional properties that is easier to work with meta-analytically. To compute the LRR index, one calculates the ratio of the intervention mean to the baseline mean, takes the natural log of this ratio, and then makes a correction for small sample bias (**?**). While the LRR values can be challenging to interpret directly, one can overcome this limitation by converting raw effect size estimates or meta-analytic results from the LRR scale

to a percentage change, using the formula:

$$\% \; change = (e^{LRR} - 1) \times 100\% \tag{9.1}$$

Converting the LRR scale to a percentage change allows researchers to take advantage of the desirable statistical properties of the LRR for meta-analytic modeling and the transparency of the percentage change scale for interpretation.

Approximate SEs for the LRR have been derived under the assumption that the observations within each phase are mutually independent and homogeneously distributed (**?**). However, as common in SCD research, observations can be autocorrelated (**??**). Positive autocorrelation will lead to downward bias in the SEs and overly narrow confidence intervals, giving the perception of greater certainty than is warranted. Although the potential for bias in the SEs is not ideal, it is still feasible to use the SEs in a meta-analysis. In situations where the accuracy of the SEs is a concern, we recommend that researchers use robust variance estimation (RVE) methods in the synthesis of the effect sizes because RVE is an approach to synthesis that can tolerate inaccuracies in the estimated SEs of individual effect sizes (**?**).

A limitation of LRRs, response ratios, and percentage change indices is that they are only appropriate for outcome measures that have a ratio scale. A ratio outcome is measured on a scale where zero corresponds to the true absence of the quantity being measured (true zero) and each unit on the scale is equal in size. Consider the variable "time spent reading." It can be measured as the number of seconds one spends reading prior to being distracted (and not reading). The ratio scale for this outcome would include 0 seconds, interpreted as zero seconds elapsed (true zero) or absolutely no time. Since a second is a consistent unit of measurement (the duration of each second on the scale is equivalent), and there can be a true zero on the scale, the outcome "time spent reading" can be measured using a ratio scale. Other examples of ratio scales include count-based variables recorded as frequencies (e.g., number of disruptive behaviors, number of pro-social behaviors) or percentages (e.g., percentage of intervals with on-task behavior). However, variables without a true zero are not ratio variables. For example, consider the measurement of a student's disruptive behavior using a behavior rating scale with the following four options: *rarely*, *sometimes*, *often*, and *almost always*. The variable *rarely* might be coded as 0, *sometimes* as 1, *often* as 2, and *almost always* as 3. However, the coded value 0 cannot be interpreted as an absolute zero (rarely $\neq$ never) and the newly coded units are not equal in size (e.g., 2 $\neq$ twice as much disruptive behavior as a 1). Thus, it is not appropriate to compute LRRs, response ratios, or percentage change for such a behavior rating scale or, more generally, for outcomes not measured on a ratio scale. However, these indices can also be difficult to interpret for ratio-based outcomes in reversal designs because the index is not symmetric with respect to the phase labels. Thus, a percentage change comparing treatment versus baseline is not necessarily the same magnitude as percentage change comparing baseline versus treatment.

Another limitation of LRRs arises in situations where the baseline has a constant value of zero because it is not possible to divide by zero or take the natural log of zero. Here it is important to consider whether we believe that the true mean level we are trying to estimate is actually zero or some value close to zero. For example, if our outcome is continuous, the true values may be close to zero but recorded as zero by our measurement instrument. As another example, consider a low-incidence behavior that happened to not be present in any of our baseline observations, but where we anticipate that we would occasionally get a non-zero value if we were to extend the baseline phase further. When the true mean is not zero, there is a method to adjust the sample mean estimate (**?**) and proceed to compute the LRR. However, when the baseline mean of zero reflects the true mean, such as when the behavior being measured is not yet in the child's behavioral repertoire, then the percentage change, response ratio, and LRR metrics are not appropriate.

In addition, for the LRR to be interpreted as an effect size, it must be assumed that there is no baseline trend and that the difference between the baseline and treatment observations arises from the treatment as opposed to some other factor. Returning to the graphed data for Denny (Figure **??** where there is a baseline trend suggesting that some other factor (e.g., maturation) is impacting the observations, LRR as an effect size is somewhat problematic. The estimated value of LRR representing a treatment effect for Denny is 0.34 (SE = 0.12, CI_{95} = 0.09, 0.58), with a corresponding percentage change index of 40% (CI_{95} = 9, 79), but these values cannot be attributed to the treatment. Put another way, to interpret LRRs or percentage changes as the effects of treatment, one must assume stability in baseline observations (i.e., there is no trend). In some synthesis contexts, studies with unstable baselines may not meet the quality inclusion criteria; however, in other contexts, such studies may be included. If included, there is a need to consider alternative effect sizes that have baseline trend adjustments. At present, we are not aware of methods in the literature that provide baseline trend corrections for LRR.

9.7 Assumptions and Limitations of Percent of Goal Obtained

Percent of goal obtained (PoGO; **?**) is a case-specific effect size that puts raw score effects onto a common scale by comparing the distance the intervention moved the outcome to the distance the intervention would have ideally moved the outcome. PoGO can only be computed if there is a goal. Goals can be set to the minimum of the outcome scale (e.g., the goal is to reduce problem behavior to 0), the maximum of the outcome scale (e.g., the goal is to increase on-task behavior to 100%), or the level of typically developing peers (e.g., the goal is to increase the number of socially appropriate interactions to the number that is typical).

The simplest and most readily available approach to estimating PoGO relies on the phase means (PoGO_M). For example, consider the study of an intervention intended to reduce problem behavior to zero occurrences (goal = 0). If the baseline mean was 40 and the treatment mean was 20 or half the level of baseline, the intervention would have moved the behavior halfway to the goal. Thus, $PoGO_M = \frac{40-20}{40-0} \times 100 = 50$. A PoGO estimate of 0 indicates no movement toward the goal, whereas a PoGO estimate of 100 indicates that the goal was achieved. However, PoGO is not bounded by 0 and 100. Values less than zero can occur if the intervention is harmful and results in observations moving the opposite direction of the goal. PoGO values can also exceed 100 if one or more of the observations exceed the specified goal. For example, if the researchers set a reading fluency intervention goal of 90 words read correctly per minute and treatment phase observations consistently exceeded 90, the PoGO estimate would be more than 100.

The assumptions underlying the calculations of approximate SEs for PoGO_M are that observations are independent and homogeneously distributed and that the sample means are approximately normally distributed (**?**). Single-case data can be autocorrelated as opposed to independent (**??**). While in theory one could add an autocorrelation parameter to the statistical model to loosen the assumption of independence, adjusting for autocorrelation is challenging. Autocorrelation parameters are difficult to estimate in short series, and the bias in these estimates would be expected to lead to biases in the SEs. The normality assumption and homogeneity assumptions may also be violated by single-case data, which are often based on frequencies or percentages (**?**), which can be non-normal and heterogeneous across phases, particularly if a phase mean is near the boundary of the variable scale (e.g., the mean baseline percent is near 0 or 100). Hopefully, future research will clarify the degree to which the commonly made assumptions of independence, normality, and homogeneity can be violated before the violations lead to substantial biases in the SEs.

An obvious limitation of PoGO is that it can only be used as an index of effects for studies that specify a goal or target criterion for the outcome under study. However, in some studies, there is no goal stated or implied. Without a goal, PoGO cannot be used as a measure of effect. In addition, to interpret PoGO_M as an effect size, it must be assumed that there is no baseline trend and that the difference between baseline and treatment observations arises from the treatment as opposed to some other factor. Returning to Figure **??**, if **?** specified a target goal of 20 correct writing sequences, then the PoGO_M value for Denny is 30.1. However, Denny's estimated PoGO_M cannot be attributed to the treatment because the trend in baseline suggests some other factor, like maturation. To interpret PoGO_M as the effect of treatment, baseline stability must be assumed (i.e., no trend). In some synthesis contexts, studies with unstable baselines may not meet the quality inclusion criteria; however, in other contexts researchers who decide to include such studies may need to consider alternative methods that have baseline trend adjustments.

To adjust PoGO for baseline trend, researchers can estimate regression models where a coefficient in the model indexes the difference between the projected baseline and actual treatment phase trend lines at a specific focal time. Next, the raw score regression coefficient is divided by the distance between the extended baseline trend line at the focal time and the outcome goal to obtain PoGO_b. The SEs for PoGO_b are then computed using the SEs from the regression coefficients (**?**). Adjusting for trends also requires additional assumptions, like a functional form for the trend (e.g., linear), and that the baseline trend would continue throughout the treatment phase in the absence of intervention. It can be challenging to argue that the unknown and unmeasured factor(s) leading to baseline instability would continue to have the same effect on the dependent variable over time (i.e., across subsequent phases) resulting in the same trend[3]. In addition to problems associated with trend extrapolation, problems of estimating and adjusting for autocorrelation can exacerbate in models with trends (**??**). Hopefully, in dealing with SCD study data having both trends and autocorrelation, future research will provide better guidance on how to estimate PoGO.

9.8 Case-Specific Effect Size Options for Synthesizing Single-Case Research

Let us suppose that a meta-analyst has chosen case-specific effect sizes based on the decision rules in Figure **??**. We suggest that the meta-analyst should then entertain four different case-specific effect size metrics: (a) nonoverlap, (b) standardizing, (c) response ratios, and (d) goal attainment. Using knowledge of the limitations of each case-specific effect size metric and the decision tree in Figure **??**, the meta-analyst can eliminate metrics clearly not appropriate for their synthesis context. For example, when the participant of an SCD study has a baseline of constant values of zero, one can rule out the use of standardizing. In addition to baselines values of zero, if the meta-analyst suspects that the true baseline mean is zero, they can also eliminate the response ratio approach from consideration. Likewise, one would eliminate the use of response ratios if the outcome scale was not a ratio scale (does not have a true zero). Finally, when the outcome does not have a goal, one would eliminate goal attainment from consideration. After this initial screening, there may be at least one approach (nonoverlap) and potentially several approaches left for further consideration.

For the approaches still under consideration, the meta-analyst should evaluate the degree to which each approach has the potential to produce effect size estimates that align with visual analysis. To do this, it may be necessary to begin estimating effect sizes. In choosing a specific effect size to estimate within an approach, researchers need to decide whether adjustments for baseline trends are necessary. When trend adjustments are unnecessary, we recommend the

[3]See for an excellent discussion of the problems associated with trying to extrapolate baselines in SCD studies, as well as some potential solutions.

use of NAP/Tau for nonoverlap, SMD_W for standardizing, LRR transformed into percentage change for response ratios, and $PoGO_M$ for goal attainment. If a linear trend adjustment appears most appropriate, we recommend Tau_{BC} for nonoverlap, β for standardizing, and $PoGO_b$ for goal attainment. Response ratio options are currently not available if baseline trend adjustments are needed.

If any of these effect size approaches result in estimates conflicting with visual analysis, we suggest that the approach and its results be ruled out for consideration in the overall synthesis. However, it would be beneficial for future researchers to have access to such computations and results. Thus, we encourage researchers to be transparent and make the effect sizes available in supplemental materials (or in an archive of replication materials) so that readers can assess for themselves the lack of alignment and decision to drop the index. At this point, a meta-analyst is left with only the effect size estimation approaches that align reasonably well to visual analysis. Figure **??** depicts a set of decision rules that correspond with these recommendations.

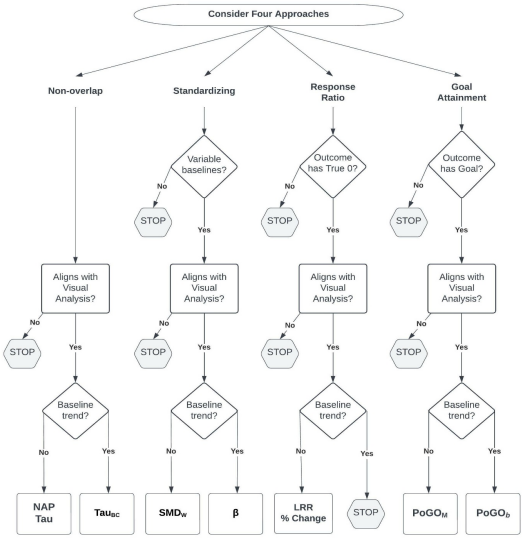


Figure 9.4: Flow Chart for the Selection of Case-Specific Effect Size(s)

After working through the decision rules, situations may arise in which multiple approaches are under consideration. At this point, a researcher must determine whether one or more indices will be selected for the meta-analysis. For example, the researcher could decide to estimate effect sizes and conduct meta-analyses using each of the viable effect size indices. The argument for including more than one index is that each has its own strengths and limitations, and using multiple indices allows us to check the sensitivity of our conclusions to the choice of effect metric (**?**). For those who prefer to select a single metric, we suggest selecting the index that best aligns with the visual analysis, or if no index aligns much

better than the others, selecting the index that tends to function the best within a meta-analytic model. For example, a study by **?** synthesizing NAP, SMD_W, and LRR found the LRR case-specific effect size functioned best (e.g., showed little to no bias and relatively accurate confidence intervals) across a variety of conditions for frequency count outcomes, followed by NAP. The SMD_W metric led to the most substantial problems. Thus, if viable, the LRR effect size is preferred over NAP, and NAP is preferred over SMD_W.

After selecting one or multiple indices for the meta-analysis, the researcher then turns to synthesizing the effects. Although multiple approaches are possible, the synthesis approach that works best depends on the effect size being synthesized. With LRRs, we recommend multilevel meta-analysis coupled with robust variance estimation (**?**). However, with NAPs and SMDs, the correlations between the effect sizes and their SEs create problems for using multilevel meta-analysis, so it is better to use simple averages with robust variance estimation (**?**). Robust variance estimation is generally helpful in synthesizing any of these case-specific effect sizes because some inaccuracies in the SEs are likely due to violations to the independence assumption that is made when estimating the SEs for NAP, Tau, SMD_W, LRR, and $PoGO_M$.

In the following chapter of this guide, we illustrate the application of case-specific effect sizes for the synthesis of single-case studies when researchers are interested in understanding the heterogeneity of effects across individuals in studies where measures and outcomes are not homogeneous. Using the flowchart in Figure **??** as a reference point, we provide step-by-step instructions for selecting one or more of the case-specific effect sizes, then demonstrate use of the *Single-Case Effect Size Calculator* (**?**) to obtain these effects and comparing them to what is found through visual inspection of participant graphs within each included study. Finally, we conclude Chapter 10 with a hypothetical synthesis using meta-analytic models.

Chapter 10

Application of Case-Specific Effect Sizes

This chapter illustrates the application of case-specific effect sizes in the synthesis of single-case studies. We provide step-by-step instructions to demonstrate the selection of the case-specific effect sizes that initially appear appropriate for the synthesis context. We then illustrate the computation of each of those effect sizes using the Single-Case Effect Size Calculator and compare the resulting effect sizes to our visual analysis. Finally, for the effect sizes that align with our visual analysis, we provide a demonstration of synthesis using meta-analytic models.

In this chapter, we illustrate the calculation and use of case-specific effect sizes in the synthesis of single-case studies. Case-specific effect sizes provide a measure of effect for each participant from each study on a common scale. Thus, these methods are particularly useful when researchers are interested in understanding variation (heterogeneity) of treatment effects across individual participants and where the outcome measurement procedures vary substantially across the included participants and studies. We present an example scenario where we want to synthesize evidence from several single-case design (SCD) studies that examine the effects of augmentative communication interventions (AAC; e.g., sign language, picture exchange communication systems, speech-generating devices) on increasing the independent requests (mands) of individuals with developmental and intellectual disabilities. For purposes of illustration, we selected three studies from the meta-analysis by (**?**).

The first study, **?**, used a reversal design to investigate the effects of an AAC intervention on increasing three adult participants' use of a switch device to communicate their desire for preferred items or activities while decreasing incidents of inappropriate behavior hypothesized to function as a request. For simplicity, we created a single outcome by combining both outcomes using this

formula:

$$\%\ switch\ activated\ responses = \frac{appropriate\ responses\ per\ minute}{inappropriate\ +\ appropriate\ responses\ per\ minute} \times 100$$

Participants received between 4 and 10 five-minute intervention sessions per day, with Jen and Tammy participating in sessions across three days and Rose completing all sessions in two days. The intervention sessions focused on switch training, which was "considered complete when the participant produced the response independently within 1 min on five consecutive opportunities," (**?**, p. 344). Therefore, we defined the outcome goal as use of the communication switch device on 100% of opportunities per session.

The second study by **?** included four student participants and used a multiple baseline design to examine AAC intervention effects on increasing the mean proportion of signed target requests in a classroom setting while simultaneously reducing the proportion of intervals engaged in problem behavior. The manual signing intervention was implemented by all classroom staff (one teacher, three aides) across the school day. The research team conducted observations of the participants during the first 30 minutes each morning at school across five weeks. Each participant's targeted behaviors were measured across a five-minute period using partial interval recording methods, in which observers checked a box next to any or all of the target behaviors that they observed at any point within a 10-second interval. The study goal was to reduce the proportion of intervals that each student engaged in problem behavior to zero (0) and increase positive communication skills to 1.0. We combined these outcomes to create a new outcome measure representing the proportion of communicative opportunities with signed requests.

The third study, by **?**, also used a multiple baseline design. The researchers examined the impact of a peer-mediated intervention along with an iPod-based speech-generating device to teach communication skills to four autistic students aged 5-12 years old who had limited or no vocal or verbal behavior and no existing successful use of an alternative or augmentative communication system. The primary outcomes were the number of independent mands (e.g., "I want juice.") as well as independent responses to a direct question (e.g., "I want juice." In response to "What do you want?"). Intervention consisted of peer assisted communication application (PACA), in which five typically developing children between 7-13 years old delivered communication training sessions with the participants that mirrored the six Picture Exchange Communication System (PECS) phases. Researchers collected frequency data on independent and prompted mands and responses to questions by each participant for as many as three sessions daily, with each session consisting of 10 trials to request known preferred items and answer the questions.

These three example studies used a range of different procedures for measuring outcomes and measured behaviors over sessions of different length. For example, **?** used frequency or count measures of the dependent variable during five-minute sessions (between four and 10 sessions per day) per participant,

whereas **?** collected data on no more than three 10-trial sessions per day (session duration was not specified). The outcomes from these studies are on a variety of different scales, and responses to these AAC interventions could vary across individual participants. Thus, it is advantageous to use case-specific effect sizes to synthesize results from these studies.

10.1 Selecting Case-Specific Effect Sizes for the Single-Case Studies

Using the decision rules in Figure **??**, we first select which of the types of case-specific effect sizes are viable for further consideration. Non-overlap indices can be computed with any type of outcome, so they are deemed initially viable. For the within-case standardized mean difference (SMD_W), we need a non-zero within-case standard deviation. The graphs of the data from the studies are provided in Figures **??** to **??**. Inspection of Figure **??** shows that each of the participants in **?** has baseline variability, and as seen in Figure **??**, the same is true for each participant reported in **?**. However, examination of Figure **??** (data from **?**) shows that the first case (Parker) has only three baseline values and that those values are all the same, such that the estimated baseline standard deviation is 0. To compute a standardized mean difference for this case, we would have to consider estimating the within-case standard deviation by pooling the deviations across the baseline and treatment phases, something we will discuss in more detail later. For all other cases, the baselines show variation, and the typical approach of standardizing using the baseline standard deviation is feasible.

Next, to determine if log response ratios (LRRs) are viable, we consider whether the outcomes are ratio-scale variables, which have a true zero and equal intervals (i.e., a unit at the lower end of the scale represents the same amount as a unit at the upper end of the scale). In **?**, the outcome is the percentage of switch-activated responses per session, in which zero would indicate absolutely no switch-activated responses and the intervals can be assumed equal. In **?**, we used both the signed request and problem behavior outcomes reported in the study to create a new dependent variable: the mean proportion of communicative opportunities with a signed request (taught behavior) per observation interval, which has a true zero. Finally, the outcome in the **?** study was the count of independent mands and responses per session, in which each session provided 10 opportunities. This count outcome has a true zero and is thus a ratio-scale variable. Because all outcomes are on a ratio scale, and none of the baselines are consistently zero, we determine that computing LRRs is feasible.

To determine if computing percentage of goal obtained (PoGO) is a viable option, we consider whether the outcomes have a goal, a scale value that would correspond to the outcome of an optimally effective or ideal intervention. For the percentage of switch-activated behaviors in **?**, we determined the goal level

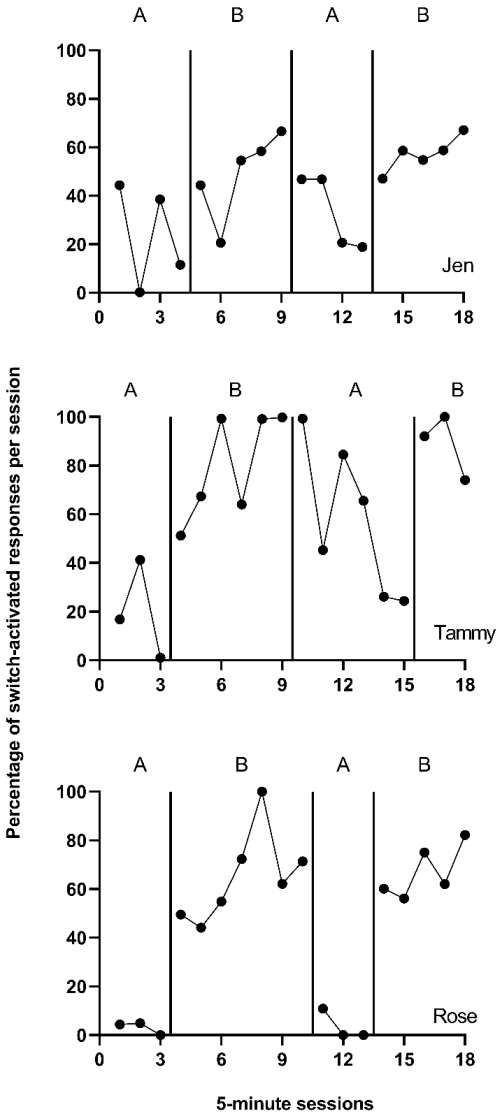


Figure 10.1: Effect of Intervention on Percent of Appropriate Responses (Byiers et al., 2014)

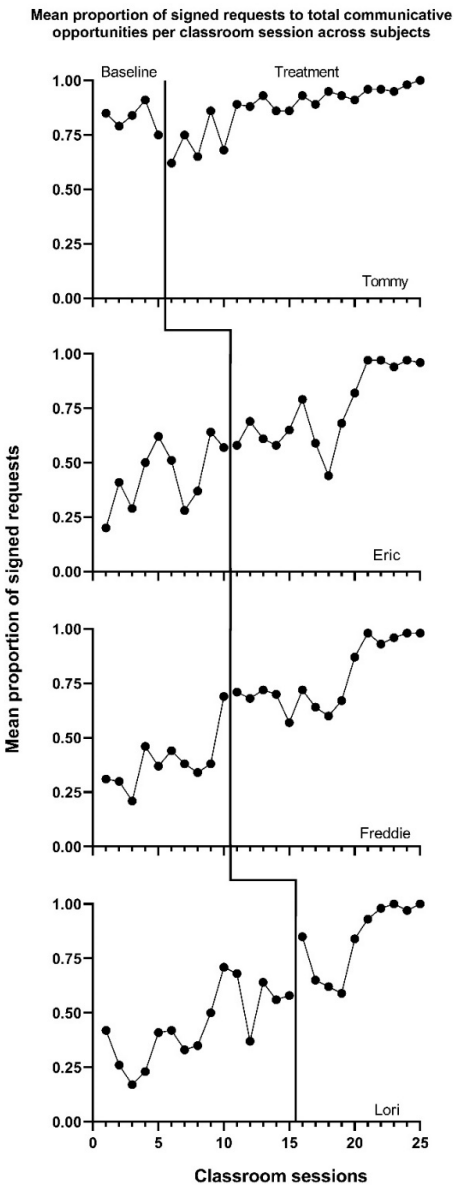


Figure 10.2: Effect of Intervention on Classroom Behavior (Casey, 1978)

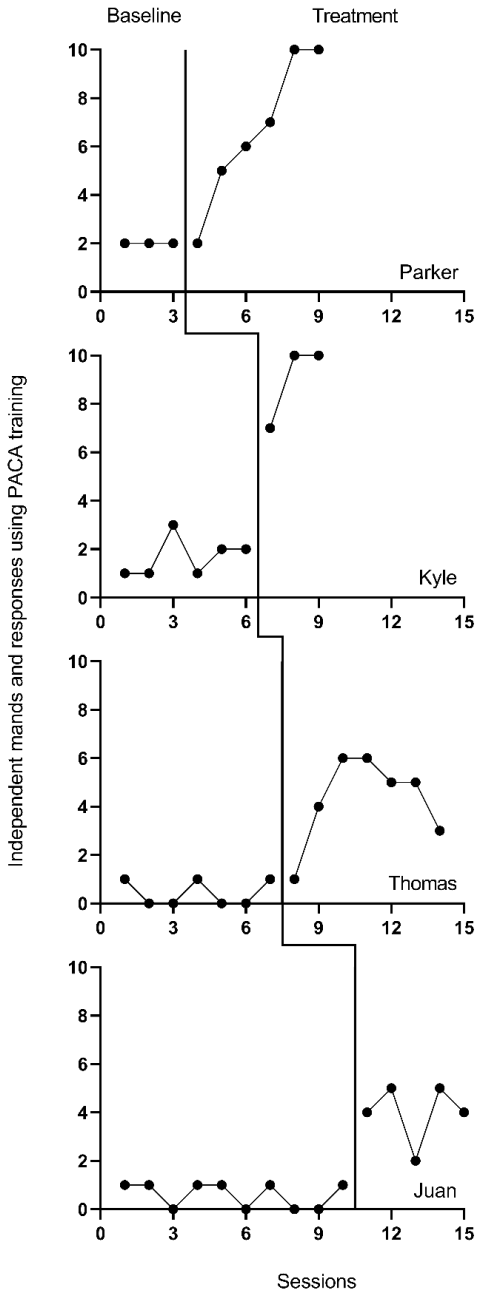


Figure 10.3: Effect of Intervention on Mands (Strasberger & Ferreri, 2014)

to be 100 percent. For the **?** study, the proportion of communicative opportunities with a signed request had a goal level of 1.0. Finally, for the count of mands and responses per session in **?**, the goal was specified as 10, a positive response to each of the 10 total opportunities presented per session. Thus, PoGO could be computed for each case in each study because the ideal value for each outcome was known.

Based on our initial consideration of the studies, each case-specific effect size category is potentially viable. The next decision we must make is whether to adjust for baseline trend. Based on our understanding of these outcomes, participants, and settings, we would expect stable baselines. When we visually analyze the plots in Figures **??**-**??**, the baselines appear stable for most cases, with the possible exceptions of Freddie and Lori in Figure **??**. Because we did not see trends for most cases and have uncertainty as to whether Freddie and Lori have trends that would continue, we decide to not make baseline trend adjustments, and to interpret Freddie and Lori's effect size estimates with caution and less confidence. Using **??** in Chapter 9 as our guide, we will estimate NAP for non-overlap, the SMD for standardizing, the LRR for response ratios (which we will transform to percentage change for interpretation), and PoGO_{M} for goal attainment. We illustrate the computation of each in the next section using the *Single-Case Effect Size Calculator* (**?**). After computing each of the effect sizes, we will then consider whether the estimates align well with our visual analysis of participant graphs in Figures **??**-**??**. Indices that align poorly with our visual analysis will be given less focus in our synthesis, whereas indices that align well with our visual analysis will be used to illustrate the averaging of case-specific effect sizes and will figure more heavily into our interpretation.

10.2 Estimating the Case-Specific Effect Sizes for the Included AAC Intervention Studies

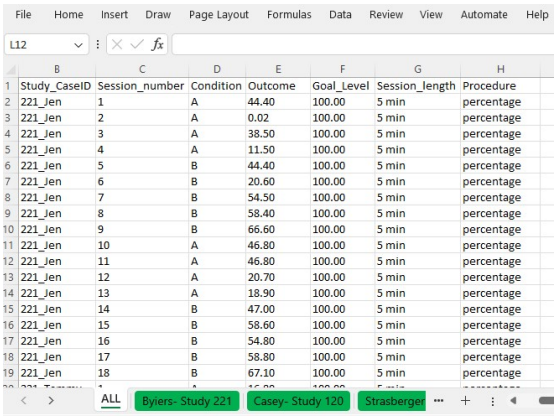
We demonstrate calculation of the case-specific effect size estimates using the web application, Single-Case Effect Size Calculator (**?**), which is a graphical user interface to the SingleCase R package. All calculation methods implemented in the app are documented at https://jepusto.github.io/SingleCaseES/.

Prior to using the app to estimate case-specific effect sizes, we show how the data from the three augmentative communication intervention studies are entered into an Excel data file. We then illustrate the process of accessing the app, uploading the Excel data file for analysis, defining the variables, examining graphs provided within the app, and estimating the case-specific effect sizes.

10.2.1 Entering the Data into Excel

Figure **??** is a screenshot of the Excel data file we use in this tutorial. These data are from three different studies examining the effects of AAC interventions on improving the requesting behavior of individuals with disabilities. There are four different spreadsheets within the file, three representing data extracted from each of the included studies (**???**). We also have a spreadsheet containing all participant data across the three studies, with the tab labeled ALL, as shown in Figure **??**.

The data in our spreadsheet are arranged using a long data format, which means that each row includes data for one observation of a participant in a given study. For example, the first row contains the first observation from the first case from the first study (e.g., Jan Session 1 in Figure **??**), and the second row contains the second observation from that case (Jan Session 2 in Figure **??**). After entering each of the observations from the first case of the first study, we enter each successive observation from the second case (e.g., Tammy in Figure **??**) of the first study, and continue until all observations for all cases in the first study have been entered. Then, beneath the data from the first study, we enter all values for the second study, again starting with the first observation from the first case and continuing through the last observation of the last case. All remaining studies are entered directly below the previous studies, with one row for each observation.



	B	C	D	E	F	G	H
1	Study_CaseID	Session_number	Condition	Outcome	Goal_Level	Session_length	Procedure
2	221_Jen	1	A	44.40	100.00	5 min	percentage
3	221_Jen	2	A	0.02	100.00	5 min	percentage
4	221_Jen	3	A	38.50	100.00	5 min	percentage
5	221_Jen	4	A	11.50	100.00	5 min	percentage
6	221_Jen	5	B	44.40	100.00	5 min	percentage
7	221_Jen	6	B	20.60	100.00	5 min	percentage
8	221_Jen	7	B	54.50	100.00	5 min	percentage
9	221_Jen	8	B	58.40	100.00	5 min	percentage
10	221_Jen	9	B	66.60	100.00	5 min	percentage
11	221_Jen	10	A	46.80	100.00	5 min	percentage
12	221_Jen	11	A	46.80	100.00	5 min	percentage
13	221_Jen	12	A	20.70	100.00	5 min	percentage
14	221_Jen	13	A	18.90	100.00	5 min	percentage
15	221_Jen	14	B	47.00	100.00	5 min	percentage
16	221_Jen	15	B	58.60	100.00	5 min	percentage
17	221_Jen	16	B	54.80	100.00	5 min	percentage
18	221_Jen	17	B	58.80	100.00	5 min	percentage
19	221_Jen	18	B	67.10	100.00	5 min	percentage

Figure 10.4: Example Data Spreadsheet (.xlsx) Formatting

The columns of the spreadsheet correspond to the variables the app will need to compute the case-specific effect sizes. Our study indicator variable appears in the first column, labeled StudyID. The values in this column can be either numeric or alphanumeric, but the variables should consistently represent the study from which the outcome observations were extracted. For example, all study data from **?** are assigned a study identifier of 221, **?** data have study identifier 120, and **?** have study identifier 158. The case indicator variable

appears in our second column, labeled *Study_CaseID*. The values in this column can also be numeric or alphanumeric. All within-case observations should have the same case indicator variable, with a unique code for each case in each study. To make it easier to track which case came from which study, we use values for each case indicator that are a concatenation of the study identifier and the case name used by the study authors. The third column, labeled *Session_number*, represents the session number for each recorded outcome value. It should only contain numerical values, with no values repeated across rows for a given case. The *Condition* column represents the phase indicator variable. It can be numeric or alphanumeric. Here we use A to indicate a baseline observation and B to indicate a treatment phase observation. The next column, *Outcome*, represents the value of our primary dependent variable in the synthesis-requests or mands. We have also included a goal level for each study, *Goal level*, which is needed to compute the PoGO_{M} effect sizes. Both outcome and goal-level values must be numeric. Finally, the last two columns, *Session_length* and *Procedure* allow us to include additional information regarding the measurement of the outcome, which can be useful in the estimation of some case-specific effect sizes. As seen in Figures **??** and **??**, the outcome for the first case from the first study (i.e., Jen) is based on a five-minute observation period, and the outcome is reported as a percentage of switch-activated responses per session.

10.2.2 Accessing the App

We can estimate each of the case-specific effect sizes for each participant, as well as the associated standard errors (SEs) and confidence intervals (CIs), using the web-based Single-Case Effect Size Calculator (**?**). The app, shown in Figure **??**, can be accessed at https://jepusto.shinyapps.io/SCD-effect-sizes/. For researchers comfortable using the R statistical computing environment, the app can also be accessed through the SingleCaseES R package (https://jepusto.github.io/SingleCaseES/). By running the app through R, one has the advantage of carrying out the calculations on their local computer, rather than on a cloud-based web server. As a result, the app will be faster and more responsive when run locally than when accessed via the website. Regardless of how it is accessed, the user interface and functioning of the app are the same.

At the top of the screen are headers for two distinct parts of the app. The *Single-Series Calculator* allows users to calculate effect sizes for individual cases (i.e., entering the data for one participant at a time)[1]. The *Multiple-Series Calculator* allows users to calculate effect sizes for several cases from one or more studies, all at once. The *Multiple-Series Calculator* is also useful for calculating more than one effect size measure for the same set of data. Meta-analytic contexts include data from multiple cases and multiple studies. Therefore, we focus on the *Multiple-Series Calculator* in this chapter[2]. Select the *Multiple-Series*

[1] A video demonstration of the Single-Series Calculator is available at https://www.youtube.com/watch?v=V_r9MEX9LwY.

[2] A video demonstration of the Multiple-Series Calculator is available at https://www.

Calculator tab (as indicated with the arrow in Figure **??**) to enter this part of the app.

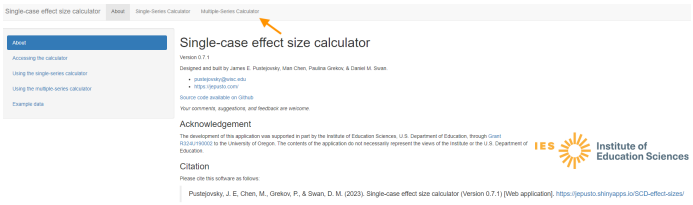


Figure 10.5: Single-case effect size calculator web application (Pustejovsky et al., 2023)

10.2.3 Loading the Data into the App

After clicking on *Multiple-Series Calculator*, we get the screen shown in Figure **??**. Towards the top are tabs for *Data, Variables, Plot, Estimate, and Syntax for R*. The first tab is *Data*, which appears black to indicate that it is the active tab, whereas the other, inactive tab labels are blue. The *Data* section allows us to select a dataset to use for effect size calculations. To upload our dataset, it must be saved as a .txt, .csv, or .xlsx file.

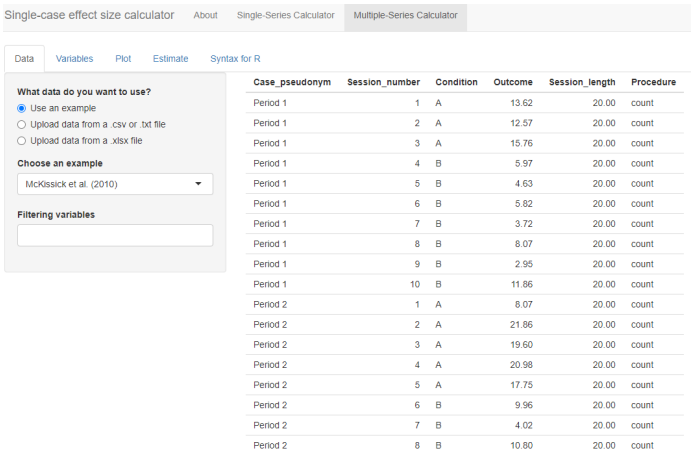


Figure 10.6: Initial Multiple-Series Calculator screen

To load our Excel data file (.xlsx) into the *Data* section of the app, we first select the choice "Upload data from a xlsx file" and then click the *Browse* option in the app (see Figure **??**) to locate and select the file as saved on our computer. The app will default to a checked box next to *File has a header?*. We leave this box checked because the top row of our Excel file contains our variable names.

youtube.com/watch?v=DSW7wuFG7og.

Because our file contains multiple spreadsheets, we must select a single sheet for analysis. Using the drop-down box, we select the sheet ALL that has the data from all three of the studies, as shown in Figure **??**.

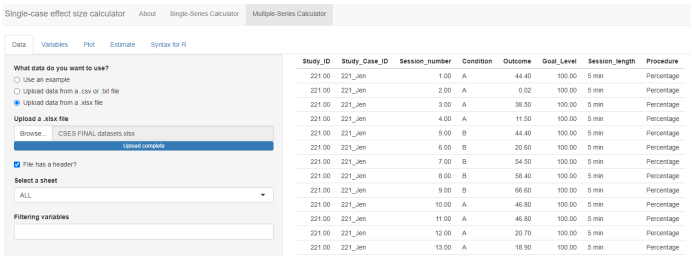


Figure 10.7: Initial Multiple-Series Calculator screen

Under the *Select a sheet* field, there is a field titled *Filtering variables*. If we click the cursor in the empty field, a menu appears from which we can select one or more variable names to use in defining a subset of the dataset. This option can be helpful for researchers who want to analyze only specific studies, cases, or phases from a larger dataset. For example, if we only wanted to compute the effect sizes for the first study, **?**, we could select *Study_ID* under filtering variables. Doing this would lead to an additional menu with the request *Please select the values for each filtering variable*. If we chose the value 120, the larger data set would be reduced to include only the data from study 120, and we could proceed to examine the graph for that specific study and to estimate the effect sizes for that specific study. As another example, some datasets might include studies with maintenance phases or multi-phase designs such as an ABCABC design. To calculate effect sizes for the comparison between phases A and B only, we could use the *Filtering variables* field to exclude data from the maintenance phases or C phases.

For present purposes, we leave the *Filtering variables* field empty because we want to estimate the effect sizes for all the cases in all three of our studies. Once the data are loaded, it automatically populates on our screen, so that we can quickly verify that the data presented correspond to the dataset we uploaded. In Figure **??**, we observe that the ID assigned to each study is listed under the *Study_ID* column (e.g., 221 for all participants in **?**), so that the measured dependent variable values (*Outcome*) for participant Jen (*Study_Case_ID*) for all sessions (*Session_number*) across phases (*Condition*; A = baseline, B = treatment) and specified goal level are accurately presented.

10.2.4 Defining the Variable within the App

After uploading our data, the next step is to move to the *Variables* tab. The menu, shown in Figure **??**, guides us through the process of indicating which of the variables in the dataset correspond to the variables needed for analysis.

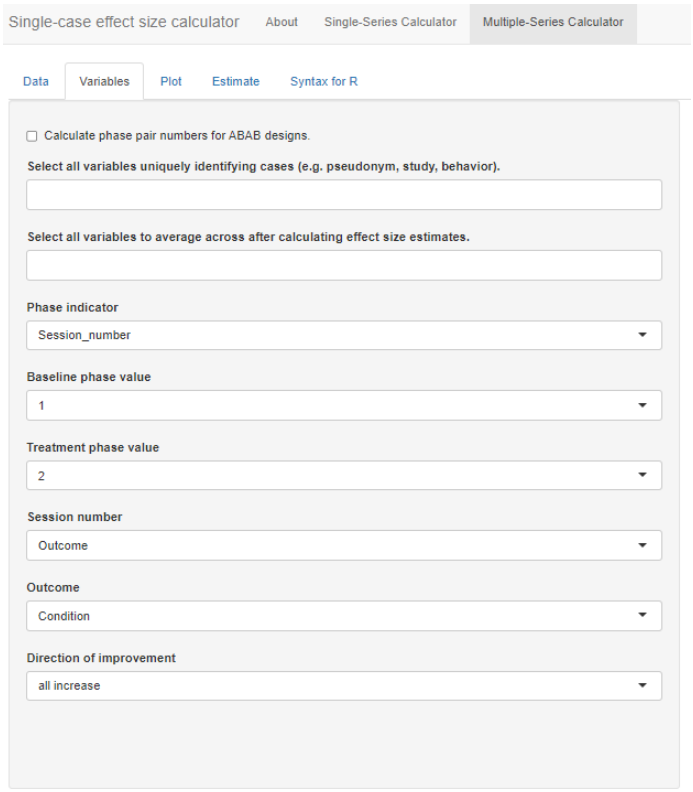


Figure 10.8: Variables Section Menu

At the top of the menu, there is a checkbox allowing us to *Calculate phase pair numbers for ABAB designs*. We check this option because one of our studies (i.e., **?**) is an ABAB design replicated across several cases. After checking the box, a new identified variable *phase_pair_calculated* is added to the dataset containing a label for each unique pair of sequential phases. Initially, the results in this variable will not make sense but they will be recalculated after we select unique identifier variables in the next step.

Below the optional checkbox, we are asked to *Select all variables uniquely identifying cases (e.g., pseudonym, study, behavior)*. This drop-down box allows us to specify the variable names that uniquely identify our included study cases. Because we have a unique name for each case (i.e., we don't use the same name for a case in two different studies), we can select our case identifier variable, *Study_Case_ID*. However, if the case labels were repeated across studies (e.g., "Case 1", "Case 2", etc. for each study), we would need to select both the *Study_ID* and *Study_Case_ID* to uniquely identify cases. In general, it is a good practice to select variables here for each relevant level of analysis. For our dataset, we select both *Study_ID* and *Study_Case_ID* so that we retain a variable with a unique ID for each study. Note that we must click outside the drop-down window or hit the Tab key to exit out of the options list.

After selecting these unique identifying variables, the app will automatically recalculate the *phase_pair_calculated* variable. The variable should now include a unique identifier for each unique pair of sequential phases within a data series. For instance, in **?**, each case will now have a label "1" for the first A phase and first B phase and a label "2" for the second A phase and second B phase. We recommend that users verify accurate calculation of *phase_pair_calculated* before continuing with variable selection.

Next, from the drop-down box *Select all variables to average across after calculating effect size estimates*, we select *phase_pair_calculated*. Particularly important for the **?** ABAB design, this option allows us to obtain an average effect size estimate across phase pairs for each case. While this variable was not in the original uploaded data file, it is added to the list because we checked the box, *Calculate phase pair numbers for ABAB designs*.

We specify our *Phase indicator* next, which is the variable we labeled *Condition* in our data file. Once selected, the app automatically populates what it believes are the baseline and treatment phase values. With this dataset, it is correct that *A* indicates a baseline observation and *B* indicates a treatment observation. However, researchers should verify the accuracy of this field. Then, from the *Session_number* menu, we select our variable named *Session_number* and we select our variable *Outcome* from the Outcome drop-down menu.

The last item in the Variables section asks us to specify the expected direction of the data paths for the included cases anticipated given our outcome of interest, and whether the measured outcome is expected to increase or decrease in the treatment phase. Because our included studies examine participants' ac-

quisition of communicative skills, specifically requesting behavior, we select *all increase* from the *Direction of improvement* field. If the studies in the dataset include outcomes measured in different ways and with different valence, then the dataset will need to include a variable indicating the direction of improvement for each case, labeled as "increase" or "decrease." If the dataset includes such a variable, we could select the "by series" option from the *Direction of improvement* field, resulting in a new field appearing, *Select variable identifying improvement direction*. We would use this field to indicate the name of the variable containing the labels for direction of improvement.

At this point, we have completed all sections on the *Variables* tab. As shown in Figure **??**, we have specified which of the variables in the data set correspond to the study indicator variable, case indicator, phase indicator variable, outcome, etc. We are now ready to click on the *Plot* tab in the left panel of the screen, so that we can visually inspect the data for each of the cases.

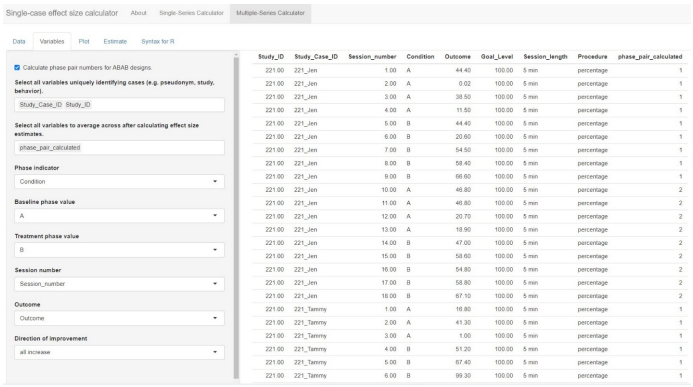


Figure 10.9: Multi-Series Calculator Variables tab

10.2.5 Examining the Graphs within the App

The *Plot* tab of the Multiple-Series Calculator displays a graph of outcomes by session number, differentiating baseline phases from treatment phases using green and red data paths, respectively. The *Display plots for each value of this variable* field allows us to specify variables with which to group the data plots. By default, the field defaults to *None*, which results in the app plotting only those data for the first pair of AB phases from the first case appearing alphanumerically in our dataset. However, we want to view more than one case at a time and select *Study_Case_ID* from the drop-down options under *Display plots for each value of this variable*. Then we can select a variable from the drop-down box under *Select a value for each grouping variable* to plot certain values for the grouping variables (e.g., study 221; Figure **??**). The app plots all cases on the same scale, so if reviewing the plots for multiple studies at once or for a study in which cases have different outcomes, a case with a relatively small

scale may appear to have near-zero levels of responding.

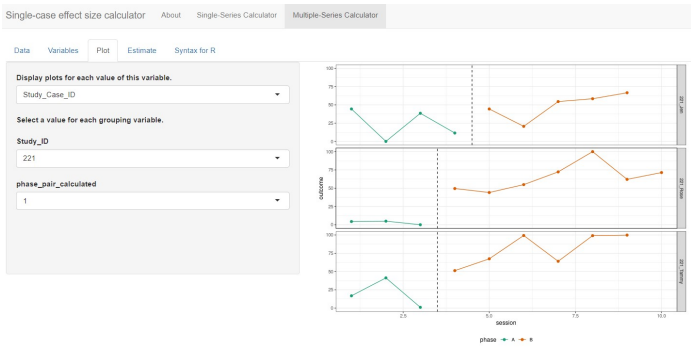


Figure 10.10: Multiple-Series Calculator Plot Section

At this point, with *Study_Case_ID* selected from the *Display plots for each value of this variable* drop-down menu, the app automatically populates *phase_pair_calculated* under the header *Select a value for each grouping variable*, with the default set to 1. Thus, the plot shown is for the first AB phase pair of the selected case. Since the **?** study is a replicated ABAB design (i.e., multiple AB phase pairs), we must change this menu option from 1 to 2 so we can examine the plot for the second AB phase pair. Although researchers will likely have already reviewed the plots from their primary studies to determine which case-specific effect sizes would be appropriate to calculate, it is useful to review them again here. An additional review of these data plots ensures that the app has read the data as intended and can verify researchers' judgements about trends or other features of the data series.

10.2.6 Estimating the Effect Sizes within the App

After reviewing the plots, we move on to the estimation of case-specific effect size indices by selecting the *Estimate* tab as shown in Figure **??**. The effect sizes that can be computed are grouped into two categories: (a) Non-Overlap Effect Sizes, and (b) Parametric Effect Sizes. Researchers can select one or more of the effect size indices that they are interested in calculating. For the set of studies included in this illustrative example, we decided to estimate non-overlap of all pairs (NAP) as a non-parametric effect size. We also selected several parametric effect sizes to estimate, including the within-case standardized mean difference (SMD), log response ratio for increasing outcomes (LRRi), and percent of goal obtained (PoGOM). When we check the box next to each effect size index, the box turns blue (see Figure 10.12). For some effect size measures, additional estimation options also appear. For example, after checking the SMD box, the Standardize SMD field appears, asking us to select a quantity by which to standardize. Researchers should choose the option that is most appropriate for their synthesis, selecting either the first radio button to standardize the

SMD using the standard deviation of baseline observations (baseline SD) or standardize the SMD using the standard deviation pooled across the baseline and treatment phases (pooled SD). We choose to leave the standardization to the default option, baseline SD, for this illustration.

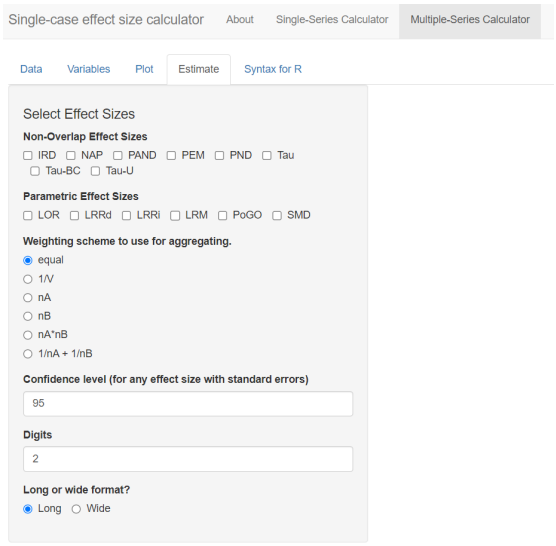


Figure 10.11: Multiple-Series Calculator Estimate Section

Additional fields also appear when we opt to calculate LRRi. These fields allow us to provide more detail about the nature of our outcome and to select additional output to aid in the interpretation of results. The first field, *Outcome Scale*, allows us to indicate how the outcome measurements are scaled. The menu defaults to count (i.e., all variables are measured by counts). However, the outcome measures vary across our included studies—for **?** the outcome is a percentage, for **?** the outcome is a proportion, and for **?** the outcome is a count. Therefore, to specify different outcome scales for our different cases, we choose the option *by series* under *Outcome Scale*. When we do so, another menu labeled *Select variable identifying outcome scale* appears. Here, we select the *Procedure* variable, which indicates the outcome scale used for each case in the dataset.

In some contexts, researchers may need to specify variables in their dataset using one or all three remaining menus on the page. These variables allow the app to estimate LRRs for cases with zero or near-zero levels during a phase. Because all studies in our example dataset have non-zero baselines and treatment levels, selections in these fields will not affect the results of this illustration, but they may influence LRRi calculations in studies with outcomes at or near zero. For the menu, *Optionally, a variable identifying the number of intervals per session*, researchers can select the corresponding variable in their dataset that contains

the number of intervals per observation. We leave this field at the default value of *NA*. Similarly, using the *Optionally, a variable identifying the length of each observation session* menu, researchers can define the observation session length in minutes. However, if this value is unknown or not relevant to the calculations at hand, researchers can opt to choose *NA*, which is what we have chosen to do. Finally, the menu labeled *Optionally, provide a floor for the log-response or log-odds ratio? Must be greater than or equal to 0* allows users to define the floor constant. This field provides the app with the information needed to estimate or specify a truncation constant. In either case, this makes it feasible to compute the LRR-increasing and LRR-decreasing values (see **?**, **?** for further details). We leave this section blank for the purposes of our illustration. However, other researchers analyzing data with near-zero baselines may find it helpful.

Six more fields remain in the *Estimate* section menu. Next, we move to specifications relevant to the PoGO_{M} case-specific effect size. Using the *Set the goal level for PoGO* drop-down menu, we can opt to set the same goal for every series in the dataset (i.e., *common goal*) or to set a different goal for each series (i.e., *different goals across series*). Selecting *common goal* triggers another field requesting us to select a single value that represents the goal level across all cases and studies. However, because we have three different studies, each with different goal levels, we select the alternative PoGO estimation option: *different goals across series*. Beneath this field, we then use the drop-down menu to specify the variable in our dataset that represents the goal level (e.g., *Goal_Level*; see Figure **??**).

After selecting the effect size indices and providing the app with additional information needed for computation where relevant, we have four additional options to consider, as shown in Figure **??**. The *Weighting scheme to use for aggregating* section defaults to *equal*. For our illustration, we are aggregating the effect sizes across the AB pairs in the ABAB design (**?**). With the calculator default as *equal*, the app will average the effect sizes from the two AB pairs by assigning them equal weights. This default is typically recommended for most syntheses and is appropriate here, so there is no need to change it. However, if there are substantial differences in the information used to compute one effect size versus another (e.g., one effect size is calculated with a larger set of observations or substantially less variable data), other options may be more appropriate. For example, the multiple-series calculator allows researchers to weight the effect sizes by the number of observations in the baseline phase (option nA), or the inverse of the error variance (option $1/V$).

The next menu, labeled *Confidence level (for any effect size with standard errors)*, lets us specify the preferred coverage level for the confidence intervals reported for effect size measures that have known sampling variances. We leave this field at the default and standard value, *95*. However, researchers can adjust the value to meet their needs. For the menu *Digits*, we choose the decimal places used in the reporting of our calculated effect size estimates (along with their SEs and confidence limits). We keep the value default of 2, which is consistent

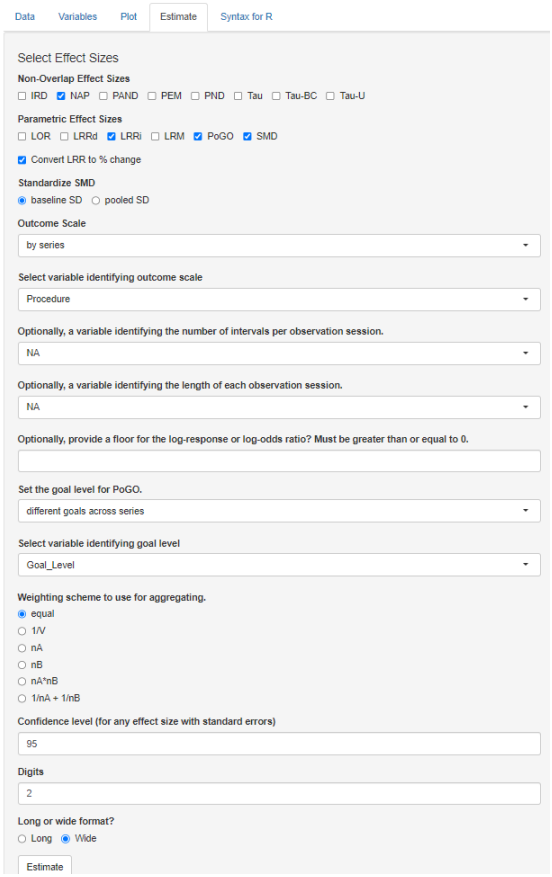


Figure 10.12: Estimate Section of the Multiple-Series Calculator

with APA reporting guidelines.

Finally, the last menu *Long or wide format?* in the Estimate section presents two options for how our output will be presented: long format or wide format. The app default is *Long*, which will arrange the output by cases, with each effect size appearing in a separate row. Since we opted to calculate four different effect size indices, there will be four rows per case. Alternatively, we can choose to view our output as *Wide*. Wide format organizes results with each case assigned to its own row, and each different effect size index represented by separate columns. In this scenario, we would view our cases in rows and scroll to the right to view each of the different effect sizes in a separate column. Although no format is better than the other, we present the wide format in Figure **??**. With a single line per data series, we can more easily complete across-case comparisons of specific effect size indices. With our data containing cases from multiple studies, the wide format will help us assess the degree to which one or more of the calculated effect size metrics provides estimates that align well with our visual analyses.

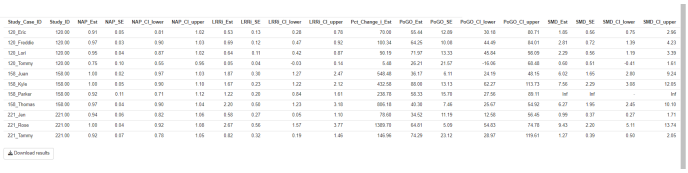


Figure 10.13: Estimate Section of the Multiple-Series Calculator

At this point, to view the results, we can click the *Estimate* button at the bottom of the screen. By clicking this button, the multiple-series effect size calculator will display the results table with each effect size estimate and their corresponding SEs and confidence limits. Figure **??** shows an example of the table that appears to the right of the *Select Effect Sizes* options we selected in Figure **??**. The table includes the NAP, SMD, LRRi, and PoGO_{M} estimates for each individual case across all three included studies, as well as associated 95% confidence intervals. Pressing the button labeled *Download results* will download the output as a .csv file, which will be useful in the further processing of the effect sizes.

10.3 Syntex for R

The *Multiple-Case Effect Size Calculator* also provides auto-generated R code, which can be useful for reproducing the effect size calculations carried out through the menu-driven interface. If we want to obtain R code, we click on the *Syntax for R* tab, as shown in Figure **??**. After clicking on the Copy button at the top of the script field, we can paste the auto-generated code into an R script[3]. We can then save the script to our computer for use in future analyses

[3]To do this using the popular RStudio program, click on the green plus button on the top left of the RStudio interface, select R script, and then paste the auto-generated code into the

or for replicating the calculations carried out within the app.

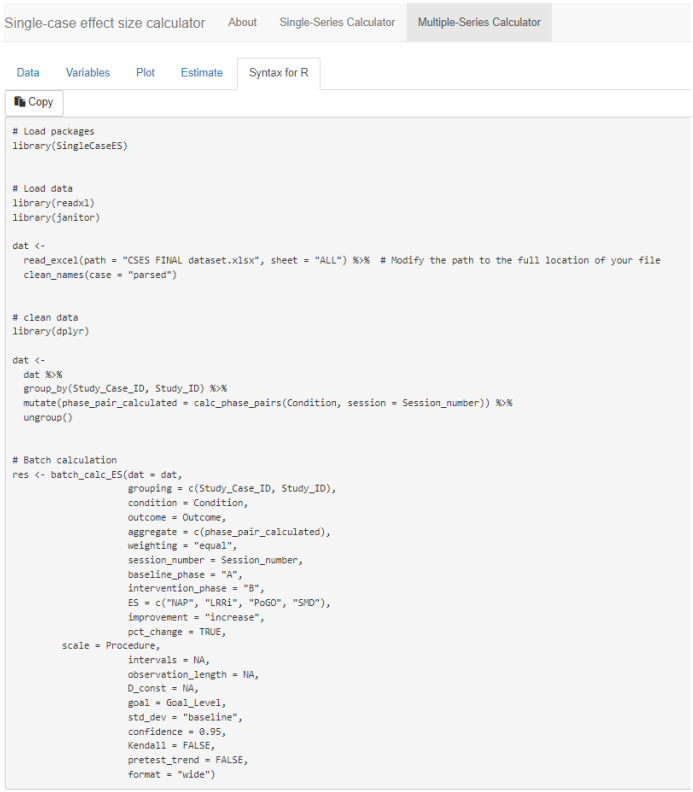


Figure 10.14: Generated Syntax for R Script

10.4 Examining the Alignment of the Case-Specific Effect Sizes with our Visual Analysis

After computing the different effect sizes, it is useful and informative to compare the effect size estimates to the graphical displays in Figures **??** through **??**. Specifically, we want to examine the degree to which the variation in effects seen in the graphs corresponds to the variation in the values of the effect estimates obtained using the *Single-Case Effect Size Calculator*. Effect size estimates in a metric that correspond well with the results of visual analysis are conceptually easier to interpret.

---

blank R script.

10.4.1 NAP

NAP values of .50 or less suggest that at least half of the paired A and B phase data points are opposite in direction to what is expected if the treatment were helpful. Thus, if a treatment were harmful or ineffective at best, we would expect NAP values between 0 and .5. Alternatively, NAP values that are higher than .5 are indicative of positive treatment effects, with the largest possible effect being a NAP of 1.

Each participant's estimated NAP effect size per study is shown in Figure **??** and listed in Table **??**. Starting with **?** participants, Tammy had a NAP effect size estimate of .92, which tells us that there is a 92% probability that a randomly selected observation from Tammy's treatment phase improved upon a randomly selected baseline phase observation. We also obtained an SE of .07 and 95% CI [0.78, 1.05][4]. Jen had a slightly larger NAP effect size of .94, with an SE of .06. Finally, the NAP effect for Rose was estimated as 1.0, with an SE of .04. Rose's NAP effect size was the largest value possible, which can be interpreted as 100% of Rose's treatment-baseline phase observational pairs having no overlap.

Next, we compare these NAP estimates to what is visually depicted in the Figure **??** graphs. Our visual analysis suggests positive effects for each participant. With Rose we see no overlap between phases and a large initial effect of the intervention. With Jen and Tammy's graphs, we see smaller initial effects, some overlap with the baseline phases, and increasing trends in the treatment phases. The NAP values of all participants being relatively high is consistent with the visual impression of positive effects for each participant, and Rose having a NAP of 1, which is the largest value is consistent with our visual inspection and conclusion that the intervention had the largest effect for Rose. Thus, for this study, there is a reasonable level of congruence between our visual analysis and the effect estimates.

Table **??** also reports the NAP effect size estimates for our included **?** study with four participants. The obtained NAP values ranged from .75 (Tommy) to .97 (Freddie). Eric's estimated NAP effect size was .91, with an SE of .05. The second largest NAP value was .95 for Lori, with an SE of .04. Visual inspection of **?** participant graphs revealed the smallest effect and the largest amount of overlap between baseline and treatment phase data for Tommy, which is congruent with the obtained NAP effect size estimate of .75 and SE of .10. The lower baseline and higher average treatment levels noted in the graphs for Eric, Lori, and Freddie resulted in larger NAP effect sizes; overall, these NAP

[4]When NAP values are aggregated across multiple phases, the confidence interval for the aggregated NAP is based on a large-sample normal approximation and can therefore include values outside the logical range of 0 to 1. Specifically, the point estimate and standard error of the aggregated NAP are calculated using the user-selected weighting scheme. An ω-level confidence interval is then calculated by adding and subtracting z_ω times the standard error from the point estimate, where z_ω is a standard normal critical value. Thus, if an aggregated NAP estimate is near the extreme of the range, the confidence interval can include values larger than 1 or less than 0.

estimates are consistent with the results of our visual analysis of participant graphs.

Finally, we review the four NAP effect size estimates for the Strasberger and Ferreri (2014) participants shown in Table **??**. Parker had the lowest NAP effect size estimate of .92 with an SE of .11. The estimated NAP effect size for Thomas was .97, with a SE of .04, while both Juan and Kyle had NAP effect size estimates of 1, meaning that there was 0% overlap between baseline and treatment observation pairs. The NAP effect sizes obtained using the app, representing the degree of non-overlap of all baseline and treatment phase pairs, appear to be consistent with the overlap we observe in the **?** graphs in Figure **??**. However, Parker, who had the lowest NAP, reaches 10 independent mands in the last two treatment sessions, whereas Juan, who has a NAP of 1.0, never exceeds 5 independent mands. Thus, there is some inconsistency between the magnitude of effect seen in the visual analysis and the values of NAP for the graphs of **?**.

10.4.2 SMD Results

We also see the SMD outcomes in Table **??** for all cases across all three included studies. Starting with the **?** study, we examine the SMD results for Jen, Tammy, and Rose. Jen's estimate of 0.99 indicates that the average percentage of switch-activated requesting behavior shifted between the treatment and baseline phases by 0.99 standard deviations. This estimated SMD effect size has an SE of 0.37. Tammy's SMD effect size was 1.27, with an SE of 0.39. The third participant, Rose, had the largest estimated SMD effect size of 9.43, with an SE of 2.2. The smallest SMD value assigned to Jen (0.99) appears consistent with the results of visual analysis, which revealed a large amount of data overlap and a smaller degree of level change between phases. To a slightly lesser extent, the between-phase data overlap and smaller average level change seen in Tammy's graph was in line with the estimated SMD effect size of 1.27. Rose's largest SMD effect size was consistent with graphed data in Figure **??** that showed no between-phase data overlap and the largest change in level between phases.

Next, we examine the SMD estimates for the four **?** study participants. The SMD effect size estimates obtained using the app ranged between 0.60 (Tommy) and 2.81 (Freddie). Eric's SMD effect size was estimated to be approximately 1.85 standard deviations from the *baseline SD*, with an SE of 0.56 and 95% CI [0.75, 2.96]. Freddie had an SMD effect size of 2.81, SE of 0.72, and 95% CI [1.39, 4.23]. The second largest SMD treatment effect estimate was 2.29 for Lori, with an SE of 0.56 and 95% CI [1.19, 3.39]. Finally, Tommy's SMD effect size was 0.60, with an SE of 0.51 and 95% CI [-0.41, 1.61]. Visual analysis of participants' graphs in Figure **??** is consistent with these results, as Tommy had the smallest degree of level change across phases, and Freddie had a larger mean level change. Also, if we compare the SMD estimates across the first two studies, we note the largest SMD estimate was from Rose, and this also aligns with our visual analysis.

We conclude our comparison of SMD results obtained using the app and visual analysis results with the four **?** participants. In order from smallest SMD effect size to largest, Juan's estimated SMD effect size was 6.02 with an SE of 1.6 and 95% CI [2.8, 9.24], Kyle's SMD effect size was 7.56 with an SE of 2.29 and 95% CI [3.08, 12.05], and Thomas had an SMD effect size of 6.27, with an SE of 1.95 and 95% CI [2.45, 10.1]. However, as shown in Table **??**, only three out of the four participants have SMD effect estimates reported; the *Multiple-Case Effect Size Calculator* reports the code "Inf" for Parker. This code means the effect size is infinite, which results from trying to divide the mean difference by zero. This problem will occur anytime the baseline is used for standardization and there is no variability in baseline, which is the case for Parker. To get an effect size for Parker, we reran the analysis choosing to standardize by the *pooled SD* across phases. Parker's resulting SMD effect size was 1.60, with an SE of 0.73 and 95% CI [0.16, 3.03]. Although an effect size is computed, the value is notably smaller than the value for the other three participants. Another option would be to pool the SD across phases for all cases in this study, or to pool the SD across phases for all cases in all studies. Parker's smaller effect size value compared to Thomas or Juan is inconsistent with our visual inspection of the graphs in Figure **??**, where Parker is observed to give 10 independent mands in his last two treatment sessions, whereas Thomas and Juan, who have similar baseline levels to Parker, never exceed 6 independent mands. This illustrates a problem that is encountered with the SMD when baselines have little to no variability for a case. Another option would be to pool the SD across phases for all cases in this study, or to pool the SD across phases for all cases in all studies. Although Parker's SMD is somewhat problematic, the SMD effect sizes for the other cases (Thomas, Juan, and Kyle) are consistent with our visual analysis and reflect the improvement in level change seen from baseline to treatment across participants. Also, if we compare the SMDs across studies, we see that Rose has the largest SMD, followed by Thomas, Juan, and Kyle, which is generally consistent with our visual analysis.

10.4.3 LRRi Results

LRRi estimated effects for all cases across all three studies are reported in Table **??**. Starting with Jen (**?**), the LRRi effect size estimate of .58 represents the natural log of the average percentage of switch-activated requesting behaviors per session in the treatment phase divided by the average percentage of the requests during the baseline phase, giving us a relationship in terms of percentage change. Jen's LRRi was 0.58 and the estimated percentage change was 79%. Tammy's LRRi treatment effect was 0.82, with an estimated percentage change of improvement of 147%. Rose had the largest LRRi effect size with 2.67 and percentage change of 1390%. Visual analysis of the participants' graphed observations across phases was aligned with the LRRi results obtained in the app. For example, the LRRi effect size and percentage change estimates arranged in order of smallest (Jen: 0.58, 78.6% respectively) to largest (Rose: 2.67, 1389.7%) reflect the same order in which we would rate the degree of level change noted

between phases for all participants.

For the **?** study participants, the LRRi treatment effect estimates ranged from smallest to largest as follows: 0.05 (Tommy), 0.53 (Eric), 0.64 (Lori), and 0.71 (Freddie). Tommy's LRRi effect size SE was .04 with a 95% CI [-.03, .14] and an estimated percentage change of 5.5%. Eric's LRRi SE was .13 with a 95% CI [0.28, 0.78], and he had an estimated percentage change of improvement of 70%. Lori's LRRi SE was .11 with a 95% CI [.42, .87], and she had a percentage change increase of 90%. Finally, Freddie's LRRi SE was .12 with a 95% CI [.47, .92], and an estimated percentage change of 100%. In the treatment phase, all participant graphs showed an increasing trend, in the direction expected. For **?** estimated effects, we find the LRRi effect size results estimated using the app to be in line with what we see when we visually inspect the original graphs.

The LRRi treatment effect estimates for the four **?** cases ranged from 1.22 (Parker) to 2.2 (Thomas). Juan had an LRRi effect estimate of 1.87, with an SE of .3 and 95% CI [1.27, 2.47], and percentage change estimated at 548%. Kyle's LRRi effect size was 1.67, with an SE of .23 and 95% CI [1.22, 2.12], and percentage change of 433%. Parker's estimated LRRi effect was 1.22, with an SE of .2 and 95% CI [.84, 1.61], and percentage change estimate of 239%. Finally, Thomas had the largest reported LRRi treatment effect among the study participants with an estimate of 2.20, SE of .5, and 95% CI [1.23, 3.18]. He also had the largest estimated percentage change of 806%. Although Thomas showed the largest percentage change, there was overlap in observations across phases, and his treatment observations did not move as far from baseline or reach as high of levels as either Parker or Kyle. Thus, we find that the LRRi effect sizes for this study are not as consistent with the results of our visual analyses as they were for the other two studies. The explanation is that the LRR makes a ratio of the treatment mean and the baseline mean. Because for Thomas and Juan the baseline means are notably lower than the mean for Kyle, the treatment phase means are divided by smaller values, leading to larger effect sizes.

10.4.4 $\mathbf{PoGO_{M\uparrow}}$

PoGO_M effect estimates for all cases across all three studies are reported in Table **??**. Starting with the **?** study participants, the case-specific estimates in order from smallest to largest are as follows: Jen had a PoGO_M effect size of 34.5 with an SE of 11.2, Tammy had a PoGO_M effect size of 74.3 with an SE of 23.1, and Rose had a PoGO_M effect size estimate of 64.8 with an SE of 5.1. Visual inspection of the participants' graphs in Figure **??** showed variability in effects across cases, with more of Tammy's treatment phase (B) observations at or approaching the goal level (100%) than other participants. This is consistent with the PoGO_M effect size results (74.3) obtained in the app. Similarly, the smallest PoGO_M effect size estimate for Jen (34.5) is also aligned with the data graphed in Figure **??**. However, for one participant, the PoGO_M results ranked from smallest effect (Jen) to largest (Tammy) are somewhat in conflict

Table 10.1: Case-Specific Effect Size Estimation Results Across Studies

		NAP		SMD~W~		LRRi	
Study	Case	Est. (SE)	95% CI	Est. (SE)	95% CI	Est. (SE)	95%
Byiers et al., 2014	Jen	.94 (.06)	[0.82, 1.06]	.99 (.37)	[0.27, 1.71]	.58 (.27)	[0.05,
	Tammy	.92 (.07)	[0.78, 1.05]	1.27 (.39)	[0.50, 2.05]	.82 (.32)	[0.19,
	Rose	1 (.04)	[0.92, 1.08]	9.43 (2.2)	[5.11, 13.74]	2.67 (.56)	[1.57,
Casey, 1978	Eric	.91 (.05)	[0.81, 1.02]	1.85 (.56)	[0.75, 2.96]	.53 (.13)	[0.28,
	Freddie	.97 (.03)	[0.90, 1.03]	2.81 (.72)	[1.39, 4.23]	.69 (.12)	[0.47,
	Lori	.95 (.04)	[0.87, 1.02]	2.29 (.56)	[1.19, 3.39]	.64 (.11)	[0.42,
	Tommy	.75 (.10)	[0.55, 0.95]	.60 (.51)	[-0.41, 1.61]	.05 (.04)	[-0.03,
Strasberger & Ferreri, 2014	Juan	1 (.02)	[0.97, 1.03]	6.02 (1.65)	[2.80, 9.24]	1.87 (.30)	[1.27,
	Kyle	1 (.05)	[0.90, 1.10]	7.56 (2.29)	[3.08, 12.05]	1.67 (.23)	[1.22,
	Parker	.92 (.11)	[0.71, 1.12]	Inf		1.22 (.20)	[0.84,
	Thomas	.97 (.04)	[0.90, 1.04]	6.27 (1.95)	[2.45, 10.10]	2.20 (.50)	[1.23,

with what we see in Figure **??**; Rose's PoGO_{M} effect size was estimated smaller than Tammy's, despite having graphs portraying the largest between-phase level change and no data overlap between phases.

The **?** $\text{PoGO}_{\text{M}\uparrow}$ effect estimates are also shown in Table **??**. Lori had the largest PoGO_{M} effect size of 72.0 with an SE of 13.3, followed by Freddie with a PoGO_{M} effect size of 65.3, SE of 10.1, and Eric with a PoGO_{M} effect size of 55.4 and SE of 12.9. Although Tommy had the largest treatment phase mean ($\beta = 0.87$), he also had the highest baseline phase mean ($\alpha = 0.83$). As a result, and consistent with the graphed observations in Figure **??**, Tommy's change in level between baseline and treatment was much smaller in magnitude than any other participant, and the smaller effect was noted in the effect estimate of 26.2 with an SE of 21.6. When combined with the results of visual analysis indicating little data overlap between phases for Lori and Freddie, the larger difference in phase means aligned with their larger PoGO_{M} values.

10.5 Averaging the Case-Specific Effect Sizes

Although there were a few instances where inconsistencies were noted between an effect size calculation and our visual analysis, each of the effect size metrics produced values that were reasonably well aligned with our visual analysis for the majority of the cases. Thus, we now illustrate the use of meta-analytic methods to summarize each of our case-specific effect sizes across participants and across studies. Two distinct approaches to meta-analysis may be useful in this situation, based respectively on fixed effects and random effects models. We present each of these approaches in turn. We carried out all calculations using the `metafor` package (**?**) for the R environment for statistical computing (**?**).

10.5.1 Fixed Effects Meta-Analysis

First, researchers might wish to simply summarize the effect sizes across the cases in each study, or across all 11 cases from the three studies. For this purpose, a fixed effects meta-analysis is a convenient approach. In fixed effects meta-analysis, we draw inferences only about the cases and studies included in the summary, without generalizing beyond the observed participants. Results are summarized by taking an average of the effect sizes across cases and calculating SEs and CIs to represent the uncertainty in the average effect *for these particular cases*—but not for any broader population of cases or studies. Conventional fixed effects meta-analysis uses a weighted average with weights inversely proportional to the squared SE of the effect size estimates (i.e., inverse-variance weighting or precision weighting). However, for some of the case-specific effect size indices we have calculated, the SEs can be correlated with the effect size estimators. Moreover, the effect size estimators and SEs are based on very few observations per phase, making inverse-variance weighting methods less appropriate. Instead, we use simple, equally-weighted averages to summarize the effect sizes for each of the studies.

Table **??** reports the average effect size for each study and across all 11 cases included in the three studies. SEs and CIs are based on the fixed effects model. The average NAP effect size across all 11 cases is .94 with an SE of .02. The average NAP effect size appears similar across all three studies, ranging from .89 to .97. Both the study-level average and overall average suggest large intervention effect sizes, with relatively little overlap between treatment and baseline phases. The average SMD_W effect size across all 11 cases is 3.70 with an SE of 0.39. Study-level average effect sizes are more heterogeneous, ranging from 1.89 for **?** to 5.36 for **?**. The average LRR effect size across all 11 cases is 1.18 with an SE of 0.09, which corresponds to an average of a 225% increase from baseline to intervention, 95% CI [173%, 287%]. Study-level average LRR effect sizes appear heterogeneous, ranging from 0.48 (corresponding to a 62% increase) to 1.74 (corresponding to a 470% increase). Finally, the average $PoGO_M$ across all 11 cases is 55.8 with an SE of 4.2, suggesting that, on average, intervention led to improvements that were about 56% of the way toward the goal levels. Average $PoGO_M$ values appear quite consistent across studies, ranging from 54.5 for **?** to 57.9 for **?**.

Examining the results across effect size metrics, the effect size indices paint somewhat different pictures regarding the strength of the average treatment effect and the degree of variation from study to study. Looking at the overall average across participants, the NAP effect size is near the upper boundary of the scale, and the SMD_W and LRRi effect size both suggest strong effects. In contrast, the average $PoGO_M$ effect size of 55.8 suggests a more moderate effect of moving outcomes less than 60% of the way towards goal levels. Furthermore, the average NAP values are quite similar across studies, as are the study-level average $PoGO_M$ effect sizes. In contrast, the study-level SWM SMD_W and LRRi values suggest that the effects might be more variable across the three studies.

Table 10.2: Study-Level Average and Overall Average Effect Sizes Across Studies Based on Fixed Effects Models with Equal Weights

		NAP		SMD~W~		LRRi	
Study	N (Cases)	Est. (SE)	95% CI	Est. (SE)	95% CI	Est. (SE)	95%
Byiers et al. (2014)	3	.95 (.03)	[.89, 1.02]	3.90 (0.76)	[2.42, 5.38]	1.36 (0.23)	[0.90
Casey (1978)	4	.89 (.03)	[.83, .96]	1.89 (0.30)	[1.30, 2.47]	0.48 (0.05)	[0.38
Strasberger & Ferreri (2014)	3	.97 (.03)	[.91, 1.03]	5.36 (0.88)	[3.65, 7.08]	1.74 (0.16)	[1.41
Overall Average	11	.94 (.02)	[.90, .97]	3.70 (0.39)	[2.93, 4.47]	1.18 (0.09)	[1.00

10.5.2 Random Effects Meta-Analysis

In summarizing the results across a very limited number of cases and studies, researchers might be wary of drawing any inferences beyond the included participants, and so will use fixed effects meta-analysis. However, in larger-scale research synthesis projects, researchers might have an explicit goal of drawing broader generalizations about the effects of a class of interventions or practices for a specific population of participants. Furthermore, researchers may be interested in understanding not only the average effects of intervention, but also the extent of variation in those effects across different participants and study contexts and whether such variation can be explained by features of the participants, intervention, setting, or other aspects of the study context. For such purposes, a random effects model provides a useful analytic approach.

Random effects models start from the premise that the studies included in a meta-analysis represent a broader population of potential studies or potential contexts in which an intervention might be used, where the goal is to draw inferences about this population. Consequently, the SEs and CIs from random effects models represent the uncertainty in average effect sizes due to having only a limited sample from the population of potential studies. Compared to fixed effects models, SEs based on random effects models will typically be larger, and CIs will be wider, because they account for further sources of uncertainty.

For synthesizing case-specific effect size indices such as NAP, SMD_W, LRRi, and $PoGO_M$, it is natural to use a hierarchical random effects model that describes an overall average effect, the degree of variation in average effect sizes across studies, and the degree of variation in effect sizes across participants nested within studies (**??**). The study-level variation in effect size is described by the standard deviation of average effects across the hypothetical population of studies. We will denote this standard deviation as τ. The participant-level variation in effect size is described by the standard deviation of the effects across the participants within each study. We will denote this standard deviation as ω. Typically, the degree of within-study variation is assumed to be homogeneous from study to study, so that ω represents the typical, or pooled, standard deviation in the hypothetical population of studies. The overall degree of variation in effect size across participants and across studies can then be described by the

total heterogeneity, $\sqrt{\tau^2 + \omega^2}$. Using the metafor package (**?**), we can estimate these heterogeneity values using restricted maximum likelihood methods.

For estimating an overall average effect size using the hierarchical random effects model, the conventional approach is to take an inverse-variance weighted average. However, inverse-variance weighting is less appropriate when effect size estimates are based on very limited numbers of observations and when the SE of the effect size is a function of the magnitude of effect size. **?** evaluated different approaches to meta-analyzing several case-specific effect size indices, including NAP, SMD_W, and LRRi, in an extensive simulation study. They found that using an inverse-variance weighted average was appropriate for LRRi (and LRRd) effect size indices, but not for NAP or for SMD_W. For the latter two indices, the strong connection between the SE and the magnitude of effect size led to bias when inverse-variance weighting was used for estimation of overall average effects. **?** also found that using simple weighted averages led to less biased estimates, and although they did not evaluate meta-analyses of $PoGO_M$, the fact that its SE is related to the magnitude of effect size suggests that a similar pattern may hold. We therefore use simple weighted averages for summarizing NAP, SMD_W, and $PoGO_M$ based on the hierarchical random effects model. As a result, the summary effect sizes here are identical to those based on the fixed effects model-only the SEs and confidence intervals differ. For LRRi, we follow the recommendations from **?** and use inverse-variance weighting.

A further complication involved in synthesizing case-specific effect sizes is that the SEs of the effect size estimates assume that each observation is independent. If there is autocorrelation in the data series, where observations that are more closely spaced in time tend to be more predictive of future observations, this assumption will be violated, and the calculated SEs will tend to understate the true degree of uncertainty. However, **?** found that this problem can be mitigated by using robust variance estimation methods for calculating SEs and CIs. Robust SEs and CIs provide accurate quantifications of uncertainty in overall average effect size estimates even for autocorrelated data series. When summarizing results for each of the four effect size metrics, we therefore report robust SEs and robust 95% CIs for overall average effect sizes. The robust SEs and robust CIs incorporate small-sample corrections (**??**) so that they perform well even when based on a very limited number of studies.

Finally, we also report prediction intervals as a further method of describing the degree of heterogeneity in effect sizes. A prediction interval is an estimate of the range of effect sizes observed across a specified part of the hypothetical population of studies and participants. Another way to think about this is that the prediction interval is a range in which we would expect to observe the effect size for a new study with a new participant drawn from the population[5].

[5]Note that prediction intervals and confidence intervals pertain to different inferences. Confidence intervals are interval estimates that will tend to cover the true average effect size in a population, whereas prediction intervals are interval estimates that will tend to cover the effect size for a new individual participant drawn from the population. Prediction intervals

Table 10.3: Overall Average Effect Sizes and Heterogeneity Estimates Based on Hierarchical Random Effects Models with Robust Variance Estimation

	NAP (equal weighting)		SMD~W~ (equal weighti	
Parameter	Est. (SE)	95% CI	Est. (SE)	95% CI
Overall average effect size	.94 (.03)	[.82, 1.05]	3.70 (1.14)	[-1.44, 8.8
Between-study SD (τ)	.02		.88	
Within-study SD (ω)	.00		2.01	
Total heterogeneity $\Big(\sqrt{\tau^2 + \omega^2}\Big)$	.02		2.19	
80% Prediction Interval	[.88, 1.00]		[-.62, 8.02]	

If a prediction interval is quite wide, it means that we expect the population distribution of effect sizes to be heterogeneous. Such prediction intervals are an especially helpful approach to describing heterogeneity because they can be transformed into different scales, such as by transforming from the LRR scale to percentage change. For each effect size metric, we calculated 80% prediction intervals to describe the range of effect sizes that we would expect to observe in the middle 80% of the population.

Applying the hierarchical random effects model to data from a small number of studies will tend to produce estimates of average effects and heterogeneity SDs that are imprecisely estimated and, therefore, must be interpreted cautiously. Although our example dataset includes just three studies and 11 cases, we will report and interpret the model estimates to illustrate the unique insights to be gleaned from the random effects approach. However, we caution readers not to draw any substantive conclusions about the effects of AAC interventions from this small illustration.

Table **??** reports the average effect size, between-study heterogeneity, within-study heterogeneity, and total heterogeneity estimates, along with 80% prediction intervals, for each of the four effect size metrics. SEs and CIs are based on robust variance estimation methods. The average NAP effect size across all eleven cases is .94 with an SE of .03, a small degree of between-study heterogeneity (τ estimate = .02), and no within-study heterogeneity. Because there is little estimated heterogeneity, the 80% prediction interval is narrow, ranging from .88 to 1.00. This suggests that a large majority of participants in the population would have strong NAP effect sizes.

The average SMD_W effect size across all 11 cases is 3.70, identical to the fixed effects meta-analysis, but the SE is 1.14, much larger than in the fixed effects meta-analysis. The larger SE arises because there is substantial heterogeneity estimated both between and within studies. The high degree of heterogeneity is also apparent in the 80% prediction interval, which ranges from less than 0 to over 8 standard deviations.

will therefore tend to be wider than confidence intervals with the same specified coverage level.

The average LRR effect size across all 11 cases is 1.06 with an SE of 0.36, which corresponds to an average 188% increase from baseline to intervention, 95% CI [-39%, 1266%]. The LRR effect size appears to be heterogeneous both between studies and across participants within studies. The 80% prediction interval of [-0.38, 2.50] corresponds to percentages changes ranging from -32% (i.e., an iatrogenic effect) to 1113%. Based on this distribution of LRR effect sizes, the intervention appears to be moderately effective, on average, but also quite variable in terms of the degree of improvement for individual participants.

Finally, the average PoGO_M across all 11 cases is 55.8, identical to the fixed effects meta-analysis, with an SE of 1.14. Unusually, the robust SE based on the random effects meta-analysis is smaller than the SE from the fixed effects meta-analysis[6]. Although the average PoGO_M values appear quite consistent across studies, the estimated within-study SD of $\omega = 14.1$ and 80% prediction interval of [30.6, 81.1] indicate that there is some variation in PoGO_M values across the participants in each study.

10.5.3 Further Directions for Synthesizing Case-Specific Effect Sizes

The examples of fixed effects and random effects meta-analysis of case-specific effect sizes that we have presented here are only a starting point. In addition to summarizing findings across multiple participants and studies and providing quantitative estimates of the degree of heterogeneity in findings, meta-analysis techniques are available for investigating many further questions. One important further direction is to explore how features of a study's design, intervention characteristics, or participant profiles explain variation in effect size magnitude. Such questions can be investigated using tools such as subgroup analysis, meta-analytic analysis of variance, and meta-regression analysis (**?**, Chapters 19-21; **?**; **?**), which is the meta-analytic analogue of multiple regression analysis for primary study data.

Another further direction is to explore whether selective reporting of primary study findings could create biases in meta-analytic summary estimates. Selective reporting and publication biases are critical concerns in synthesis of group design studies (**?**) and there is good reason to anticipate that similar biases may also affect syntheses of case-specific effect sizes in single-case research (**????**). Although many methods are available for detecting and correcting the biases created by selective reporting, nearly all available tools have been developed in the context of group design studies, where primary analyses are based on inferential statistical approaches. These tools have been applied in syntheses of single-case research as well, but they may not be as informative in this context because of reliance on visual analysis (rather than statistical inference). There

[6] We suspect that this is because the robust standard error is based on only a few studies. When applied to results from a larger collection of studies, robust standard errors will tend to be larger than fixed effects standard errors when the population distribution is heterogeneous.

remains an outstanding need to develop methods adapted to the context of single-case research (**??**).

More broadly, tools for synthesis of single-case designs remains an active area of methodological development, with many recent statistical innovations that are beyond the scope of the present guide. In light of the rapid pace of methodological developments, we especially encourage researchers conducting syntheses of single-case designs to seek collaborations with methodologists working in the area. Such collaborations can not only improve the methodological rigor of single-case syntheses, but also spur further methodological developments to improve existing methods, making them more relevant and better suited for empirical application. In turn, researchers will be able to better apply the tools of synthesis to inform theory, policy, and practice on topics where single-case research is prevalent.