NORTHWESTERN UNIVERSITY

Operationally Comparable Effect Sizes for Meta-Analysis of Single-Case Research

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Statistics

By

James E. Pustejovsky

EVANSTON, ILLINOIS

June 2013

ABSTRACT

Operationally Comparable Effect Sizes for Meta-Analysis
of Single-Case Research

James E. Pustejovsky

This thesis studies quantitative methods for summarizing and synthesizing single-case studies, a class of research designs for evaluating the effects of interventions through repeated measurement of individuals. Despite long-standing interest in meta-analytic synthesis of single-case research, there remains a lack of consensus about appropriate methods, even about the most basic question of what effect size metrics are useful and appropriate. I argue that operational comparability, or invariance to heterogeneous operational procedures, is crucial property for an effect size metric. I then consider two problems with operational comparability that arise in single-case research. The first problem is to find effect sizes that can be applied across studies that use different research designs, such as single-case designs and two-group randomized experiments. The second problem is to find effect sizes that can be applied across studies that use varied operations for measuring the same construct. To address each of these problems, I propose structural models

that capture essential features of multiple relevant operations (either design-related operations or measurement-related operations). I then use these structural models to precisely define target effect size parameters and to consider identification issues and estimation strategies.

Chapter 1 defines operational comparability and situates the concept within the broad methodological concerns of meta-analysis, then reviews relevant features of single-case research and previously proposed effect sizes. Chapter 2 describes an abstract set of modeling criteria for constructing design-comparable effect sizes. Chapters 3 applies the general criteria to the case of standardized mean differences and proposes an effect size estimator based on restricted maximum likelihood. Chapter 4 presents several applications of the proposed models and methods. Chapter 5 proposes measurement-comparability model and defines effect size measures for use in studies of free-operant behavior, one of the most common classes of outcomes in single-case research. Chapter 6 extends the proposed effect size models to incorporate more complex features, including time trends and serial dependence, and studies a method of estimating those models through a combination of marginal quasi-likelihood and Gaussian pseudo-likelihood estimating equations. Chapter 7 collects various further extensions, areas for further research, and concluding thoughts.

Acknowledgements

I have been fortunate to receive guidance and support from a number of people in the course of completing this work. I would like to thank the members of my committee, Larry Hedges, Tom Cook, and Noelle Samia, for agreeing to oversee this project. I have benefited especially from Larry Hedges' thoughtful mentorship and encouragement, which extended far beyond purely statistical matters to include lessons in the many aspects of academic life. Tom Cook's seminar on research design was among the most valuable classes that I have ever taken, and I credit it–and him–for giving me the intellectual tools to assess research designs as a whole, rather than just the statistical details in isolation. I am grateful as well to Bruce Spencer and James Spillane for taking a chance on an under-qualified student and offering me opportunities that led to Northwestern's graduate program.

Special thanks are due to Will Shadish for many, many helpful discussions and for inspiring me by his tenacity and boundless curiosity. This thesis would not have been conceived without his agenda-defining work on single-case research methods. Thanks also to Kristynn Sullivan, Austin Mulloy, and David Rindskopf for sharing insights about single-case research, to fellow dissertators Kelly Hallberg, Christina LiCalsi-Labelle, and Francie Streich for providing feedback on draft chapters, and to Beth Tipton and Patti Ferguson for offering counsel and perspective throughout my graduate career.

Finally, I am forever grateful to my wife Mary for her love, support, and fortitude throughout the past five years. Whatever accomplishment this work represents, it is ours rather than mine.

Dedication

For my parents, John and Susan Pustejovsky

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Operational comparability and single-case research

This thesis considers quantitative methods for synthesizing single-case research. Single-case research is both a set of research methods and a body of empirical research that applies those methods. In the former sense, single-case research is a set of designs and procedures for evaluating the effects of interventions, practices, or programs on individual cases. These single-case designs (SCDs) have in common the use of repeated outcome measurements on each case, the deliberate manipulation of the treatment, and the use of each case as its own control. In the latter sense, empirical single-case research appears in many areas of psychology and education, particularly in special education, school psychology, clinical psychology, psychotherapy, social work, and applied behavior analysis (Horner et al., 2005; Kazdin, 2011; Kennedy, 2004). Methodologically similar research also appears in medicine and public health (Gabler, Duan, Vohra, & Kravitz, 2011).

By nature of the research methods employed and how study results are analyzed, single-case research emphasizes individual change. Single-case designs intend to identify individual treatment effects through comparison of outcomes measured on the same individual at different points in time. Studies often report separate results for each case, with little emphasis on overall averages across cases. This ideographic orientation presents a trade-off: while study designs may be tailored so that their results are highly relevant to the individual cases involved, each study provides meager evidence for drawing generalized inferences.

Meta-analytic synthesis has long been considered as an approach for summarizing and generalizing from single-case studies. Just as other fields began to take growing interest in meta-analysis, Gingerich (1984) argued that synthesis of single-case research could improve the precision of individual treatment effect estimates, bolster the internal validity of single studies through replication, and provide a means for studying variation in treatment effectiveness and generalizing from a collection of studies. Others noted the drawbacks of excluding single-case studies from comprehensive syntheses (e.g., Allison & Gorman, 1993; Scruggs & Mastropieri, 2012). In response, several effect size statistics and meta-analytic approaches were developed specifically for single-case research (Busk & Serlin, 1992; Center, Skiba, & Casey, 1985; Gorsuch, 1983; Scruggs, Mastropieri, & Casto, 1987).

The past two decades have seen a growing emphasis across the fields of education, medicine, and social policy on tying professional practice more closely to empirical research. This emphasis has spurred interest among scholars and policy-makers in using quantitative synthesis methods to identify evidence-based practices and programs (Shavelson & Towne, 2002; Slavin, 2008). In response, fields that use single-case research have articulated standards of scientific evidence and attempted to codify synthesis methods (Chambless & Ollendick, 2001; Gast, 2010; Horner et al., 2005; Kratochwill & Stoiber, 2002; Odom et al., 2005). As interest has grown, syntheses of single-case research now appear with increasing frequency (Maggin, O'Keeffe, & Johnson, 2011). Furthermore, large research synthesis projects such as the What Works Clearinghouse (WWC) have recently broadened the scope of their evidence standards to include evidence from single-case research (Kratochwill et al., 2012).

Despite long-standing interest among single-case researchers and increased attention due to the evidence-based practice movement, there is little consensus regarding how single-case studies should be synthesized. Even the most basic question of what effect size metric to use for meta-analysis remains unresolved, though proposals have proliferated (for reviews of candidate effect size metrics, see Beretvas & Chung, 2008b; Wolery, Busick, Reichow, & Barton, 2010). Existing approaches are nearly all subject to serious conceptual or technical criticisms (Shadish, Rindskopf, & Hedges, 2008), a situation that led the authors of the WWC pilot standards to refrain from recommending any specific effect size metrics or particular statistical approaches to analysis of single-case data.

This thesis addresses two problems related to effect size definition and meta-analysis of single-case research. The first problem is to find effect sizes that can be applied across studies that use different research designs, such as single-case designs and two-group randomized experiments; I term such effect size metrics *design-comparable*. The second problem is to find effect sizes that can be applied across studies that use varied operations for measuring the same construct, such as different tests of math achievement or different procedures for measuring the prevalence of self-injurious behavior; I term such effect size metrics *measurement-comparable*. Both problems can be understood as specific dimensions of *operational comparability*, or the extent to which a metric can be interpreted in terms of scientifically meaningful constructs, across multiple, heterogeneous instances of study operations. To address each of these problems, I propose structural models that capture essential features of multiple relevant operations (either design-related operations or measurement-related operations). I then use these structural models to precisely define target effect size parameters and to consider identification issues and estimation strategies.

The remainder of this chapter provides additional context regarding operational comparability and single-case research. In the next section, I explain the motivation for seeking operationally comparable effect sizes, noting other areas of meta-analysis where similar problems appear. In the following section, I survey the characteristics of single-case research that are relevant to the problems at hand, including the major types of single-case designs and common approaches to outcome measurement. I then review existing proposals for single-case effect sizes and meta-analysis. In the final section, I outline the broad organization of this thesis.

1.1. Operational comparability

One of the central questions in any quantitative research synthesis is how to operationally define the effect size metric, the basic unit of analysis in a meta-analysis (Cooper, 2009). In some cases, the choice of effect size may be straight-forward. If the empirical evidence to be synthesized is confined to a specialized field or a specific, narrow topic, research practices may be homogeneous to such an extent that the choice of effect size becomes a matter of disciplinary conventions. Yet a research synthesis may have a more ambitious aim and broader scope than to synthesize narrowly delineated sets of studies. For syntheses that seeks to combine evidence across varied or complex study designs, based on disparate operational procedures, or across diverse fields, operational definition of an effect size becomes more challenging.

The choice of effect size metric involves, implicitly or explicitly, an assumption about the comparability of results from different studies that may use various participant inclusion criteria, treatment procedures, outcome measurement instruments, or experimental

designs (Hedges, 2008). An appropriate operational definition of effect size is therefore crucial for maintaining the construct validity of the synthesis–that is, the extent to which the operations employed in individual studies can be taken as measures of a common underlying phenomenon and can be meaningfully compared.[1] In synthesizing results from studies that use different operational procedures, one would like to use an effect size that is on the same metric across all of them; I term such an effect size metric *operationally comparable*.

Operational comparability is essential in that it allows the meta-analyst to control for incidental characteristics related to study procedures and to focus instead on variation that is of scientific interest (cf. Rubin, 1992). Without operational comparability, a collection of effect sizes will exhibit heterogeneity due merely to procedural differences in how the study was carried out. For instance, imagine a set of studies that are exact replications–including using samples of units from the same population–except for one aspect of the studies' procedures. Unless the effect size metric used to summarize the studies is operationally comparable with respect to that procedural aspect, the results will differ by more than would be expected due to sampling variation alone. In the more realistic circumstance that studies differ along many different dimensions, lack of operational comparability will

[1]The construct validity of an individual study is an assessment of the extent to which the theoretical constructs employed in formulating the research question are appropriately represented by the specific operations employed (Shadish, Cook, & Campbell, 2002). The construct validity of a research synthesis goes further, because a meta-analyst must classify not just one but multiple and diverse studies according to whether their operations fall under common construct domains. In this sense, the inclusion and exclusion criteria of a meta-analysis constitute an operational definition of the relevant constructs referenced by the research question at hand. The effect size metric chosen for a synthesis provides the scale on which results of different studies are weighed; the assumption is not only that the various study operations all fall under relevant constructs, but that the resulting effect size estimates are comparable *on the chosen metric*.

tend to obscure substantive differences among study results and will reduce not only the precision of meta-analytic summaries, but their basic interpretability.

Complete operational comparability may seem an impossible standard to achieve given the array of procedural decisions that must be made in any empirical study. In practice, meta-analysts use effect size metrics that are operationally comparable for only the most salient features of a collection of studies. Take, for example, the standardized mean difference, a ubiquitous effect size for measuring a difference between two groups. The standardized mean difference is the ratio of the difference in mean outcomes between two groups to the standard deviation of one of the groups (or of both, if their variances are assumed to be equal). A strong rationale for using the standardized mean difference rather than the equally ubiquitous p-value as a measure of group differences is that only the former is operationally comparable across studies employing different sample sizes or treatment group allocations (Borenstein, 2009).

A further example can be found in syntheses of education- and employment-selection test validity studies. In this field, data often exhibit range restriction as a result of some selection process, such as when an aptitude test is given to a sample of job applicants but job performance measures are available only for those who are subsequently hired. Corrections for range restriction are applied because scientific interest is in the correlation between the test and performance measures in the unrestricted population (Hunter, Schmidt, & Le, 2006; Mendoza & Mumford, 1987). The unrestricted correlation is the operationally comparable effect size metric, to be estimated across studies that use various selection criteria.

Inevitably, one must rely on a set of modeling assumptions in order to establish the operationally comparability of an effect size metric across variations of a particular operational feature. If those modeling assumptions do not hold, then the effect size metric may not be operationally comparable, or may be only approximately so. For example, the correction for range restriction of correlation coefficients relies on the assumption that the two variables are bivariate normally distributed in the unrestricted population; if this assumption does not hold, the corrected effect size statistic may be inaccurate even in large studies, with a sampling distribution that depends on population characteristics other than the bivariate correlation.

In circumstances where modeling assumptions do not hold or where no model for operational comparability can be found, the meta-analyst might turn to meta-regression techniques to explain variation in effect sizes due to study procedures. This strategy can provide useful insights about the methodological assumptions and practices employed in a field (c.f. Shadish & Ragsdale, 1996; Shager et al., 2012). Still, lacking a model of operational comparability, results that depend on arbitrary procedural characteristics remain difficult to interpret in terms of the substantive scientific questions that motivate a research synthesis. I now consider specific dimensions of operational comparability related to study design and to outcome measurement procedures.

1.1.1. Design-comparability

The design of a study can be understood as the planned pattern of measurements: which variables are measured at which times on which units; in experimental designs, this includes how units are selected (i.e., sampling issues) and assigned to treatment conditions.

The need to synthesize studies that use different designs arises in many areas of research synthesis. The following examples highlight areas in which design-comparability problems arise:

- For psychological experiments, Morris and DeShon (2002) discuss different effect size metrics that can be used to summarize single-group or two-group pre-test/post-test designs, emphasizing comparisons between such designs and between-groups designs that use only a post-test.
- In correlational studies, meta-analysts occasionally encounter designs (such as extreme groups designs) that do not yield conventional correlation coefficients yet still provide information about the bivariate association between two variables (Preacher, Rucker, MacCallum, & Nicewander, 2005). In order to synthesize such studies, one must convert reported effect sizes into the same metric as the correlation coefficients that are generated by a conventional, bi-variate sampling design (Pustejovsky, 2012).
- Though the standardized mean difference is an unambiguous parameter in simple between-groups designs, many field experiments use randomized-block or cluster-randomized designs in which multiple variance components are identified. In such designs, several different standardized mean difference effect sizes might be defined, necessitating careful consideration of their design-comparability (Donner & Klar, 2002; Hedges, 2007, 2011).
- In epidemiology, various study designs are used to measure the effect of a binary exposure on a binary outcome measure, including prospective cohort designs and case-control designs. In contrast to metrics such as risk ratios or risk differences,

odds ratios allow a direct comparison between results of these designs, at least under certain modeling assumptions (Fleiss & Berlin, 2009).[2]

- In medical trials, meta-analysts may need to combine evidence from parallel-group trials with that from cross-over trials, in which each unit receives multiple treatment conditions at different points in time (Elbourne et al., 2002). Curtin, Altman, and Elbourne (2002a, 2002b) study the comparability of various effect size metrics for syntheses that include both designs. Others have developed models and methods for combining data from varied, often idiosyncratic trial designs in specific areas of application (e.g., Frost, Clarke, & Beacon, 1999; Tvete et al., 2012).

Although Sutton and Higgins (2008) characterize combination of evidence from multiple types of study designs as "complex synthesis" (p. 633) in order to draw a contrast with conventional meta-analysis, design-comparability is nonetheless an important consideration in many areas of application.

The design-comparability of an effect size estimate is distinct from its internal validity. In studies of causal treatment effects, an effect size might be design-comparable under a certain model yet internally invalid if that model does not adequately account for confounding factors. Conversely, a treatment effect size estimate may be internally valid yet lack design-comparability with a between-groups randomized experiment. Consequently, if a meta-analysis seeks to evaluate empirically the internal validity of a research design by comparing results to randomized experiments in the same field (e.g., Shadish & Ragsdale, 1996), design-comparable effect sizes are required in order to separate any bias of

[2]For a critique of odds ratios as a design-comparable metric, see Greenland (1987).

treatment effects from differences in metric. Without design-comparability, the effect size metric itself will confound true differences in internal validity across study types.[3]

In single-case research, the WWC standards highlight the need for design-comparable effect sizes, noting in particular that "an [effect size] estimator for SCDs that is comparable to those used in traditional group studies is badly needed" (Kratochwill et al., 2012, p. 10).[4] The need for design-comparable effect sizes is further demonstrated by potential for application. There exists evidence from both single-case and between-groups designs on a number of topics, including phonological awareness training programs (What Works Clearinghouse, 2012), repeated reading interventions (Chard, Ketterlin-Geller, & Baker, 2009; O'Keeffe, Slocum, Burlingame, Snyder, & Bundock, 2012), reading fluency interventions (P. L. Morgan & Sideridis, 2006), writing interventions (Graham & Perin, 2007; Rogers & Graham, 2008), positive behavioral interventions and supports (Bradshaw, Mitchell, & Leaf, 2009; Horner et al., 2009; Marquis et al., 2000), and picture exchange communication systems (Hart & Banda, 2009). Past syntheses on such topics have either reported separate meta-analyses for each type of design or limited their scope to only one type of design. Design-comparable effect sizes are required in order to combine evidence from both types of designs. I describe a general approach to defining design-comparable effect sizes for single-case designs in Chapter 2, then apply this approach to standardized mean differences in Chapter 3.

[3]Questions about the design-comparability of effect sizes are less crucial when one can make within-study comparisons of designs, particularly when outcomes can be measured using the same procedures for all participants (Cook, Shadish, & Wong, 2008). This is one reason that empirical studies of research design might privilege within-study comparisons over across-study meta-analysis.

[4]Horner, Swaminathan, Sugai, and Smolkowski (2012) also note the need for design-comparable effect sizes, though they define comparability in a restrictive sense as being relative to the standardized mean difference.

1.1.2. Measurement-comparability

One of the primary difficulties in operationally defining an effect size is that studies in a collection to be synthesized often use a variety of different measurement instruments, such as different measures of spatial ability (e.g., Uttal et al., 2013) or reading achievement tests from different states (e.g., Reschly, Busch, Betts, Deno, & Long, 2009). To account for differences in measurement instruments, it is desirable that the magnitude of an effect size not depend on the units of the measurement instrument. However, simply because an effect size is unit-free does not imply that it is measurement-comparable. Rather, measurement-comparability is addressed by a theory of the relationship between different scales, perhaps expressed formally as a statistical model. I note three examples of such theories.

First, consider again the standardized mean difference. Because both the mean difference (numerator) and standard deviation (denominator) are in the same units, their ratio is unit-free. The standardized mean difference can be understood as a measurement-comparable effect size for linearly-equatable, interval scale measures (Hedges & Olkin, 1985).[5] That different instruments produce linearly equatable measures may seem a rather tenuous theory, but in many circumstances, lack of further information prohibits the use of any more elaborate model.

Next, methods for converting among different effect size metrics have been proposed that rely on explicit theories of measurement-comparability. When groups are based on dichotomization of a continuous, normally distributed variable, the standardized mean

[5]The standardized mean difference can also be justified as a more general scale capturing distributional overlap (Hedges & Olkin, 1985).

difference between the groups can be transformed to the same metric as the Pearson correlation for measuring the bi-variate association between the outcome variable and the underlying continuous variable on which the groups are based (Hunter & Schmidt, 1990). Similarly, methods exist for converting odds ratios for binary outcomes into standardized mean differences when the binary outcomes are based on dichotomization of a continuous, latent scale (Chinn, 2000; Dominici, Parmigiani, Wolpert, & Hasselblad, 1999; Sánchez-Meca, Marín-Martínez, & Chacón-Moscoso, 2003).

Finally, Hunter and Schmidt (2004) have proposed methods for correcting correlation coefficients and standardized mean differences based on the measured reliability of the outcome measures. These reliability correction procedures are motivated by an implicit model of measurement-comparability. For instance, one set of assumptions that justify reliability corrections for the standardized mean difference are 1) that groups differ on latent constructs that are imperfectly measured by the outcome variables and 2) that outcome variables differ on the magnitude of their unique error variances.

Issues of measurement-comparability have been little explored in the context of single-case research. As described further in Section 1.2.2 and in Chapter 5, the outcome measures in single-case studies are often intended to capture some aspect of an individual's behavior. Though several different procedures are commonly employed to measure behavior, previous work has addressed neither the comparability of measurements across procedures nor the implications for effect sizes and meta-analysis. In Chapter 5, I propose a measurement model for one domain of behavioral measures (free-operant behavior) and use that model to define and estimate effect sizes that are directly comparable across measurement procedures.

1.2. Features of single-case research

Models for the operational comparability of effect sizes must be appropriate and reasonable for the areas of application. Thus, to motivate the development of design-comparability and measurement-comparability models in later chapters, this section describes relevant characteristics of single-case research. I first survey the main types of single-case designs because later analyses of design-comparability must be grounded in the specific structures and internal logics of the designs. I then discuss the types of outcomes and measurement procedures that are used in empirical work, in order to provide a sense of the range of operations that measurement-comparability models will need to encompass.

1.2.1. Varieties of single-case designs

Single-case research methods comprise several related designs, including multiple baseline, treatment reversal, multiple (or alternating) treatment, and changing criterion designs (Kazdin, 2011; Shadish et al., 2002). All of these designs share three characteristics: (1) repeated measurement of an outcome over time on a set of individual cases; (2) planned, deliberate manipulation of treatment assignment by the investigator; and (3) identification of treatment effects through comparison of each individual's outcomes across different points in time (Horner et al., 2012). Beyond these defining features, the designs differ in the patterns of treatment assignment within and across cases, with consequences for the circumstances under which each design is feasible and appropriate.

The multiple baseline design is the most common and perhaps most conceptually straight-forward SCD. The multiple baseline design involves multiple individual cases,

on each of which an outcome is measured repeatedly during an initial baseline phase; a new treatment is then introduced and remains in place while outcome measurement is continued. The treatment is taken to have an effect if the stable pattern of outcomes differs between baseline and treatment phases and if the change coincides with the introduction of treatment. By staggering the point at which treatment is introduced for different cases, the design rules out certain alternative explanations for observed changes between phases (Kazdin, 2011; Shadish et al., 2002). Multiple baseline designs can also be understood as collections of interrupted time series (Shadish et al., 2002). The single-case design literature describes several types of multiple baseline designs, including multiple baselines across individuals and multiple baselines across settings (Kazdin, 2011). In some instances, multiple baseline designs also employ control cases that do not receive treatment at all (e.g., Musser, Bray, Kehle, & Jenson, 2001).

Treatment reversal designs involve repeatedly introducing and then removing a treatment. In the most common form, the ABAB design begins with a phase of baseline measurement (A), followed by a treatment (B) phase, a removal of the treatment (A), and a re-introduction of treatment (B). Clearly, such designs are limited to treatments that can feasibly be removed. In many situations, the design is employed under the assumption that treatment effects are transitory, so that removing the treatment will lead to a return to baseline conditions. Horner et al. (2005) noted that, when treatment effects are transitory, repeated introduction and removal provides "within-subject replication," which strengthens the internal validity of the inference that the treatment has an effect.

Multiple treatment designs compare two or more treatments by alternating between treatment conditons over short periods of time, such as over multiple sessions on the same

day (morning versus afternoon) or over different, commonly occuring stimulus conditions (e.g., lining up for recess, preparing for snack time). The term "multiple" refers to the use of multiple treatments within one phase of the study, rather than in separate phases. Often, the design begins with a baseline phase in which an outcome is measured without intervention. Following this, the alternating treatment phase is implemented; on the basis of the data from this phase, the researcher might choose to continue only one intervention into a further phase. Compared to the other types of SCDs, multiple treatment designs are useful only under more restrictive circumstances. As discussed in Kazdin (2011, Chp. 9), the design is most useful in studying treatments with rapid onset and rapid dissipation. Since different treatment conditions are implemented in rapid succession, the possibility of carry-over effects or interference between treatments may threaten the internal validity of inferences from multiple treatment designs. One important application of this design is to conduct functional assessment of problem behaviors (Hanley, Iwata, & McCord, 2003), a diagnostic procedure that seeks to measure the effects of different categories of antecedent conditions and contingent consequences on the behavior of an individual in order to prescribe an appropriate intervention.

Finally, changing criterion designs are unique in that the intervention is defined in terms of the performance level on a target outcome, and typically involves a consequence or incentive for the participant to meet a performance criterion (Kazdin, 2011). As the participant meets criterion, the criterion is shifted to a more stringent standard. The relationship between intervention and outcome is demonstrated by a close correspondance between the performance criterion and the level of the outcome.

Randomization of treatment assignment can be introduced into any of these designs, though doing so is rare in most areas of application (Kazdin, 2011). Kratochwill and Levin (2010) illustrated how random starting points can be determined for use in the multiple baseline design, and how random assignment of phases can be incorporated into the $(AB)^k$ design. By analogy to between-subject randomized trials, they argued that randomization improves the scientific credibility of the designs. Random assignment is somewhat more common in medical applications of alternating treatment designs, in the form of either completely random assignment or randomization within phase pairs (Gabler et al., 2011; Guyatt et al., 1990, 1988; Larson, Ellsworth, & Oas, 1993). Guyatt et al. (1988) established limited conditions in which randomized treatment reversal designs are appropriate, namely when treatment has rapid onset and transience, when treatment would be used for a long period of time, if effective, and where efficacy is truly in doubt. In medical and public health settings, there is a also greater emphasis on use of random assignment in conjunction with multiple baseline designs and the closely related stepped wedge designs (C. A. Brown & Lilford, 2006; Hussey & Hughes, 2007; Rhoda, Murray, Andridge, Pennell, & Hade, 2011).

In addition to the defining features of SCDs, it is also important to note characteristics of the designs as applied in practice, as these empirical features are also relevant to the application of statistical methods. Huitema (1985), Busk and Marascuilo (1988), Hammond and Gast (2010), Shadish and Sullivan (2011), and Smith (2012) have surveyed the design characteristics of published SCDs in the behavioral and social sciences.[6] Shadish

[6]I cite the results of Shadish and Sullivan (2011) throughout because the other reviews from the past decade were either not as detailed (Hammond & Gast, 2010) or appear to have included some extreme outliers (Smith, 2012).

and Sullivan (2011) described a census of SCDs published in 21 journals during 2008; the authors identified over 100 studies containing over 400 individual cases. On average, each study reported 3.64 cases; nearly 75% of the studies used three or more cases. The availability of multiple cases per study allows for identification of both within- and between-case variation in the outcome, which is crucial for identifying design-comparable effect sizes. Shadish and Sullivan (2011) also reported that SCDs tend to measure a small or moderate number of observations per case. They found that each case contained an average of 27 observations across multiple phases; furthermore, only 55% of relevant cases contained five or more observations in the baseline phase. Such short phase lengths make it challenging to rely on the data alone in order to determine appropriate models for the outcome process. Therefore, statistical analysis of SCD data will in many instances depend on relatively strong prior assumptions about the baseline process.

1.2.2. Outcome measurement

To guide development of measurement-comparability models, it is useful to first understand the types of outcomes used in empirical single-case research, both at the level of broad construct domains and in terms of operational procedures. Kazdin (2011) categorized outcome variables used in single-case research according to whether they are based on observation of overt behavior, on psychophysiological assessments, on rating scales, or on specific features of a target behavior. He noted that the vast majority of single-case studies use observation of overt behaviors. Indeed, the focus on outcome measurements based on direct observation is considered a hallmark of single-case methodology, in that treatment impacts on behavioral outcomes often have immediate and recognizable social

implications for individual participants and the broader populations that they represent (Hartmann & Wood, 1990; Horner et al., 2005).

Within the domain of overt behavior, I will distinguish further between behaviors in free-operant versus restricted-operant contexts. *Free-operant contexts* are defined by a setting or time-frame in which behaviors are free to occur at any time, without prompting or restriction by the investigator.[7] For instance, an investigator might observe the bullying behavior of a child during lunch recess (e.g. Ross & Horner, 2009), recording incidents as they occur over the course of the child's natural interactions with her peers. In contrast, *restricted-operant contexts* are defined by a specific behavioral stimulus, often under the control of the investigator. For instance, an investigator might observe whether or not a child complies with verbal requests (e.g. Zuluaga & Normand, 2008); because compliance is contingent on the investigator's requests, such observations are in a restricted-operant context.

A sense of the relative frequency with which studies measure different outcome domains can be gleaned from the database of single-case studies published in 2008, as assembled by Shadish and Sullivan (2011). Using the categories described by Kazdin (2011), I classified each study by the general domains of outcomes employed. As shown in Table 1.1, the vast majority of studies measured behavioral outcomes, with 56% of studies measuring behaviors in free-operant contexts and 41% of studies measuring behaviors in restricted-operant contexts. My development of measurement-comparability models focuses on the domain of free-operant behavior because it is the most common.

[7]The behavior-analytic tradition defines free-operant behavior slightly differently, as behavior stimulated by its consequences, rather than by antecedent prompts or cues (Johnston & Pennypacker, 1993, p. 366).

Table 1.1. Outcome domains employed in single-case studies

Outcome	Studies[a]	%[b]
Free-operant behavior	68	56
Restricted-operant behavior	50	41
Psycho-physiological	7	6
Rating scales	10	8
Other	4	3

[a] $N = 122$ single-case studies published in 2008, as identified by Shadish and Sullivan (2011).
[b] Total is greater than 100% because some studies used more than one type of outcome measure.

An array of methods are commonly used to measure overt behavior in free-operant or restricted-operant contexts. Detailed surveys of measurement methods are available from several sources (Ayres & Gast, 2010; Barlow & Hersen, 1984; Hartmann & Wood, 1990; Kahng, Ingvarsson, Quigg, Seckinger, & Teichman, 2011; Kazdin, 2011; Primavera, Allison, & Alfonso, 1996). For a given target behavior, any of these measurement methods might conceivably be applied; different investigators might choose different methods to measure the same behavior based on personal preferences or convenience. It will therefore be important to understand the comparability of these different procedures for measuring outcomes within each domain.

For measuring behavior in free-operant contexts, there are four major recording methods. *Continuous recording* measures the proportion of session time during which a behavior is observed. *Event counting measures* the number of behavioral event occurrences per session. *Momentary time sampling* measures whether or not a behavior occurs at each of a set of fixed moments in time, and is typically typically summarized using a proportion. Finally, interval recording techniques, including *partial interval recording* and *whole interval recording*, involve dividing an observation session into short time intervals, scoring each interval according to whether or not the behavior occurs for some part of

that interval, and summarizing as a proportion of intervals. These methods are described further in Chapter 5.

For measuring behavior in restricted-operant contexts, there are also four major methods. The method of *fixed-trial proportion of successes* (also called discrete trial recording) involves measuring the proportion of time that a behavior occurs, where the total number of trials is set by the researcher. For instance, an observer might measure the proportion of math problems on a test that the subject answered correctly or the proportion of chores completed without misbehaving. The method of*variable-trial proportion of successes* differs from fixed-trial methods in that the total number of trials is not under the control of the researcher, but may depend on other aspects of the subjects behavior. For instance, an observer might measure the proportion of homework problems correctly completed during a fixed amount of time, or the proportion of conflict situations in which the subject acquiesces. *Response latency* methods involve measuring the time between a stimulus or cue and the subject's response. Finally, some behaviors can be broken up into discrete steps or tasks, which may or may not be contingent on one another. Such behaviors are sometimes measured using a *task check-list*, recording which of the tasks are completed in the setting of interest. For example, investigators might evaluate how well a teacher implements a teaching technique by using a checklist consisting of the component tasks involved in correct use of the technique (e.g. Downs, Downs, & Rau, 2008).

Several methodological surveys have been conducted that provide indication of how frequently different measurement procedures are used, though none of the surveys is comprehensive. In one older survey, Kelly (1977) examined research articles published in the *Journal of Applied Behavior Analysis* (*JABA*) between 1968 and 1975. He reported

that 222 articles (76%) used direct observation of behavior to measure outcomes; of studies using direct observation, the proportion using event recording, trial scoring, interval recording, time-sampling, and response duration methods were further described. Mann, Ten Have, Plunkett, and Meisels (1991) reviewed observational studies (including group designs) published between 1980 and 1989 in the journal *Child Development*. They found that the three most common measurement procedures were continuous recording, interval recording, and (indirect) rating scales; approximately one third of the studies used an interval recording method. More recently, an informal survey of articles published in *JABA*, *Behavioral Interventions*, or *Research in Developmental Disabilities* between 2002 and 2005 identified 65 studies that used partial interval recording or momentary time sampling when measuring behaviors with non-negligible duration (Rapp et al., 2007); however, the authors did not report the proportion of all articles meeting their criteria, making their finding difficult to interpret. Finally, Mudford, Taylor, and Martin (2009) surveyed research articles published in *JABA* between 1995 and 2005 to assess whether the investigators used continuous (i.e. continuous duration recording, event-counting) or discontinuous (momentary time sampling, interval recording) methods to measure free-operant human behavior. They reported that 55% of the 168 articles identified used continuous-duration recording or event counting methods, with the remainder using discontinuous methods.

A more systematic perspective on the use of different measurement procedures can be obtained by returning to Shadish and Sullivan's database of single-case studies published in 2008 (Shadish & Sullivan, 2011). Table 1.2 displays the results of categorizing each study that included at least one measure of overt behavior according to the measurement

Table 1.2. Measurement procedures in single-case studies of overt behavior

Procedure	Studies[a]	%[b]
Free-operant behavior	68	
Continuous duration recording	7	10
Event counting	41	60
Momentary time sampling	5	7
Interval recording	13	19
Other	11	16
Restricted-operant behavior	50	
Fixed-trial proportion of successes	28	56
Variable-trial proportion of successes	7	14
Response latency	3	6
Task check-list	7	14
Other	6	12

[a] Subset single-case studies published in 2008, as identified by Shadish and Sullivan (2011).
[b] Totals are greater than 100% because some studies used more than one procedure.

procedures employed. In free-operant contexts, event counting is by far the most common procedure, followed by interval recording methods; in restricted-operant contexts, the majority of studies used fixed-trial proportion of successes, while all other procedures were far less common.

1.3. Effect sizes and meta-analysis for single-case research

Single-case research maintains a unique approach to analytic methodology. While placing a heavy emphasis on systematic collection of quantitative data, the field does not correspondingly emphasize statistical methods of data analysis. Instead, visual inspection of graphed outcome data is the dominant method of data analysis (Gast & Spriggs, 2010; Johnston & Pennypacker, 1993; Kazdin, 2011; Smith, 2012).[8] In arriving at a

[8]The prevalence of visual inspection can be attributed in part to the historical roots of single-case research in behaviorist psychology (Johnston & Pennypacker, 1993). Also, some fields use single-case research primarily as an aid in immediate diagnosis and decision-making regarding study participants; to that end,

determination of effectiveness, visual analysts are instructed to assess and weigh the level, trend, variability, immediacy, and consistency of treatment effects, as well as the extent to which data points from adjacent phases overlap (Horner et al., 2005). Visual inspection is a direct inferential technique, in that the analyst's goal is to determine whether a specific treatment has an effect on the outcome rather than to provide an estimate of the magnitude of the effect. Despite its widespread use and acceptance in certain areas of applied work, serious questions have been raised regarding the validity, reliability, and sensitivity of visual inspection (Allison & Franklin, 1992; Fisch, 2001; Franklin, Allison, & Gorman, 1996).

Many statistical procedures have been proposed as substitutes for or supplements to inferences based on visual inspection and as means for estimating quantitative effect sizes. Statistical analysis of single-case research is viewed as challenging (or even problematic) due to the need to account for three aspects of single-case data: first, that single-case series often display time trends; second, that repeated measurements of the same case should be treated as serially dependent, rather than independent; and third, that within-case replications (such as in ABAB designs) should be analyzed rather than discarded. Much of the quantitative methodology in single-case research has focused on addressing one or more of these problems, and recent discussions of effect sizes have emphasized the importance of accounting for all three (Horner et al., 2012; Maggin, Swaminathan, et al., 2011; Wolery et al., 2010).

---

applied researchers may consider visual inspection of graphed outcome data to be a sufficient inferential method, particularly because it is rapid and (seemingly) straight-forward.

Some proposals for statistical analysis and meta-analysis of single-case research are premised on the availability of raw data for secondary analysis. Syntheses of between-groups designs must often be based only on published summary statistics or regression results, which may be incomplete or sub-optimal for meta-analytic purposes. In contrast, an advantageous feature of single-case research is that raw data are frequently available. Many journals that publish SCDs require that the data be published in the form of a single-case graphs, and reliable methods exist for extracting data from published graphs (Shadish et al., 2009). Thus, it will often be feasible for a research synthesist to gain access to raw data from a set of studies and to calculate whatever summary statistics are desired, or even to perform a meta-analysis based on individual participant data. This approach has been applied by Van den Noortgate and Onghena (2008), who demonstrated the increased flexibility of modeling raw data rather than just summary statistics. Unlike in meta-analysis of between-groups designs, secondary analysis and meta-analysis of single-case designs could conceivably involve considerably more sophisticated or computationally complex methods than were used in the primary analysis (e.g., visual inspection) of the component studies.

Proposed statistical procedures for effect size estimation fall into three broad categories: non-overlap statistics, parametric regression models of single-cases, and hierarchical models of collections of cases.[9] The remainder of this section briefly describes each of these.

[9]Randomization tests (Edgington & Onghena, 2007; Levin & Wampold, 1999) are another distinct approach to inference from single-case data. Such tests base inference on the distribution induced by an actual or assumed randomization mechanism, and can provide valid inferences without relying on certain of the modeling assumptions that are necessary for inference in parametric regression models. To my knowledge, however, no work to date has demonstrated how such tests could provide effect size metrics that are suitable for meta-analysis. Therefore, I do not consider them further.

1.3.1. Non-overlap statistics

A variety of non-overlap statistics have been proposed as quantitative summaries of effect size for single-case data, with the stated aim of finding statistics that are not sensitive to serial dependence in the outcome measures. In one of the earliest such proposals, Scruggs et al. (1987) suggested using the percentage of non-overlapping data (PND); in a simple AB design, PND is calculated as the percentage of data points in phase B that exceed the highest (or lowest) data point in phase A. Elaborating on PND, Ma (2006) proposed instead calculating the percentage of data points in phase B that exceed the median of points in phase A. Other proposals exist as well, including several motivated by or connected to robust statistics and non-parametric tests (Parker & Vannest, 2009; Parker, Vannest, & Brown, 2009; Parker, Vannest, & Davis, 2011; Parker, Vannest, Davis, & Sauber, 2011). These approaches have come under criticism for several reasons, including for their lack of correct sampling distributions, for not being sensible metrics for measuring effect size magnitude, for their lack of design-comparability, and for not being sensitive to trends or other features of the data (Beretvas & Chung, 2008b; Shadish & Rindskopf, 2007; Shadish et al., 2008; Wolery et al., 2010). Despite such criticisms, non-overlaps statistics (particularly PND) remain the most commonly used effect size for meta-analysis of single-case designs (Beretvas & Chung, 2008b; Maggin, O'Keeffe, & Johnson, 2011).

1.3.2. Parametric models for individual cases

Parametric approaches to meta-analysis of single-case data can be classified by whether each case is analyzed separately or whether a collection of cases is modeled jointly. Most

proposals falling into the former category are based on piece-wise linear regression models (for instance Center et al., 1985; Huitema & McKean, 2000). Some proposals also include simple models for serial dependence within cases, such as lag-1 auto-correlation (Crosbie, 1993; Maggin, Swaminathan, et al., 2011). Under a given parametric model, different authors propose different methods of forming effect size estimates. Here I highlight briefly some of the more prominent approaches; more detailed reviews can be found in Gorman and Allison (1996), Jenson, Clark, Kircher, and Kristjansson (2007), Shadish and Rindskopf (2007), and Beretvas and Chung (2008b).

Busk and Serlin (1992) proposed an effect size estimator for data from a single case that resembles algebraicly the formula for Cohen's d effect size: the mean difference across phases, standardized by the pooled, within-phase sample variance. It has been noted that this effect size is appropriate only if phases do not contain trends and if observations within phases are independent of one another. Hershberger, Wallace, Green, and Marquis (1999) proposed a similar statistic, but based on only the last three data points in each phase.

Center et al. (1985) proposed to use a piece-wise linear regression to account for within-phase trends and to calculate an effect size based on the F-statistic (or equivalently, the change in R^2) corresponding to a null hypothesis of no change in level or slope due to treatment. They and others have argued that the F-statistic can then be converted into a d-type statistic, using the algebraic relationship $d = 2\sqrt{F/(n-2)}$, as given in Rosenthal (1994). Critiques of and elaborations upon this approach were developed by Allison and Gorman (1993); Faith, Franklin, Allison, and Gorman (1996); and Beretvas

and Chung (2008a). Beretvas and Chung (2008a) note that the transformed F-statistic is not necessarily comparable with d-type effect sizes from a between-subjects design.

More recently, Van den Noortgate and Onghena (2003b) and Van den Noortgate and Onghena (2008) demonstrated the use of hierarchical linear models for meta-analysis of SCDs. Though they used a hiearchical model for statistical analysis, they nonetheless formulated the effect sizes based on the parameters of the individual-case model, taking an ad-hoc approach of standardizing the raw data from each case by its pooled within-phase standard deviation. This approach is nearly equivalent to the standardized mean difference proposed by Busk and Serlin (1992).

Maggin, Swaminathan, et al. (2011) described a novel set of effect size measures and estimation procedures that can account for differential time trends and auto-correlation among the observations for an individual case. They proposed to measure the magnitude of treatment effects by comparing predicted values of the outcome in the presence and absence of treatment at a fixed point in time, mid-way through the treatment phase, and to standardize this treatment effect by the within-phase standard deviation of the errors. The authors claimed that the resulting effect size is "consistent with conventional group design effect size measures," (Swaminathan et al., 2010, p. 2), but added the caveat that this is true in the limited circumstance that all subjects have identical means in baseline.

Finally, there have been several proposals to measure effect sizes using proportionate changes. Hershberger et al. (1999) defined several different effect size metrics, all involving predicted outcomes at the end of a treatment phase, where predictions are made based on linear extrapolation from the baseline phase data or by linear regression of the date in the treatment phase. One effect size was a proportionate change index, defined as the

mean difference between predicted outcomes, divided by the predicted outcome based on extrapolation from baseline (Hershberger et al., 1999). Marquis et al. (2000) employed a similar effect size, dubbed the "suppression index," along with several others in a large meta-analysis of intervention research. Finally, J. M. Campbell (2004) and J. M. Campbell and Herzinger (2010) discussed a simpler, related effect size, termed the mean baseline reduction or mean baseline difference and defined as the difference in mean outcomes between phases, scaled by the mean outcome in the baseline phase. Though all of these authors note the intuitive appeal of proportionate change measures for applied researchers, the approach lacks any statistical development in the context of single-case research.

1.3.3. Hierarchical models

Hierarchical linear modeling has been recognized recently as a promising tool for the analysis of SCDs, particularly for combining results from multiple cases within single studies (Jenson et al., 2007; Shadish et al., 2008; Van den Noortgate & Onghena, 2003a; Zucker et al., 1997). In a medical context, Zucker et al. (1997) used an HLM with normal error distributions and Bayesian estimation methods to analyze a set of individually randomized, multiple treatment reversal designs. Their analysis demonstrated that the combined, hierarchical analysis provides improved estimates of treatment effects for individual patients. Zucker, Ruthazer, and Schmid (2010) extended the earlier analysis to a larger set of trials and emphasized the advantages of analyzing raw data rather than summary statistics from individual cases. Dealing with behavioral data, Van den Noortgate and Onghena (2003a) used Gaussian-error HLMs to combine data from multiple cases within a single study. Kyse (2010) applied a hierarchical generalized linear model to single-case

data, comparing results from a model with binomial errors and overdispersion to results using other meta-analytic approaches. Kyse, Rindskopf, and Shadish (2011) presented a tutorial on the application of HGLMs to data from single-case designs. The authors discussed four examples of increasing complexity, covering one-, two-, and four-phase designs and normal, Poisson, and binomial error distributions; they emphasized the flexibility of multilevel models for testing various hypotheses about the data.

While published work demonstrates the great utility of hierarchical modeling of single-case data, attentions has focused largely on analysis of single studies or collections of studies with identically measured outcome variables. Consequently, hierarchical analyses of single-case designs have not attended adequately to the question of how to construct summary effect sizes, and whether these are design-comparable or measurement-comparable. In fact, Kyse et al. (2011) argued that a drawback of HGLMs is that they lack effect sizes that are equivalent to those from between-case designs.

One of the primary strengths of the hierarchical approach is that it allows explicit modeling of variation both within and across cases. Hedges, Pustejovsky, and Shadish (2012a, 2012b) relied on this aspect in order to define a d-type effect size and explicitly demonstrate its design-equivalence. The central idea is that design-comparable effect sizes can be constructed only in a model that is general enough to encompass both the single-case design and a between-subjects randomized experiment. Based on one such model, Hedges and colleagues provided methods for estimating the effect size from data collected in treatment reversal designs (Hedges et al., 2012b) and multiple baseline designs (Hedges et al., 2012a). However, the model under which the proposed effect size is defined relies on strong assumptions, including that (a) the outcome measures are continuous, (b)

baselines are stable, lacking any trends, (c) the treatment effect can be modeled by a shift in the mean level of the outcome, and (d) the treatment effect is homogeneous across cases. In practice, multiple baseline studies frequently exhibit trends in both baseline and treatment phases. Also, the assumption that treatment effects are constant across cases may be undesirable, given that one of the expressed goals of many multiple baselines studies is to study variation in treatment effectiveness (Hersen, 1990). Though use of their specific estimation procedures may be limited due to reliance on these strong assumptions, the general approach of Hedges et al. (2012a, 2012b) has much broader application. In Chapters 2 and 3, I elaborate on their approach, proposing a set of general modeling criteria for defining design-comparable effect sizes and providing extensions for handling a variety of specific models, more complicated than those studied in previous work.

1.4. Overview of this thesis

This thesis proposes approaches to defining and estimating operationally comparable effect sizes for single-case research. I first address design-comparability, extending the approach proposed by Hedges et al. (2012a, 2012b). In Chapter 2, I propose a general framework for defining design-comparable effect sizes through the use of a statistical model that is sufficiently broad so as to encompass both the single-case design at hand and a hypothetical between-groups design. In Chapter 3, I apply the general framework to models for continuous, interval-scale outcomes and d-type effect sizes, discuss a variety of model specifications for multiple baseline and treatment reversal designs, and detail an estimation strategy. In Chapter 4, I demonstrate my proposed models and estimation methods in several applications. I turn to issues of measurement-comparability of effect

sizes in Chapter 5, which focuses on measurement models for free-operant behavioral outcomes while making simple assumptions about the stability of behavior over time. In Chapter 6, I extend the proposed effect size models to incorporate more complex features such as time trends and serial dependence. In the final chapter, I discuss some extensions and areas for further research, then offer some concluding thoughts.

CHAPTER 2

A general framework for design-comparability models

Hedges et al. (2012a, 2012b) demonstrated that a design-comparable standardized mean difference effect size can be obtained from multiple baseline or treatment reversal designs by specifying a model that is "broad enough to encompass both a between-subjects experiment and a single-case design with replications across individuals" (Hedges et al., 2012b, p. 225). This chapter explicates the general logic behind this approach, in which a design-comparable effect size is understood as one that is directly comparable to an effect size from a cross-sectional randomized experiment. Abstracting from specific distributional models, I outline a set of criteria for judging whether a model is "broad enough" and describe how a sufficiently broad model can be used to construct a design-comparable effect size.

The difficulty in finding a sufficiently general model–one that can encompass both a cross-sectional experiment and an SCD–arises from the possibility that treatment effects can vary over the operational dimensions of a study that relate to time. Many aspects of a study's operations include some temporal dimension: a study is conducted over a specific period of calendar time; study participants might be restricted to those at a certain time in their life-course; treatment procedures are scheduled and executed over time; and outcome measurement occurs at specific points in time (c.f. Cronbach, 1982). One might theorize that the effect of a given treatment varies with any of these dimensions; for instance, a treatment could have a gradual or abrupt effect after it is first introduced, the effect could

be temporary or permanent, and the effect of sustained intervention could differ from that of one-time or intermittent intervention (Shadish et al., 2002, Chp. 6).

The treatment effects identified in cross-sectional randomized experiments have a very simple form: they are contrasts (such as differences or ratios) between averaged values of an outcome variable when the treatment is present versus when it is absent, where the average is taken over the units in the experiment. This simplicity stems from temporal specificity: in a typical cross-sectional experiment, treatment procedures are assumed to begin at the time of random assignment and outcome measurements are made at the same time on all units.[1] Thus, there is neither the need nor the possibility to model temporal variation in the treatment effect.

In contrast, single-case designs involve measuring change over time in individual cases. As a consequence of repeated measurement, it becomes possible to observe and model how treatment effects evolve over time. Moreover, the timing of treatment administration is an important aspect in all of the major types of SCDs. For example, multiple baseline designs introduce a treatment at different points in time across units in order to weaken history and maturation threats (Kazdin, 2011; Shadish et al., 2002), and treatment reversal designs emphasize patterns of treatment that are not homogeneous in time. The presence of temporal variation means that the analyst will have to be specific about operational details in order to describe a design-comparable treatment effect size. At the same time, describing design-comparable effect sizes will require models that invoke stronger assumptions than necessary just to describe the observed data from individual cases.

[1]Or if not at the same time, at least according to a schedule that was balanced across treatment groups. Without such balance, the internal validity of the experiment is open to challenge by threats of history and maturation (D. T. Campbell, 1957).

A model needs to meet three criteria in order to be sufficiently general that it can describe both an SCD and a cross-sectional randomized experiment. First, the model must adequately describe the observed data from the SCD under analysis, including capturing the functional form of the outcome process and allowing for possible serial dependence among repeated observations on the same unit. Second, the model must describe a population broad enough that one could conceivably perform an experiment on it, and must capture variation between the units of treatment assignment. Third, the model must be causally interpretable at the level of the unit of treatment assignment. These criteria are quite broad, in that they can be satisfied by a wide variety of analytic models. Likewise, the development in this chapter proceeds in broad terms, describing the nature of each of the assumptions without limiting to any particular model; subsequent chapters will consider specific models and provide detailed applications. After discussing each of the criteria, I explain how a model meeting them can be used to construct design-comparable treatment effects by specifying certain details related to a cross-sectional experiment.

2.1. Adequately describe the observed data

One of the chief advantages of randomized experiments is that–at least in the theoretical ideal–they yield treatment effect estimates and inferences that are internally valid and *model-free*, meaning not contingent on any assumptions regarding functional forms, parametric distributions, or independence of observations (Rosenbaum, 2002; Rubin, 1974, 1978). Single-case designs do not share the same advantage; instead, the internal validity of treatment effect estimates are contingent on having a good model for the process that

generated the data.[2] The first modeling criterion is therefore that *the model must adequately describe the observed data, including capturing the functional form of time trends and allowing for possible serial dependence among repeated observations on the same unit.*

As discussed in Section 1.3, models for single-case data often posit that the outcome process involves deterministic time trends. It is crucial that the functional form of such trends be correctly specified because, in most single-case designs, treatment effects are identified by extrapolating baseline trends forward in time (Horner et al., 2012). As a minimal standard, a model should provide a reasonable fit to the observed data (within phases) if it is to be the basis for extrapolation. Furthermore, the model should capture essential features of the measurements, such as range limitations. For example, many single-case studies use outcomes measured as proportions, which can only range from zero to one; linear trends in the mean of a process measured by proportions will often be implausible, particularly when proportions are near the scale extremes.

In addition to adequately describing the deterministic component of the outcome process, a model must also make assumptions about the stochastic component.[3] The assumption that repeated measurements of an individual may be dependent is one of the basic tenets of longitudinal data analysis, and an important aspect of what distinguishes the field of from other areas of statistics (Fitzmaurice, Laird, & Ware, 2011). Particularly when repeated measurements are closely spaced in time, as in most single case studies, serial dependence should be considered rather than ruled out *a priori*. Further, serial

[2]In this respect, SCDs are like all interrupted time series designs (see Shadish et al., 2002, Chp. 6).

[3]The classical decomposition of time series data involves not only deterministic trends and stochastic errors, but also periodic patterns (such as seasonal trends). This final component has seldom been scrutinized in single-case settings (Beasley, Allison, & Gorman, 1996).

dependence models that include serial independence as a special case can be used to test one's assumptions against data.

In applications to single-case studies, models for serial dependence have been focused largely on lag-one auto-regressive processes with Gaussian errors or other simple, highly structured processes (e.g. Crosbie, 1993; J. W. Harrop & Velicer, 1985; Velicer & McDonald, 1984). Some authors have questioned the existence of serial dependence in behavioral time series data (Huitema, 1985; Huitema & McKean, 1998), while others argue that the existence of lag-one autocorrelation only scratches the surface of larger questions about serial dependence in behavioral time series (Matyas & Greenwood, 1996; Velicer, 1994). In recent discussions of statistical modeling procedures for single-case studies, there appears to be a growing concensus that serial dependence must be considered (Horner et al., 2012; Wolery et al., 2010).

2.2. A population on which one could experiment

Much of education research is concerned with populations that are characterized by hierarchical structure, such as students within classrooms within schools, and interventions on such populations may be implemented at different levels of the hierarchy (Bryk & Raudenbush, 1988).[4] In some instances, interventions can be administered directly to individuals (i.e., the lowest level units), but this is not always the case. Many experiments in education use cluster-randomization, in which the experimenter assigns higher-level units (often called clusters) to treatment or control conditions but measures outcomes on lower-level units (Bloom, 2005; Bloom, Bos, & Lee, 1999; Mosteller & Boruch, 2002). A

[4]In what follows, I refer to the levels of the hierarchy as ordered from lowest to highest, with lower-level units nested within higher-level units.

cluster-randomized design may be preferred over individual-level randomization due to logistical constraints on individual assignment, issues of political feasibility, or features of the intervention that make it impossible to do anything else (Orr, 1999). For instance, if it is to provide a true test of the program, a novel school-wide reform can only be implemented by an entire school; it is implausible that some staff members in a school could faithfully implement such a program but leave their colleagues unaffected.

In order to construct a design-comparable effect size, the analytic model must describe a population in which a cross-sectional experiment could be performed. Thus, if the nature of the intervention would necessitate cluster-randomization, the analytic model must describe multiple clusters. For a given intervention, I posit a *minimum natural level of assignment*, by which I mean the lowest level of a hierarchically structured population to which an intervention could feasibly be assigned (cf. Cronbach, 1982, pp. 92-93). The second modeling criterion is therefore that *the analytic model must describe multiple units at the minimum natural level of assignment, including both within- and between-unit variation.* This criterion ensures that it is possible to use the model to consider a hypothetical randomized experiment. If the model describes only a population of lower-level units within one unit at the minimum natural level of assignment, it will be impossible to define a design-comparable effect size. In many single-case studies, the minimum natural level of assignment will be the individual participant (i.e., the single case); in such studies, the criterion requires a model for multiple participants.

For an example where the minumum level is not the individual, consider a study conducted by Ross and Horner (2009) to evaluate the effects of a school-wide bullying

prevention program. The study used a multiple baseline design across six students, including two students from each of three separate schools. In order to provide a valid test of the intervention's effects, the program was implemented with all staff and students in each school; without a school-wide implementation, one could not claim to have evaluated the program as designed. Thus, the minimum natural level of assignment is the school, and an analytic model for these data must describe variation in a population of multiple schools. Further, a model describing only the students within these three schools, without considering variation across schools, cannot produce design-comparable effect sizes.

2.3. Causal interpretability

In order to construct a design-comparable effect size, the analytic model must clearly describe not just the observed outcome data from the longitudinal study, but also the outcomes that would be observed under variations in how treatment is assigned. These are sometimes called potential outcomes (Holland, 1986; Rubin, 2005). Based on the second criterion, treatment assignment is assumed to take place at the minimum natural level of assignment, leading to the third and final criterion: *the analytic model must be causally interpretable for units at the minimum natural level of assignment.* Different forms of single-case designs warrant separate consideration here, because different patterns of treatment assignment are possible under each of them. I discuss the multiple baseline, treatment reversal, and alternating treatment designs in turn, assuming that the case is the minimum natural level of assignment. Subsequently, I discuss causally interpretable models at higher levels of assignment.

2.3.1. Multiple baseline designs

In a multiple baseline design, each unit receives treatment at a given point in time and continues in the treatment phase thereafter. It may be that the treatment procedures take place instantaneously or take time to administer, so long as they unfold according to a fixed plan. The pattern of treatment assignment for a given unit is specified by the time at which treatment begins. To fix ideas, suppose that case i (a unit at the minimum natural level of assignment) is to be measured at times $j = 1, ..., n$. T_i measurements are made during a baseline phase (in the absence of treatment), where $0 \leq T_i \leq n$; the case then receives treatment, and $n - T_i$ further outcome measurements are made during the treatment phase. A causally interpretable model entails writing the outcome for case i at time j as a function of the treatment assignment time for that case, that is, as $Y_{ij}(T_i)$. The treatment phase for case i might begin after any given measurement occasion, so that T_i takes values from $0, ..., n$, where $T_i = 0$ if the case is assigned to treatment just prior to the first outcome measurement and $T_i = n$ if the case is not assigned to treatment during the study period.

Clearly, a causally interpretable model describes many potential outcome values, only a few of which are observed in a given design. It will often be reasonable to assume that outcomes do not depend on future treatment assignment times (that is, the outcome on Sunday should not depend on whether a case will start treatment on the following Thursday versus on the following Friday). Even with this restriction, a causally interpretable model for case i would describe $Y_{ij}(T_i)$ for $j = 1, ..., n$ and $T_i = 0, ..., j$, a total of $n(n+1)/2$ unique values. Further assumptions could be made as well. For instance, Holland (1986) describes an assumption of temporal stability, meaning that treatment effects

depend only on the length of time since the introduction of treatment, rather than the point at which treatment is introduced. This assumption further restricts the potential outcomes function so that, for instance,

$$Y_{ij}(T_i) - Y_{ij}(n) = \delta_{i(j-T_i)} \tag{2.1}$$

for $j = 1, ..., n$ and $T_i = 0, ..., j$. Assuming that the measurement times are equally spaced, the temporal stability assumption constrains the potential outcome function to at most $2n$ unique parameters. Yet more restrictive, the potential outcomes function could be modeled using linear functions of time, restricting (2.1) so that $\delta_{i(j-T_i)} = \delta \times (j - T_i)$. I consider this and several other simple, causally interpretable case-level models for multiple baseline data in later chapters.

Using potential outcomes to express causally interpretable models is a novel approach for analysis of single-case studies, though the approach is increasingly common in a variety of social science disciplines (S. L. Morgan & Winship, 2007, p. 4). However, there is a close relationship between potential outcomes models and certain analytic methods that are commonly used in single-case research, including both visual and statistical methods. Analysis of single-case designs is concerned with whether a "functional relationship" exists between the treatment and the outcome variable (Horner et al., 2005). Linear extrapolation from the baseline phase is sometimes employed to facilitate comparison with treatment phase outcomes (Allison & Gorman, 1993; Kyse, 2010; Shadish et al., 2008; D. M. White, Rusch, Kazdin, & Hartmann, 1989). Consider a linear model the observed

data from the baseline phase for case i, as in

$$Y_{ij} = \alpha_i + \beta_i \times j + \epsilon_{ij}, \qquad j = 1, ..., T_i. \tag{2.2}$$

The potential outcomes model posits that the trend holds over not only the observed baseline, but that it would also hold over the entire study, in the absence of treatment:

$$Y_{ij}(n) = \alpha_i + \beta_i \times j + \epsilon_{ij}, \qquad j = 1, ..., n. \tag{2.3}$$

The potential outcomes model is simply a method of expressing the implied assumption behind an extrapolation.

2.3.2. Treatment reversal designs

Whereas multiple baseline designs can be used to study interventions that cannot be undone once introduced (such as those that could have permanent effects), treatment reversal designs are appropriate only for studying interventions that can be removed and re-introduced. In such designs, it will sometimes be necessary (and perhaps also reasonable) to assume that treatments have only transient effects. Suppose that a treatment reversal design on a single case i measures an outcome at n equally-spaced times. Per the experimental design, the treatment is present or absent during each outcome measurement. Let T_{ij} indicate whether treatment is present or absent for case i just prior to time j, and let $\tilde{T_{ij}} = (T_{i1}T_{i2}\cdots T_{ij})$ denote the history of treatment up to time j.[5] Again assuming that outcomes cannot be affected by future treatment decisions, a completely general potential outcomes model is described by the function $Y_{ij}(\tilde{T_{ij}})$ for $\tilde{T_{ij}} \in \{0,1\}^j$

[5]This notation is similar to that used by Robins and Hernán (2009).

and $j = 1, ..., n$. The outcome at time j may depend on the presence or absence of the treatment at any point in time up to j. In this most general formulation, there are 2^j potential outcomes at time j, for a total of $2^{n+1} - 2$ potential outcomes over the full set of measurement times.

For the model to be tractable, the very large number of potential outcomes will need to be constrained by further assumptions. If the treatment can be considered homogeneous, so that treatment at time 1 is in some sense the same as treatment at time j, the temporal stability assumption may be appropriate. Letting 0^j denote the sequence $\tilde{T}_{ij} = (0 \cdots 0)$, one version of temporal stability is expressed by the restriction

$$Y_{ij}\left(t_1 \cdots t_j\right) - Y_{ij}\left(0^j\right) = Y_{i,j+k}\left(0^k t_1 \cdots t_j\right) - Y_{i,j+k}\left(0^{j+k}\right) \tag{2.4}$$

for $j = 1, ..., n$, $k = 1, ..., n - j$, and $t_1, ..., t_j \in \{0, 1\}$.

Depending on the treatment, other assumptions might be entertained instead of or in addition to temporal stability. One might assume that the treatment effect is a function of cumulative exposure, so that

$$Y_{ij}(t_1 \cdots t_j) - Y_{ij}(u_1 \cdots u_j) = \beta \sum_{k=1}^{j} (t_k - u_k). \tag{2.5}$$

Alternately, one might maintain an assumption of causal transience (Holland, 1986), meaning that the treatment effect "wears off" some time after its removal. If treatment completely wears off after $k \in \{0, ..., n-1\}$ periods, this assumption might be expressed as

$$Y_{ij}\left(t_1 \cdots t_j, 0^k\right) = Y_{ij}\left(0^{j+k}\right) \tag{2.6}$$

for all $j = 1, ..., n-k$ and $t_1, ..., t_j \in \{0, 1\}$. In the most extreme form of causal transience, potential outcomes at time j depend only on T_{ij}:

$$Y_{ij}(\tilde{T}_{ij}) = Y_{ij}(T_{ij}). \tag{2.7}$$

Combining this assumption with temporal stability implies that there is a single, uniform treatment effect δ_i and that $Y_{ij}(\tilde{T}_{ij}) = Y_{ij}(0^n) + \delta_i T_{ij}$.

In many treatment reversal designs, rapid introduction and removal of the treatment is not considered. Thus, an alternative approach to making tractable the most general form of the potential outcomes model is to restrict its domain, asserting that not all possible patterns of treatment assignment are of interest. Rather than allowing the domain of $Y_{ij}(\tilde{T}_{ij})$ to be $\tilde{T}_{ij} \in \{0, 1\}^j$, one could restrict the domain to include only those patterns where treatment is in place or absent for a minimum number of periods k, so that the maximum number of treatment reversals through time j is $1 + \lfloor j/k \rfloor$. For a given number of treatment reversals r, let

$$\tilde{T}^r_{ij}(t_0, ..., t_r) = \begin{cases} (0^{t_0} 1^{t_1} 0^{t_2} \cdots 0^{t_{r-1}} 1^{t_r}) & r \text{ odd} \\ (0^{t_0} 1^{t_1} 0^{t_2} \cdots 1^{t_{r-1}} 0^{t_r}) & r \text{ even} \end{cases}$$

for $t_0 \geq 0$, $t_1, ..., t_r > k$, and $\sum_{p=0}^{r} t_p = n$. The potential outcomes function can then be restricted to $Y_{ij}(\tilde{T}^r_{ij})$ for $\tilde{T}^r_{ij}$, $r \in \{0, ..., (1 + \lfloor j/k \rfloor)\}$.

2.3.3. Alternating treatment designs

Alternating treatment designs are very similar to treatment reversal designs in terms of the general structure of their potential outcomes functions; the distinction is a matter of

degree. Treatment reversal designs are often limited to a small number of reversals, with each phase containing multiple data points in which the same treatment is in place. In comparison, alternating treatment designs use more rapid introduction and removal of a treatment. It is also somewhat more common for alternating treatment designs to consider more than two treatment conditions, and for the design to use random assignment of treatment times. For k distinct treatment conditions, the potential outcomes function can be written in most general terms as $Y_{ij}(\tilde{T}_{ij})$ for $\tilde{T}_{ij} \in \{0, ..., k-1\}^j$. Random assignment schemes may place restrictions on which treatment sequences are possible, for instance by ensuring that all k treatment conditions are used an equal number of times or that each treatment is used once per day (i.e., blocking on day). For example, Himle, Woods, and Bunaciu (2008) used a randomized alternating treatment design to evaluate the effects of two different reinforcement schedules relative to a no-reinforcement condition on tic suppression in four children with Tourettes syndrome. Across twelve successive five-minute sessions, the three conditions (no reinforcement, differential reinforcement, and non-contingent reinforcement) were randomized such that each condition occurred four times per child.

Another occasional feature of alternating treatment designs is that outcome measurements may be clustered together in time. For instance, an experimenter might conduct s sessions of treatment and outcome measurements within two hours and repeat this process daily for n days. Such a structure can be described heirarchically, with session time jk denoting the k^{th} session on the j^{th} day. For each session, treatment T_{ijk} is administered and outcome measurement Y_{ijk} is measured. Let $\tilde{T}_{ij}(k) = (T_{ij1} \cdots T_{ijk})$ denote the history of treatment for case i on day j, up through session $k = 1, ..., s$, and let

$\tilde{T}_{ijk} = (\tilde{T}_{i1}(s)\tilde{T}_{i2}(s)\cdots\tilde{T}_{i(j-1)}(s)\tilde{T}_{ij}(k))$. With such a schedule, and with certain treatments, it might be reasonable to assume that causal transience applies across days but not across sessions within days, implying that $Y_{ijk}(\tilde{T}_{ijk}) = Y_{ijk}(\tilde{T}_{ij}(k))$. Combining this assumption with temporal stability implies that

$$Y_{ijk}(\tilde{T}_{ij}(k)) - Y_{ijk}(0^k) = Y_{i1k}(\tilde{T}_{i1}(k)) - Y_{i1k}(0^k) \tag{2.8}$$

for $j = 1, ..., n$ and $k = 1, ..., s$. The full set of potential outcomes can then be described by the sn outcomes in the absence of any treatment, $Y_{i11}(0), ..., Y_{isn}(0^{sn})$ plus $2^{s+1} - s - 2$ treatment effect parameters.

For an example of a hierarchically structured alternating treatment design, consider a study by Horrocks and Higbee (2008) that evaluated the effect of preferred versus non-preferred auditory reinforcement stimuli on the performance of free-operant tasks by six adolescents with developmental disabilities. Three five-minute sessions were conducted in succession each day for between six and eight days; each of three reinforcement stimulus conditions (high-preference, low-preference, and no-reinforcement) was used in one session per day, in a "semi-random" order. To describe this design in terms of potential outcomes, let n_i be the number of days on which participant i was studied, and let $T_{ijk} \in \{0, 1, 2\}$ indicate the stimulus condition used in session $k = 1, ..., 3$ of day $j = 1, ..., n_i$. Due to the restriction that each condition occur once per day, the treatment history of case i on day j is restricted to $\tilde{T}_{ij}(3) \in \{(012), (021), (102), (120), (201), (210)\}$, rather than the more general $T_{ij}(3) \in \{0, 1, 2\}^3$.

The three designs discussed thus far are the most commonly used in single-case research, but my coverage has been far from exhausitive. Single case researchers create

novel designs by combining elements from multiple baselines, treatment reversals, and alternating treatments; often this is done in order to study moderating effects or make comparisons among more than two treatments. For instance, the study by Horrocks and Higbee (2008) described above actually combined the alternating treatment design with a multiple baseline design, in which the reinforcement ratio was varied across days for some participants. These more complex designs will likely require modeling on a case-by-case basis; for present purposes, I do not consider them further.

2.3.4. Causally interpretable models at higher levels of assignment

Thus far my discussion of causally interpretable models has focused on those for interventions where the minimum natural level of assignment is the case, that is, where the level of assignment is the same as the level on which outcomes are (repeatedly) measured. For other interventions, where the intervention is assigned at the level of a cluster of cases, the above models can be extended under the assumption that the clusters are intact. By intact clusters, I mean that the cluster to which each individual case belongs is not affected by treatment assignment. Under this assumption, the potential outcomes for each individual case become a function of a cluster-level assignment indicator. I illustrate this below for multiple baseline and treatment reversal designs. For both designs, I use the following notation: let cluster h contain m_h cases; for case i in cluster h, let Y_{hij} denote the outcome measurement on occasion $j = 1, ..., n$.

In a multiple baseline design, the treatment assignment pattern in cluster h can be described by the occasion of the last baseline measurement for cases in the cluster, $T_h \in \{0, ..., n\}$, where $T_h = 0$ means that the cluster begins treatment prior to the first

measurement, and $T_h = n$ means that the cluster does not receive treatment during the study. The potential outcomes at time j for case i in cluster h are then described by $Y_{hij}(T_h)$. Restrictions on the potential outcomes model follow just as in the model with case-level assignment. To see the implications of the cluster-level model, consider again Ross and Horner (2009), which used a multiple baseline design across six students at three separate schools, measuring outcomes on each of 60 days. For each school, the treatment was implemented simultaneously for all students and faculty (including for many students not actually measured). Thus, the model would describe $Y_{hij}(T_h)$ for $h = 1, 2, 3$, $i = 1, 2$, and $j = 1, ..., 60$, from which it is apparent that the treatment assignment times for the first two students are the same due to necessity rather than coincidence. The cluster-level model illustrates that the set of potential assigment patterns is considerably more limited than a multiple baseline design in which individuals are separately assigned to treatment.

In a treatment reversal design, a model for a cluster-level treatment assignment pattern has to allow that the treatment may be in effect or not, depending on the measurement occasion. Let T_{hj} be an indicator variable for whether or not the treatment is in effect in cluster h and time j; following the case-level model, let $\tilde{T}_{hj} = (T_{h1}T_{h2} \cdots T_{hj})$ indicate the history of treatment in cluster h up to time j. The potential outcomes at time j for case i in cluster h are then described by $Y_{hij}(\tilde{T}_{hj})$, for $j = 1, ..., n$ and $i = 1, ..., m_h$. For example, consider a study evaluating the effect of using response cards during question-and-answer sessions on the disruptive behavior of fourth grade students (Lambert, Cartledge, Heward, & Lo, 2006). The study reported repeated measurements of the level of disruptive behavior among nine target students in two classrooms. Since the intervention in this case is a teaching technique, it must be present or absent for all students in a classroom at

once. Thus, the model in this case would describe $Y_{hij}(\tilde{T}_{hj})$ for $\tilde{T}_{hj} \in \{0,1\}^j$, $h = 1, 2$, $i = 1, ..., m_h$, and $j = 1, ..., 34$.

2.4. Target effect size parameters

A model meeting the three criteria described in this chapter allows one to specify precisely a design-comparable effect size. To describe this effect size, it is helpful to consider a hypothetical, cross-sectional experiment, in which treatment assignment begins at a fixed point in time, a fixed schedule of treatment follows, and outcomes are measured at a fixed, later point in time. Though this is clearly a stylization of how any actual experiment occurs, it is nonetheless useful in that it makes apparent how a design-comparable effect size can depend on any or all of these time-related operations. Before I define effect sizes for multiple baseline, treatment reversal, and alternating treatment designs, I first review how effect sizes are constructed in cross-sectional experiments.

In a cross-sectional experiment comparing treatment versus control conditions, and assuming unit-level treatment assignment, a sufficient potential outcomes model consists of just two quantities: the outcome under control $Y_i(0)$ and the outcome under treatment $Y_i(1)$. An individual treatment effect is a comparison between these two potential outcomes, such as their difference or ratio. However, only one of the potential outcomes can be observed on an individual unit at a given point in time; this is what Holland (1986) terms the fundamental problem of causal inference. Consequently, effect sizes for cross-sectional experiments are limited to parameters of the marginal distributions of the potential outcomes (rather than parameters of the joint distribution) because only the former are fully

identified. For instance, the d-type effect size $[\mathrm{E}(Y_i(1)) - \mathrm{E}(Y_i(0))] / \sqrt{\mathrm{Var}(Y_i(0))}$ is identified by a cross-sectional experiment, but $\mathrm{E}[Y_i(1) - Y_i(0)] / \sqrt{\mathrm{Var}[Y_i(1) - Y_i(0)]}$ is not; $\Pr(Y_i(1) < c) - \Pr(Y_i(0) < c)$ is identified, but $\Pr(Y_i(1) - Y_i(0) < c)$ is not. Design-comparable effect sizes are therefore quite limited relative to the set of treatment effect parameters that may be of theoretical interest; these limitations are all the more relevant when individual treatment effects are heterogeneous.

I now turn to the construction of design-comparable effect sizes. Imagine a cross-sectional experiment in which treatment begins immediately following a certain measurement occasion A (the *implementation time*) and in which outcomes are measured on all units at a later time B (the *follow-up time* or *end-point*). For multiple baseline designs, these design parameters are all that is needed to describe a design-comparable effect size. Given times A and B, the causally interpretable model for the multiple baseline design can be used to describe the treatment effect that would be observed in the hypothetical experiment: the effect at time B of introducing treatment at time A. In terms of potential outcomes, the effect size is a contrast between the distribution across cases of $Y_{iB}(T_i = A)$ and that of $Y_{iB}(T_i = n)$. For interventions at higher levels of assignment, the effect size is a contrast between the distribution of $Y_{hiB}(T_h = A)$ and that of $Y_{hib}(T_h = n)$. In designs other than the multiple baseline, some further operational details must be specified. For treatment reversal and alternating treatment designs, I assume for purposes of exposition that the effect of interest is based on a schedule of continuous treatment between times A and B, meaning that the treatment is not removed once it is introduced. In terms of potential outcomes, the effect size is a contrast between the distributions of

$Y_{iB}\left(\tilde{T}_{iB} = 0^A 1^{B-A}\right)$ and of $Y_{iB}\left(\tilde{T}_{iB} = 0^B\right)$; for interventions at higher levels of assignment, the effect size compares the distribution of $Y_{hiB}\left(\tilde{T}_{hB} = 0^A 1^{B-A}\right)$ versus that of $Y_{hiB}\left(\tilde{T}_{hB} = 0^B\right)$. Later chapters deal in detail with specific effect sizes, all constructed from parameters of the marginal potential outcomes distributions.

2.4.1. Choosing A and B

The design-comparability model analyzed by Hedges et al. (2012a, 2012b) assumed that the effect size was stable across measurement occasions, and therefore functionally independent of the design parameters A and B. As shall be seen in later chapters, certain more general models do not share this property; instead, chosen values of A and B will influence the magnitude of the design-comparable effect size to a greater or lesser extent, depending on the model specification used. This sensitivity naturally raises the question of how one should choose these design parameters. I offer several general comments, while deferring any definitive answer until later, in the context of specific applications.

First, I would argue that the choice of which times A and B to use should be informed by scientific concerns, rather than solely by methodological ones. An investigator planning an actual cross-sectional randomized experiment would face essentially the same design choices, and would need to choose implementation and follow-up times that are both feasible and relevant to extant scientific theory regarding the intervention. The choice of end-points might also be guided by the conventions of past research in the same disciplinary context or therapeutic area.

Second, the choice of A and B should be tempered by the extent of extrapolations from the observed data. Increasing the time between implementation and follow-up times

will require longer extrapolations, and will in general decrease the credibility of the effect size estimate. Given that most SCDs consist of relatively short time series, issues of model identification might be inconsequential if short follow-up times are used, but could create great ambiguities when longer follow-ups are specified. In fact, such sensitivity provides one rationale for the study of design-comparability. For a given choice of design parameters, a design-comparable effect size is useful as a focal parameter for studying the sensitivity of estimates and inferences to variations in modeling assumptions. By focusing on the design-comparable effect size, one can limit the scope of causal inference to a specific parameter, and perhaps even weaken modeling assumptions as a consequence.

Finally, in the context of a meta-analysis, it might be desirable to choose the design parameters A and B so as make effect size estimates as operationally comparable as possible. For instance, suppose that a meta-analysis is to include several single-case studies and one randomized trial. One could choose design parameters for the single-case studies that mimic the operations of the randomized trial, thus reducing the heterogeneity of study effects that is due solely to incidental design choices.

2.4.2. Related effect size proposals

The approach described in this chapter bears a certain resemblance to other proposals in the literature on single-case effect sizes. For example, Hershberger et al. (1999) proposed measuring effect sizes by comparing the last three observations in the baseline phase to the last three observations in the treatment phase, using a standardized mean difference. Swaminathan et al. (2010) proposed comparing the observed outcomes at the mid-point of the treatment phase to the projected trend from the baseline phase. While the exact

interpretation of these effect sizes would depend to some extent on the form of the outcome process being modeled, it can be seen that they are not design-comparable. By defining the effect size metric relative to the length of treatment phase, these proposals introduce *design-dependence*, or heterogeneity among effect sizes due to variation in study design. Effect sizes so formulated might vary from study to study due only to differences in the length of the treatment phases employed in each study. In contrast, as I have argued in this chapter, using fixed implementation and follow-up times (as in a hypothetical cross-sectional experiment) to define the effect size of interest controls for irrelevant design-related variation and thus improves the interpretability of averages across and contrasts between study results.

CHAPTER 3

Design-comparable standardized mean differences: Modeling and estimation

The standardized mean difference, often called Cohen's d, is the most well-known and widely used family of effect size measures for summarizing treatment effects. Standardization using some measure of variation is intended to solve a ubiquitous measurement-comparability problem: that different studies use outcome measurements on different scales, for which no other means of equating may be available. Standardized mean differences are therefore distinctly useful for interval-scale, continuous measurements, in which ratio comparisons are not meaningful. In this chapter, I apply the abstract modeling criteria outlined Chapter 2 to the specific case of d-type effect sizes. I survey a set of models for multiple baseline designs and treatment reversal designs under which design-comparable standardized mean differences can be defined, then propose one method of estimating those effect sizes.

In Chapter 2, I described three criteria for modeling of single-case designs that, if met, allow a design-comparable effect size to be defined. For evaluating an intervention that can be assigned to individual cases, the components of such a model are: 1) a causally interpretable case-level model, 2) a description of the variation across cases, and 3) an adequate description of the observed data. Hierarchical models with normal (Gaussian) errors provide a broad and flexible framework for satisfying these criteria, particularly for

studies using continuous, interval-scale outcome measurements. As noted in Section 1.3.3, hierarchical linear models has recently received increased attention as tools for modeling single-case data. Here, I demonstrate their application for purposes of defining design-comparable effect sizes, focusing on a few models that are likely leading candidates for analysis of data from multiple baseline and treatment reversal designs.

As is often the case in statistical matters, there are many possible approaches to estimation of hierarchical models, with different rationales and different properties. For purposes of design-comparable effect size estimation, the foremost concern is finding estimators that are approximately unbiased, because any bias will propogate through later meta-analysis of the estimates. Finding unbiased estimators is particularly challenging when dealing with studies containing few cases and relatively few observations per case, as are typical in single case research. Additionally, one would prefer estimation methods that are accessible, meaning that they can be executed using widely available software, and extensible, meaning that they can be used to estimate a variety of models, tailored to the specifics of a study, rather than only a few pre-specified models. Different approaches to estimation satisfy the criteria of unbiasedness, accessibility, and extensibility to varying degrees.

Hedges, Pustejovsky, and Shadish (hereafter HPS) proposed a specialized approach to effect size estimation, assuming a certain model for the data from treatment reversal designs (Hedges et al., 2012b) or from multiple baseline designs (Hedges et al., 2012a). For both designs, the approach provides close-to-unbiased estimates of effect sizes while being fairly insensitive to the method used to estimate nuisance parameters that are not of direct interest. The estimation methods are designed to be accessible, in that they

consist only of closed-form algebraic formulas. However, the HPS approach relies on restrictive assumptions about the model and data under consideration (Models MB1 and TR1, described further in following sections), making it difficult to generalize for more complex models.

Shadish et al. (2012) outlined a fully Bayesian approach to estimation of a similar model (Model MB2), although their analysis did not allow for serial dependence between outcomes measured on the same case. The authors implemented their method using the WinBUGS software (Lunn, Thomas, Best, & Spiegelhalter, 2000), which uses Markov Chain Monte Carlo (MCMC) computational methods. Though this approach appears to be fully extensible, the frequentist properties of the resulting effect size estimator are not well understood in this setting. Furthermore, the specialized software needed for MCMC computations could limit the accessibility of this approach.

In settings other than single-case research, hierarchical linear models are often estimated using likelihood-based methods, which include the well-known and widely applied methods of full information maximum likelihood (FML) and restricted maximum likelihood (RML) estimation. In FML estimation, parameter estimates are set equal to the parameter values that maximize the likelihood function for a specified model. RML estimation, proposed by Patterson and Thompson (1971), uses a penalized likelihood function that often produces better estimates of variance components from small datasets. Chi and Reinsel (1989), Jennrich and Schluchter (1986), Laird and Ware (1982), and Lindstrom and Bates (1988) have applied RML estimation in the context of models for longitudinal repeated measurements. Singer and Willett (2003, Section 4.3.2) and Kreft and De Leeuw

(1998, Section 5.6) both provide non-technical discussions of RML and related estimation approaches.

FML and RML estimation have several great advantages. First, these approaches can be applied even in complex hierarchical models and in designs with missing data or unevenly spaced measurement occasions. Furthermore, using statistical techniques that are so well-developed gives the analyst access to a broad array of other statistical tools that are useful for model development and interrogation, including likelihood ratio tests and techniques for estimating individual random effects (see for example Pinheiro & Bates, 2000; Raudenbush & Bryk, 2002). Finally, many software packages provide implementations of full and restricted maximum likelihood estimation for hierarchical linear models, including `SPSS` version 11.0 and following, the `nlme` package in `R` (Pinheiro, Bates, DebRoy, & Sarkar, 2012), the `xtmixed` (StataCorp, 2011) and `gllamm` (Rabe-Hesketh, Skrondal, & Pickles, 2004) commands in `Stata`, `PROC MIXED` in `SAS` (SAS Institute Inc., 2008), and the stand-alone programs `HLM` (Raudenbush, Bryk, & Congdon, 2011) and `ASReml` (Gilmour, Gogel, Cullis, & Thompson, 2009). West, Welch, and Galecki (2007) demonstrate the use of many of these programs. Although in the abstract, FML and RML involve more complex computations than the HPS methods, the wide availability of software implementations for the latter makes those differences immaterial in terms of practical accessibility.

Given its advantages, RML estimation would seem a natural starting point for estimation of the hierarchical models on which the design-comparable effect sizes are based. I focus on RML estimation rather than FML on the assumption that the former criteria will lead to less-biased estimates of the variance components that are a main component

of the effect size. More generally, by basing effect size estimation on RML, the method I describe can be viewed as one step in an integrated statistical analysis of the data from a single-case study, rather than as a separate and isolated procedure.

The remainder of this chapter is organized as follows. Section 3.1 describes a set of hierarchical models for multiple baseline designs and examines the form of the design-comparable standardized mean difference parameter under each model. Section 3.2 proceeds along the same lines for treatment reversal designs. Section 3.3 reviews RML estimation and proposes an effect size estimator that can be applied to any of the models just described. Section 3.4 presents several simulation studies that examine the operating characteristics of the proposed effect size estimator for sample sizes typical of single-case designs. Finally, Section 3.5 concludes by discussing open questions and further extensions to the models and estimation methods considered in this chapter.

3.1. Models for multiple baseline designs

Suppose that a multiple baseline design measures outcomes at each of n equally-spaced times, on each of m cases, where the treatment is introduced to case i just after time T_i. As discussed in Section 2.4, a design-comparable effect size for the multiple baseline design can be operationally defined by specifying a time-point for treatment introduction A and a time-point for outcome measurement B. Once these are specified, a design-comparable d-type effect size is given by

$$\delta_{AB} = \frac{\mathrm{E}[Y_{iB}(A)] - \mathrm{E}[Y_{iB}(n)]}{\sqrt{\mathrm{Var}[Y_{iB}(n)]}} \tag{3.1}$$

This parameter represents the difference between the average outcome if treatment is introduced just after time A and the average outcome if treatment is not introduced (i.e., introduced after time n), scaled by the standard deviation of the outcome if treatment is not introduced, where all outcomes are measured at a fixed time B. Note that this definition relies on having a causally interpretable model for the data, in order for the quantities $Y_{iB}(A)$ and $Y_{iB}(n)$ to be defined.

This section presents a catalog of model specifications for multiple baseline designs, along with the corresponding effect size parameters. I write the models using a two-level formulation similar to that used by Singer and Willett (2003), where level one (the case level) describes a regression model for observations $j = 1, ..., n$ on the i^{th} individual and level two (the group level) describes how the case-level regression coefficients vary across cases $i = 1, ..., m$. As will be seen, the specifications under consideration can all be written using the same case-level assumptions; only the group-level assumptions change. For each model, I will re-express the general parameter defined by (3.1) in terms of model components. As will be seen, in some models the effect size may depend only on $B - A$, only on B, or on both A and B.

3.1.1. Case-level assumptions

The third modeling criterion described in Chapter 2 requires a case-level model that is causally interpretable. I therefore specify a structural model where the outcome Y_{ij} is a function of treatment assignment time T_i, as follows:

$$Y_{ij}(T_i) = \beta_{0i} + \beta_{1i}1(j > T_i) + \beta_{2i}(j - C) + \beta_{3i}\left((j - T_i) \times 1(j > T_i)\right) + \epsilon_{ij} \tag{3.2}$$

Here and following, $1(j > T_i)$ is an indicator variable equal to 0 for $1 \leq j \leq T_i$ and to 1 for $T_i < j \leq n$. Equation (3.2) is a piece-wise linear regression model, a very common specification for analysis of multiple baselines (Center et al., 1985; Gottman, 1981; Huitema, 2011; Huitema & McKean, 2000). In this model, time is taken to be equivalent to measurement occasion and is centered at the constant C. In the case level model, the choice of centering time affects only the interpretation of the the intercept, but it will have larger implications in some of the group-level specifications discussed subsequently.

To interpret the coefficients, it is helpful to first consider the form of the regression if case i does not receive treatment, so that $T_i = n$. The model then reduces to $Y_{ij}(n) = \beta_{0i} + \beta_{2i}(j - C) + \epsilon_{ij}$, and it can be seen that β_{0i} represents the average level of the outcome at time $j = C$ in the absence of treatment while β_{2i} represents the linear change in the outcome per measurement occasion, also in the absence of treatment. For arbitrary T_i, β_{1i} and β_{3i} describe the effect of the treatment on case i. Specifically, β_{1i} represents the immediate change in the level of the outcome due to introducing the treatment, while β_{3i} represents additional change in the outcome per measurement occasion that is due to the treatment. If the treatment began at time $A + 1$, then the individual treatment effect for case i at time B would be $\beta_{1i} + \beta_{3i}(B - A)$. Note that if β_{3i} is assumed to be zero (i.e., not included in the regression model), then the individual treatment effect reduces to β_{1i}, a constant effect not depending on the choice of A or B.

It remains to specify assumptions about the error term ϵ_{ij}. Because measurements on each case are taken over time, the assumption that the errors are independent is usually considered implausible. In the literature on statistical analysis of single-case designs, the

most common assumption is that the errors are auto-correlated and follow a stationary, first-order auto-regressive process. I follow this convention as well, by assuming that the errors have expectation zero, variance σ^2, and first-order autocorrelation ϕ. The last assumption implies that $\text{Cov}(\epsilon_{ij}, \epsilon_{ik}) = \phi^{|k-j|}\sigma^2$. Further, all errors are assumed to be independent across cases, so $\text{Cov}(\epsilon_{hj}, \epsilon_{ik}) = 0$ if $h \neq i$. Having described the case-level assumptions, the remainder of this section examines several group-level specifications that make different assumptions about variation across individuals.

3.1.2. Group-level assumptions

Model MB1: Varying intercepts, no trends

Hedges et al. (2012a) considered perhaps the simplest possible model for multiple baseline data, assuming that baseline outcomes are stable (lacking trend) and that the treatment causes a shift in the level of the outcome that is constant across individuals. Using the case-level model from (3.2), their model is equivalent to assuming that

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \quad \beta_{1i} = \gamma_{10}, \quad \beta_{2i} = 0, \quad \beta_{3i} = 0, \tag{3.3}$$

where η_{0i} is normally distributed with mean zero and variance τ_0^2. Here, γ_{00} is the average level of the outcome across individuals in the absence of treatment, $\eta_{0i} = \beta_{0i} - \gamma_{00}$ is the deviation from this average level for case i, and γ_{10} is the treatment effect, assumed to be constant across individuals.[1] The coefficients for the time trends β_{2i} and the time-by-treatment interactions β_{3i} are both assumed to be zero.

[1] In the notation of Hedges et al. (2012a), $\gamma_{00} = \mu^C$, $\gamma_{10} = \mu^T - \mu^C$, $\eta_{0i} = \eta_i$, and $\tau_0^2 = \tau^2$.

To find the effect size parameter for Model MB1, substitute the assumptions of (3.3) into the case-level regression model (3.2) to arrive at the mixed-model specification:

$$Y_{ij}(T_i) = \gamma_{00} + \gamma_{10} 1(j > T_i) + \eta_{0i} + \epsilon_{ij}. \tag{3.4}$$

From the mixed-model specification, it can be seen that $\mathrm{E}(Y_{iB}(N)) = \gamma_{00}$, $\mathrm{E}(Y_{iB}(A)) = \gamma_{00} + \gamma_{10}$, $\mathrm{Var}(Y_{iB}(N)) = \mathrm{Var}(\eta_{0i} + \epsilon_{ij}) = \tau_0^2 + \sigma^2$, and from Expression (3.1),

$$\delta_{AB} = \frac{\gamma_{10}}{\sqrt{\tau_0^2 + \sigma^2}}. \tag{3.5}$$

Because both β_{2i} and β_{3i} are assumed to be zero, the treatment effect does not depend on the choice of A or B. In Section 3.3, I will describe a method for estimating δ_{AB} that differs from the one proposed in Hedges et al. (2012a).

Model MB2: Varying treatment effects

Model MB1 makes the restrictive assumption that the treatment effect β_{1i} is constant across cases. This assumption can be relaxed by allowing the treatment effect to vary across individuals, while retaining the assumptions regarding the stability of baseline and treatment phases. Such a model was studied by Ferron, Bell, Hess, Rendina-Gobioff, and Hibbard (2009). The group-level specification becomes:

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \quad \beta_{1i} = \gamma_{10} + \eta_{1i}, \quad \beta_{2i} = 0, \quad \beta_{3i} = 0, \tag{3.6}$$

where (η_{0i}, η_{1i}) is multi-variate normally distributed, with mean $(0, 0)$ and covariance matrix

$$\mathbf{T} = \begin{bmatrix} \tau_0^2 & \tau_{10} \\ \tau_{10} & \tau_1^2 \end{bmatrix}.$$

Because the effect size is a linear contrast and is scaled by the standard deviation of the outcome in the absence of treatment, allowing β_{1i} to vary randomly does not alter the parameter of interest; instead, its form is the same as the effect size under MB1, as given in (3.5). Though the parameter is identical, the assumption that the treatment effect is not constant across individuals does have implications for how the parameter is estimated.

Model MB3: Varying intercepts, fixed trends

Multiple baseline data often exhibit trends during the baseline phase, treatment phase, or both; Model MB1 might be criticized as overly restrictive for ignoring this possibility. A slightly less restrictive model would allow for trends in both the baseline and treatment phase, but assume that those trends are common across individuals. Along with the assumption that the treatment has a constant effect across cases, the group-level model becomes:

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \quad \beta_{1i} = \gamma_{10}, \quad \beta_{2i} = \gamma_{20}, \quad \beta_{3i} = \gamma_{30}, \tag{3.7}$$

where $\eta_{0i} \sim N(0, \tau_0^2)$, as in Model 1. The parameters γ_{00} and η_{0i} have the same interpretation as in Model MB1, but now γ_{10} represents the immediate change in the outcome after introducing treatment, γ_{20} represents the change in the outcome per measurement

occasion in the absence of treatment, and γ_{30} represents additional change in the outcome per measurement occasion due to introducing treatment. All of γ_{10}, γ_{20}, and γ_{30} are assumed to be constant across individuals.

Following the same procedure as before, it can be seen that $\mathrm{E}(Y_{iB}(N)) = \gamma_{00}+\gamma_{20}(B-C)$, $\mathrm{E}(Y_{iB}(A)) = \gamma_{00}+\gamma_{10}+\gamma_{20}(B-C)+\gamma_{30}(B-A)$, $\mathrm{Var}(Y_{iB}(N)) = \mathrm{Var}(\eta_{0i}+\epsilon_{ij}) = \tau_0^2+\sigma^2$, and

$$\delta_{AB} = \frac{\gamma_{10} + \gamma_{30}(B - A)}{\sqrt{\tau_0^2 + \sigma^2}}. \tag{3.8}$$

Here, the effect size parameter depends on the difference $B-A$, the length of time between treatment introduction and outcome measurement. However, if $B - A$ is held constant, the parameter does not depend on the choice of B alone, because the variance is constant across measurement occasions, regardless of the pattern of treatment assignments.

Model MB4: Varying trends

In some multiple baseline studies, the assumption in Model MB3 that the baseline slopes are constant across cases may itself be overly restrictive. To relax this assumption, let

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \quad \beta_{1i} = \gamma_{10}, \quad \beta_{2i} = \gamma_{20} + \eta_{2i}, \quad \beta_{3i} = \gamma_{30}, \tag{3.9}$$

where (η_{0i}, η_{2i}) is multi-variate normally distributed, with mean $(0, 0)$ and covariance matrix

$$\mathbf{T} = \begin{bmatrix} \tau_0^2 & \tau_{20} \\ \tau_{20} & \tau_2^2 \end{bmatrix}.$$

The mixed specification for Model MB4 is

$$Y_{ij}(T_i) = \gamma_{00} + \gamma_{10}1(j > T_i) + \gamma_{20}(j - C) + \gamma_{30}\left((j - T_i) \times 1(j > T_i)\right) + \eta_{0i} + \eta_{2i}(j - C) + \epsilon_{ij}. \tag{3.10}$$

As in Model MB3, $\mathrm{E}(Y_{iB}(N)) = \gamma_{00} + \gamma_{20}(B - C)$ and $\mathrm{E}(Y_{iB}(A)) = \gamma_{00} + \gamma_{10} + \gamma_{20}(B - C) + \gamma_{30}(B - A)$. However, the variance now changes over time, with

$$\mathrm{Var}(Y_{iB}(N)) = \mathrm{Var}(\eta_{0i} + \eta_{2i}(B - C) + \epsilon_{ij}) = \tau_0^2 + (B - C)^2\tau_2^2 + 2(B - C)\tau_{20} + \sigma^2.$$

It follows that the design-comparable effect size is

$$\delta_{AB} = \frac{\gamma_{10} + \gamma_{30}(B - A)}{\sqrt{\tau_0^2 + (B - C)^2\tau_2^2 + 2(B - C)\tau_{20} + \sigma^2}}. \tag{3.11}$$

This parameter depends on the choice of both A and B, rather than just their difference. A simpler algebraic expression can be obtained by choosing to center at time $C = B$; in this case, $\mathrm{Var}(Y_{iB}(N)) = \tau_0^2 + \sigma^2$ and the effect size parameter reduces to (3.8). However, even with this simplification, the effect size parameter still depends implicitly on B though the choice of centering point.

Model MB5: Varying intercepts, varying trends, varying treatment-by-time interaction

To illustrate how hierarchical models can be tailored to specific contexts, I consider one further model here, to be applied in the third example of Chapter 4. Model MB5 elaborates on MB4 by assuming that the treatment-by-time trend interaction varies across

cases, while holding fixed the immediate treatment effect (the β_{1i}). The case-level assumptions for this model are identical to those in the previous models, as given in 3.2. The group-level assumptions are as follows:

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \quad \beta_{1i} = \gamma_{10}, \quad \beta_{2i} = \gamma_{20} + \eta_{2i}, \quad \beta_{3i} = \gamma_{30} + \eta_{3i}, \tag{3.12}$$

where $(\eta_{0i}, \eta_{2i}, \eta_{3i})$ is multi-variate normally distributed, with mean $(0, 0, 0)$ and covariance matrix

$$\mathbf{T} = \begin{bmatrix} \tau_0^2 & \tau_{20} & \tau_{30} \\ \tau_{20} & \tau_2^2 & \tau_{32} \\ \tau_{30} & \tau_{32} & \tau_3^2 \end{bmatrix}.$$

In other words, in the absence of treatment, cases vary in their average levels of the outcome and in their rates of change; furthermore, the treatment has variable effects, altering the rate of change by more for some cases and less for others. Finally, note that the design-comparable effect size is equivalent to that given in (3.11), because MB5 differs from MB4 only in the variability of the treatment effect, rather than in how baseline variability is described.

3.1.3. Further models for multiple baseline designs

The five models presented are far from an exhaustive list of the possible specifications. For example, one might assume in MB4 that the time trend is constant across phases, so that $\beta_{3i} = 0$. Alternately, any of the models might be extended through the addition of polynomial time trends. In principle, one could specify a model in which any or all of the case-level regression coefficients vary randomly, though in practice the number of

randomly varying coefficients will need to be tempered by the number of cases m being modeled. I return to the question of how the number of random efffects relates to the number of cases in Section 3.4.

3.2. Models for treatment reversal designs

Compared to models for the multiple baseline design, causally interpretable models for the treatment reversal design need to accomodate a larger number of possible treatment assignment schedules, in order to allow that treatments can be removed and re-introduced. As a consequence, the treatment reversal design requires more highly structured models for individual cases. Suppose that the design measures outcomes at equally-spaced times $j = 1, ..., n$, on each of m cases. Recall from Section 2.3.2 that a causally interpretable model specifies a structural form for the function $Y_{ij}\left(\tilde{T}_{ij}\right)$, where T_{ij} indicates whether treatment is present or absent for case i and time j and $\tilde{T}_{ij}$ denotes the treatment history through time j. This section describes several such models, along with possible group-level specifications and the corresponding design-comparable effect sizes. I again present a catalog of model specifications, though the focus is on different case-level assumptions, rather than only group-level assumptions as with the multiple baseline design.

Given a causally interpretable case-level model and assumptions about how the model varies across cases, a design-comparable treatment effect can be specified by an implementation time A and a treatment schedule to be followed until outcome measurement time B. As discussed in Section 2.4, I will assume that the schedule of primary interest is one of continuous treatment between times A and B, so that $\tilde{T}_{iB} = 0^A 1^{(B-A)}$, compared to a schedule of no treatment through time B. A design-comparable d-type effect size is

then given by

$$\delta_{AB} = \frac{\mathrm{E}\left[Y_{iB}\left(0^{A}1^{(B-A)}\right)\right] - \mathrm{E}\left[Y_{iB}\left(0^{B}\right)\right]}{\sqrt{\mathrm{Var}\left[Y_{iB}\left(0^{B}\right)\right]}} \tag{3.13}$$

3.2.1. Models with transient treatment effects

Hedges et al. (2012b) studied the simplest possible model for the treatment reversal design. At the individual level, they assumed that baseline outcomes are stable (lacking trend) and that the treatment causes a shift in the level of the outcome that is entirely transient (c.f., Equation (2.7) in Section 2.3.2). The case-level model is then:

$$Y_{ij}\left(\tilde{T}_{ij}\right) = Y_{ij}\left(T_{ij}\right) = \beta_{0i} + \beta_{1i}T_{ij} + \epsilon_{ij}, \tag{3.14}$$

where, following conventional assumptions for statistical models of single-case data, the errors $(\epsilon_{i1}, ..., \epsilon_{in})$ are normally distributed with mean zero, variance σ^2, and first-order autocorrelation ϕ. Models TR1 and TR2 use this case-level model but describe different assumptions about variation across cases.

Model TR1: Transient, constant treatment effect

Hedges et al. (2012b) assumed that the mean outcomes in the absence of treatment β_{1i} varied across cases, but that the treatment effect is constant:

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \qquad \beta_{1i} = \gamma_{10} \tag{3.15}$$

where η_{0i} is normally distributed with mean zero and variance τ_0^2.[2] As noted by Hedges et al. (2012a), Model MB1 for the multiple baseline design can be seen as a special case of model TR1. The latter model includes a very wide set of possible treatment schedules, but reduces to the former model if restricted to schedules where the treatment is never removed once it is introduced. As a result, the effect size parameter for TR1 has the same form as that for MB1. Noting that $E\left[Y_{iB}\left(0^B\right)\right] = \gamma_{00}$, $E\left[Y_{iB}\left(0^A 1^{B-A}\right)\right] = \gamma_{00} + \gamma_{10}$, and $\text{Var}\left[Y_{iB}\left(0^B\right)\right] = \text{Var}(\eta_{0i} + \epsilon_{ij}) = \tau_0^2 + \sigma^2$, it follows from (3.13) that

$$\delta_{AB} = \frac{\gamma_{10}}{\sqrt{\tau_0^2 + \sigma^2}}, \tag{3.16}$$

which is identical to (3.5). Also note that, just as when the model is specialized to the multiple baseline design, the effect size does not depend on the chosen design parameters A and B. In Section 3.3, I will describe a method for estimating the effect size that differs from the method described by Hedges et al. (2012b).

Model TR2: Transient, varying treatment effects

The constant treatment effect assumption may be overly restrictive, and can be relaxed by allowing the treatment effect to vary across cases (just as Model MB2 relaxes MB1). Retaining the case-level specification of (3.14), the group-level specification becomes

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \quad \beta_{1i} = \gamma_{10} + \eta_{1i}, \tag{3.17}$$

[2] In the notation of Hedges et al. (2012b), $\gamma_{00} = \mu^C$, $\gamma_{10} = \mu^T - \mu^C$, $\eta_{0i} = \eta_i$, and $\tau_0^2 = \tau^2$.

where (η_{0i}, η_{1i}) is multi-variate normally distributed, with mean $(0, 0)$ and covariance matrix

$$\mathbf{T} = \begin{bmatrix} \tau_0^2 & \tau_{10} \\ \tau_{10} & \tau_1^2 \end{bmatrix}.$$

As this modification does not affect the variance in the absence of treatment, the effect size for Model TR2 is identical to that for TR1.

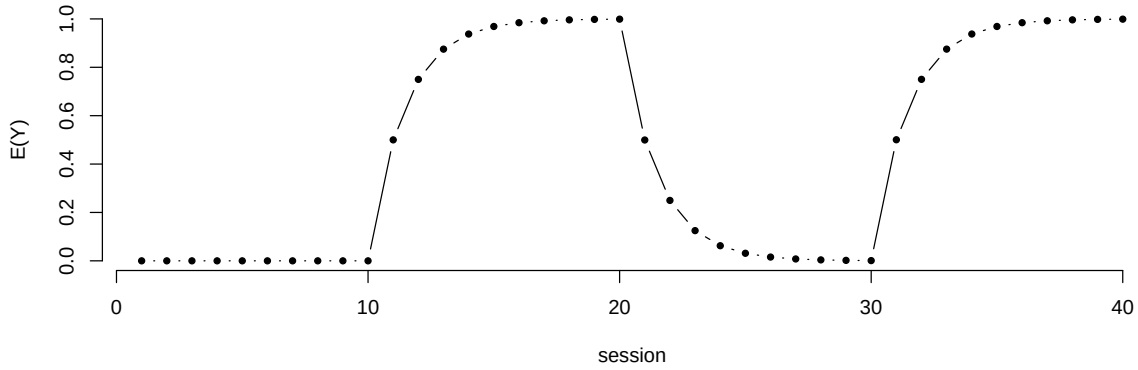
3.2.2. Models with decaying treatment effects

The assumption that treatment effects are entirely transient is a strong one, and is particularly restrictive because it assumes that the full effect of the treatment is realized immediately upon introduction. It is useful to entertain models that relax this structural assumption while retaining a causally interpretable structure. In one such model, treatment impacts are only fully realized after the treatment is in place for a length of time. The potential outcome for case i at time j depends on the treatment schedule as follows:

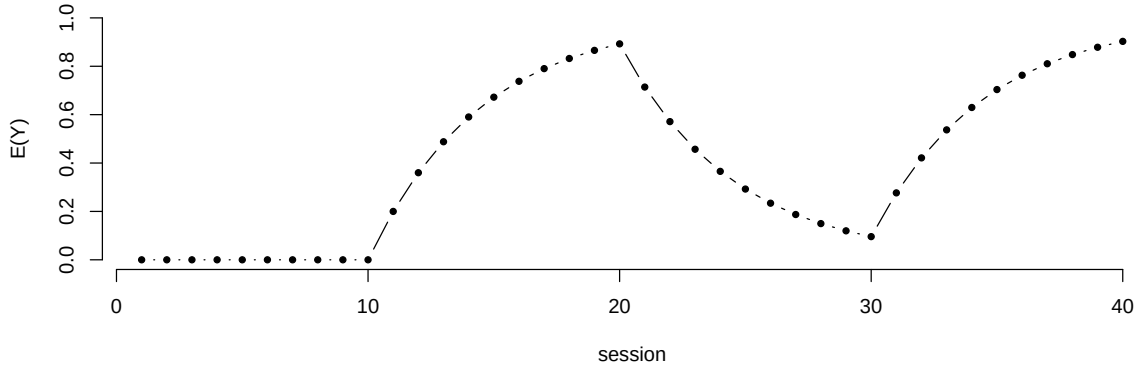
$$Y_{ij}\left(\tilde{T}_{ij}\right) = \beta_{0i} + \beta_{1i}(1-\omega)\sum_{k=1}^{j}\omega^{(j-k)}T_{ik} + \epsilon_{ij}, \tag{3.18}$$

for $\omega \in [0, 1)$ and $(\epsilon_{i1}, ..., \epsilon_{in})$ following an auto-regressive process with first-order auto-correlation ϕ and variance σ^2, as in previous models. This case-level regression is a simple example of the intervention analysis model described by Box and Tiao (1975).

Equation (3.18) is somewhat more complicated than any of the previous case-level models due to its non-linearity in the parameter ω. I offer several remarks about interpretation. First, one can view this model as an expression of a weaker form of causal transience, in which the presence of the treatment at a given time wears off according to



(a) $\omega = 0.5$



(b) $\omega = 0.8$

Figure 3.1. Mean outcome process for Model TR3, with $\beta_{0i} = 0$, $\beta_{1i} = 1$, and varying values of ω.

a pattern that does not depend on the presence of treatment at any other time. Specifically, the presence of the treatment at time k leads to an immediate impulse response of $\beta_{1i}(1 - \omega)$, the magnitude of which decays geometrically over time at a rate of ω per measurement occasion. These individual responses accumulate if the treatment is present without interruption over subsequent time-points, eventually reaching an asympote of β_{1i}. Therefore, β_{1i} represents the long-run or equilibrium effect of sustained intervention, while β_{0i} retains the interpretation of the mean level of the outcome for case i in the absence of any intervention.

Second, note that the decay parameter ω controls the degree of curvature in the path of the mean outcome over time. At one extreme, $\omega = 0$ implies that $Y_{ij}\left(\tilde{T}_{ij}\right) = \beta_{0i}+\beta_{1i}T_{ij}+\epsilon_{ij}$, equivalent to Model TR1, in which the potential outcomes depend only on the contemporaneous treatment status. At the opposite extreme, as ω increases towards one, the process takes a longer time to reach equilibrium, and the mean path more closely resembles a linear trend.[3] Figure 3.1 illustrates how the decay rate affects the outcome process: the mean outcome over time is plotted for an ABAB design in which treatment is introduced after time 10, removed after time 20, and re-introduced after time 30; the top panel uses a decay rate of $\omega = 0.5$, while the bottom panel uses a more gradual decay rate of $\omega = 0.8$.

Third, a geometrically decaying impulse response can be justified both by its relative simplicity and by its close connection to first-order auto-regression, which is the most prominent assumption entertained about the error processes in single-case data. Observe that if $\omega = \phi$, so that the impulse decays at the same rate as the correlation among the errors, then (3.18) reduces to a particularly simple auto-regressive equation:

$$Y_{ij}\left(\tilde{T}_{ij}\right) = \beta^*_{0i} + \beta^*_{1i}T_{ij} + \phi Y_{i,j-1}\left(\tilde{T}_{i,j-1}\right) + \epsilon^*_{ij}, \tag{3.19}$$

for $j = 2, ..., n$, where $\beta^*_{0i} = (1-\phi)\beta_0$, $\beta^*_{1i} = \beta_1(1-\phi)$, and $\epsilon^*_{i2}, ..., \epsilon^*_{in}$ are independent, normally distributed errors with mean zero and variance $(1-\phi^2)\sigma^2$.

It remains to specify the group-level assumptions for the model. As in Models TR1 and TR2, one might entertain the assumption that the equilibrium treatment effect is constant or that it varies across cases. Models TR3 and TR4 presents each of these

[3]In the limit, the model comes to resemble a stochastic counterpart of the cumulative exposure model given in Equation (2.5).

assumptions in turn. In each model, the curvature ω is treated as a nuisance parameter and so assumed to be constant across cases, similar to how the auto-correlation ϕ has been treated in this and previous models.

Model TR3: Decaying, constant treatment effect

Assuming that the equilibrium treatment effect is constant across cases, the group-level specification is

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \quad \beta_{1i} = \gamma_{10}, \tag{3.20}$$

where η_{0i} is normally distributed with mean zero and variance τ_0^2. As in Models TR1 and TR2, the mean outcome (across cases) in the absence of treatment is $E\left[Y_{iB}\left(0^B\right)\right] = \gamma_{00}$, with variance $\text{Var}\left[Y_{iB}\left(0^B\right)\right] = \text{Var}(\eta_{0i} + \epsilon_{ij}) = \tau_0^2 + \sigma^2$. However, unlike in other models the mean outcome at time B depends on the length of the treatment course:

$$E\left[Y_{iB}\left(0^A 1^{B-A}\right)\right] = \gamma_{00} + \gamma_{10}(1-\omega)\sum_{k=A+1}^{B}\omega^{(B-k)} = \gamma_{00} + \gamma_{10}\left(1-\omega^{B-A}\right).$$

The effect size parameter for TR3 therefore has the form

$$\delta_{AB} = \frac{\gamma_{10}\left(1-\omega^{B-A}\right)}{\sqrt{\tau_0^2+\sigma^2}}. \tag{3.21}$$

Note that if $B-A$ is sufficiently large–that is, if the treatment schedule of interest involves sustained intervention–then the multiplicative term involving ω will be close to one and

the effect size parameter will be close to

$$\delta_{A\infty} = \frac{\gamma_{10}}{\sqrt{\tau_0^2 + \sigma^2}}. \tag{3.22}$$

This might be called the "equilibrium" effect size parameter, and is identical to (3.16).

Model TR4: Decaying, varying treatment effect

As with previous models, the constant treatment effect assumption in TR3 may be overly restrictive, and can be relaxed by allowing the treatment effect to vary across cases. With the case-level specification of (3.18), the group-level specification becomes

$$\beta_{0i} = \gamma_{00} + \eta_{0i}, \quad \beta_{1i} = \gamma_{10} + \eta_{1i}, \tag{3.23}$$

where (η_{0i}, η_{1i}) is multi-variate normally distributed, with mean $(0, 0)$ and covariance matrix

$$\mathbf{T} = \begin{bmatrix} \tau_0^2 & \tau_{10} \\ \tau_{10} & \tau_1^2 \end{bmatrix}.$$

Because the term involving the decay parameter in the numerator of the effect size does not vary across cases, allowing the equilibrium treatment effect to vary across cases does not change the form of the mean treatment effect across cases. Also, as with previous models, the modification does not affect the variance in the absence of treatment. All together, this means that the effect size for Model TR4 is identical to that for TR3, as given in (3.21).

3.2.3. A causally ambiguous model

Some proposals for parametric models of treatment reversal designs are not causally interpretable, even if they might be adequate descriptions of observed data. For instance, consider a piece-wise linear regression model for an ABAB design, as described by Hershberger et al. (1999) or Swaminathan et al. (2010), among others. Suppose that the case is in baseline phase A_1 for measurements $j = 1, ..., n_1$, treatment phase B_1 for measurements $j = n_1 + 1, ..., n_2$, return-to-baseline phase A_2 for measurements $j = n_2 + 1, ..., n_3$, and treatment re-introduction phase B_2 for measurements $j = n_3 + 1, ..., n$. The model can then be written as

$$
\begin{aligned}
(3.24) \quad Y_{ij} = {} & \beta_{i0} I(j \leq n_2) + \beta_{i1} j \times I(j \leq n_2) \\
& + \beta_{i2} T_{ij} \times I(j \leq n_2) + \beta_{i3} T_{ij} \times (j - n_1) \times I(j \leq n_2) \\
& + \beta_{i4} I(j > n_2) + \beta_{i5} j \times I(j > n_2) \\
& + \beta_{i6} T_{ij} \times I(j > n_2) + \beta_{i7} T_{ij} \times (j - n_3) \times I(j > n_2) + \epsilon_{ij}.
\end{aligned}
$$

Here, the mean outcome in each phase is described by a linear trend. In (3.24), β_{i2}, β_{i3} describes the change in level and change in trend from phase A_1 to phase B_1, while β_{i6}, β_{i7} describe the same quantities from phase A_2 to phase B_2. Some authors have used slightly different parameterizations, such as describing each phase using one term for the level at the beginning of the phase and another term for the slope during the phase.

While this model may fit the observed data reasonably well, it is in a sense too weak as a model for the potential outcome process, regardless of how it is parameterized. It is

inadequate because it does not express unambiguously how the process would change if the phase lengths were altered or if additional treatment reversals were added. For instance, consider the outcome at the beginning of phase A_2, when $j = n_2 + 1$. The observed outcome is $Y_{i,n_2+1}\,(0^{n_1}1^{n_2-n_1}0)$. What if instead of this pattern of treatment assignment, the experimenter had assigned the case to continue in baseline for the entire time, so $\tilde{T}_{in} = \{0^{n_2+1}\}$? Does $\mathrm{E}\,[Y_{ij}(0^{n_2+1})] = \mathrm{E}\,[Y_{ij}(0^{n_1}1^{n_2-n_1}0)]$? Or should one assume that the initial baseline phase would be continued up to time n_2+1, so that $Y_{ij} = \beta_1 + \beta_5 j + \epsilon_{ij}$ for $j = 1, ..., n_2 + 1$? A model that cannot address this and similar questions is not causally interpretable.

3.3. Restricted maximum likelihood (RML) estimation

The previous sections have presented a variety of models for single case designs with continuous, interval-scale outcome measures and demonstrated how to use those models to identify design-comparable standardized mean difference effect sizes. Most of the models discussed fall into the general category of hierarchical linear models, the exceptions being Models TR3 and TR4, which involve one non-linear parameter. This section describes one method for estimating the hierarchical linear models, focusing specifically on how to estimate the effect size parameter of interest. I begin by laying out notation that is general enough to encompass any of the hierarchical linear models described in this chapter. Next, I review RML estimation, focusing in particular on the estimates generated by most software packages, and discuss methods for handling software convergence problems that arise due to estimates on the boundary of the parameter space. I then describe a bias-corrected effect size estimator, discuss issues related to parameterization of the variance

components, and note that a certain re-parameterization leads to simplified calculations. Finally, I briefly discuss a rough approximation for estimating the non-linear models TR3 and TR4.

3.3.1. Notation and general model

Mixed-effect specifications of the models described in previous sections can be expressed compactly using matrix notation. Let $\mathbf{y}_i = (y_{i1}, ..., y_{in_i})'$ be the $(n_i \times 1)$ vector of outcome data from case i, excluding missing observations. Because these data points did not necessarily occur on subsequent measurement occasions, some notation is needed to indicate the measurement occasion corresponding to each non-missing observation; thus let $(j_{i1}, ..., j_{in_i})$ denote the measurement occasions of the non-missing observations. Let $\mathbf{X}_i$ be the $(n_i \times p)$ design matrix corresponding to the terms in the regression model. Let $\mathbf{Z}_i$ be an $(n_i \times q)$ design matrix with one column for each of the regression coefficients that is allowed to vary at random; $\mathbf{Z}_i$ will typically consist of a subset of the columns of $\mathbf{X}_i$. Let $\boldsymbol{\gamma} = (\gamma_{00}, ..., \gamma_{(p-1)0})^T$ be the $(p \times 1)$ vector of fixed effects and $\boldsymbol{\eta}_i = (\eta_{0i}, ..., \eta_{(q-1)i})'$ be the $(q \times 1)$ vector of random effects for case i. Finally, let $\mathbf{A}_i$ be the $(n_i \times n_i)$ correlation matrix of the individual errors $(\epsilon_{i1}, ..., \epsilon_{in_i})'$, with st^{th} cell $[\mathbf{A}]_{st} = \phi^{|j_{is} - j_{it}|}$, and let $\mathbf{T}^{(q)}$ be the $(q \times q)$ covariance matrix of $\boldsymbol{\eta}_i$.

Conditional on the random effects, $(\mathbf{y}_i | \boldsymbol{\eta}_i)$ is multivariate normally distributed with mean $\mathbf{X}_i\boldsymbol{\gamma} + \mathbf{Z}_i\boldsymbol{\eta}_i$ and covariance matrix $\sigma^2\mathbf{A}_i$. After taking expectations over the distribution of random effects, the unconditional distribution of $\mathbf{y}_i$ is also multivariate normal, with mean $\mathbf{X}_i\boldsymbol{\gamma}$ and covariance matrix $\mathbf{Z}_i\mathbf{T}^{(q)}\mathbf{Z}_i' + \sigma^2\mathbf{A}_i$. The unconditional model for all

m cases can therefore be written as

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\gamma}, \mathbf{V}) \tag{3.25}$$

where $\mathbf{y} = (\mathbf{y}_1', ..., \mathbf{y}_m')'$, $\mathbf{X} = (\mathbf{X}_1', ..., \mathbf{X}_m')'$, $\mathbf{Z} = \oplus_{i=1}^m \mathbf{Z}_i$, $\mathbf{T} = \mathbf{I}_m \otimes \mathbf{T}^{(q)}$, $\mathbf{A} = \oplus_{i=1}^m \mathbf{A}_i$, and $\mathbf{V} = \mathbf{Z}\mathbf{T}\mathbf{Z}' + \sigma^2\mathbf{A}$; here $\oplus$ creates a block-diagonal matrix out of a sequence of sub-matrices, $\otimes$ is the Kronecker product, and $\mathbf{I}_m$ is an $(m \times m)$ identity matrix. Let $N = \sum_{i=1}^m n_i$ be the total number of non-missing observations, so that $\mathbf{y}$ and $\mathbf{X}\boldsymbol{\gamma}$ are $N \times 1$ and $\mathbf{V}$ has dimension $N \times N$. As a last point of notation, let $\boldsymbol{\theta}$ be an $(r \times 1)$ vector that collects all of the parameters related to the covariance matrices; below, $\mathbf{A}$, $\mathbf{T}$, and $\mathbf{V}$ are to be understood as functions of $\boldsymbol{\theta}$.

With this notation established, any of the effect sizes from the linear models in Sections 3.1 and 3.2 can be expressed as the ratio of a linear combination of the fixed effects to the square root of a linear combination of the variance parameters. Assuming an appropriate parameterization of the variance components,

$$\delta_{AB} = \frac{\mathbf{p}'\boldsymbol{\gamma}}{\sqrt{\mathbf{r}'\boldsymbol{\theta}}} \tag{3.26}$$

for suitably chosen vectors $\mathbf{p}$ and $\mathbf{r}$. For example, in Model MB1, take $\boldsymbol{\gamma} = (\gamma_{00}, \gamma_{10})'$ and $\boldsymbol{\theta} = (\sigma^2, \phi, \tau_0^2)'$. Setting $\mathbf{p} = (0, 1)'$ and $\mathbf{r} = (1, 0, 1)'$ makes (3.26) equivalent to (3.5). In Model MB4, take $\boldsymbol{\gamma} = (\gamma_{00}, \gamma_{10}, \gamma_{20}, \gamma_{30})'$ and $\boldsymbol{\theta} = (\sigma^2, \phi, \tau_0^2, \tau_{20}, \tau_2^2)'$. For a given choice of A, B, and centering point C, setting $\mathbf{p} = (0, 1, 0, B - A)'$ and $\mathbf{r} = (1, 0, 1, 2[B - C], [B - C]^2)'$ makes (3.26) equivalent to (3.11).

3.3.2. Estimation of fixed effects and variance components

As previously noted, many software packages provide implementations of restricted maximum likelihood estimation for hierarchical linear models. All of these software packages provide RML estimates of both the fixed effects $\boldsymbol{\gamma}$ and variance component parameters $\boldsymbol{\theta}$, as well as an approximate covariance matrix for the fixed effects. All of the packages also provide some form of approximate covariance matrix of the variance components, though they differ in how this covariance matrix is estimated. I now briefly review how these parameter estimates and covariances are generated.

RML estimation involves two stages, beginning with variance parameter estimation, followed by fixed effect estimation. For explanatory purposes, it is helpful to begin with the second stage, supposing that the variance parameters in $\boldsymbol{\theta}$ are all known (or equivalently, that $\mathbf{V}$ is known). In this case, the only unknowns are the fixed effects $\boldsymbol{\gamma}$, which can be estimated efficiently using weighted least squares. Define the weighted least squares estimate

$$\tilde{\boldsymbol{\gamma}} = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \tag{3.27}$$

which is multivariate normally distributed with mean $\boldsymbol{\gamma}$ and covariance matrix

$$\mathrm{Cov}(\tilde{\boldsymbol{\gamma}}) = \left(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\right)^{-1} \tag{3.28}$$

if $\mathbf{V}$ is known. Of course, in the present problem, the variance parameters must be estimated. RML estimates of the fixed effects use estimates of the variance parameters $\hat{\boldsymbol{\theta}}$ to form an estimate $\hat{\mathbf{V}} = \hat{\mathbf{V}}(\hat{\boldsymbol{\theta}})$, which is then used in place of $\mathbf{V}$ in (3.27) and (3.28).

Thus the RML fixed effect estimates are

$$\hat{\boldsymbol{\gamma}} = \left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \tag{3.29}$$

with estimated covariance matrix

$$\mathbf{C}(\hat{\boldsymbol{\gamma}}) = \left(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}\right)^{-1} \tag{3.30}$$

It is known that $\mathbf{C}(\hat{\boldsymbol{\gamma}})$ tends to underestimate the true covariance of $\hat{\boldsymbol{\gamma}}$ in small samples; Kenward and Roger (1997, 2009) provide more elaborate covariance estimators intended for use in small samples.

Expressions (3.29) and (3.30) are based on estimated values of the variance components that have yet to be discussed. Except in certain special cases and simple models, it is not possible to obtain closed-form expressions for these estimates; instead, RML estimates of $\boldsymbol{\theta}$ are obtained by maximizing the log of the restricted likelihood corresponding to (3.25) via iterative numerical methods. As given for instance in Lindstrom and Bates (1988), the log of the restricted likelihood is

$$-2l_R(\boldsymbol{\theta}|\mathbf{y}) = \log\left|\mathbf{X}'\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{X}\right| + \log|\mathbf{V}(\boldsymbol{\theta})| + \mathbf{y}'\mathbf{Q}(\boldsymbol{\theta})\mathbf{y}, \tag{3.31}$$

where $\mathbf{Q}(\boldsymbol{\theta}) = \mathbf{V}^{-1}(\boldsymbol{\theta}) - \mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{X}\left[\mathbf{X}'\mathbf{V}^{-1}(\boldsymbol{\theta})\mathbf{X}\right]^{-1}\mathbf{X}'\mathbf{V}^{-1}(\boldsymbol{\theta})$. Let $\hat{\boldsymbol{\theta}}$ denote the values that maximize $l_r(\boldsymbol{\theta}|\mathbf{y})$.

It is possible for the RML estimate of one or more variance components to lie on the boundary of its parameter space; such boundary estimates are particularly common when estimates are based on data from only a few independent cases. Depending on the

software package and the parameterization used to maximize the restricted likelihood, boundary estimates can lead to algorithmic convergence issues. I discuss methods for addressing these issues in Section 3.3.3.

Estimates of the sampling variance of $\hat{\boldsymbol{\theta}}$ are needed to calculate a small-sample correction for the effect size estimator and to estimate the variance of the effect size. Different software implementations use different methods to estimate $\text{Cov}(\hat{\boldsymbol{\theta}})$. This is often done using the inverse of the observed, expected, or average Fisher Information matrix (see for instance Gilmour, Thompson, & Cullis, 1995); any of these provides an approximate estimate of the covariance, valid as sample size grows large. The observed information matrix has entries

$$\begin{aligned}(3.32)\qquad \left[\mathcal{I}_O^{\boldsymbol{\theta}}\right]_{st} &= -\frac{\partial^2 l_R(\boldsymbol{\theta}|\mathbf{y})}{\partial\theta_s\partial\theta_t}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &= \frac{1}{2}\mathbf{y}'\mathbf{Q}\left(\dot{\mathbf{V}}_s\mathbf{Q}\dot{\mathbf{V}}_t + \dot{\mathbf{V}}_t\mathbf{Q}\dot{\mathbf{V}}_s - \ddot{\mathbf{V}}_{st}\right)\mathbf{Q}\mathbf{y} - \frac{1}{2}\text{tr}\left(\mathbf{Q}\dot{\mathbf{V}}_s\mathbf{Q}\dot{\mathbf{V}}_t - \mathbf{Q}\ddot{\mathbf{V}}_{st}\right)\end{aligned}$$

for $s,t = 1,...,r$, where $\dot{\mathbf{V}}_s = \partial\mathbf{V}/\partial\theta_s|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, $\ddot{\mathbf{V}}_{st} = \partial^2\mathbf{V}/\partial\theta_s\partial\theta_t|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$, and $\mathbf{Q} = \mathbf{Q}(\hat{\boldsymbol{\theta}})$. The expected information matrix is the expected value of $\mathcal{I}_O$ over the distribution of $\mathbf{y}$ (conditional on $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$), with entries

$$(3.33)\qquad \left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]_{st} = \frac{1}{2}\text{tr}\left(\mathbf{Q}\dot{\mathbf{V}}_s\mathbf{Q}\dot{\mathbf{V}}_t\right)$$

for $s,t = 1,...,r$. The average information matrix, used in the program `ASReml`, is the arithmetic average of $\mathcal{I}_O^{\boldsymbol{\theta}}$ and $\mathcal{I}_E^{\boldsymbol{\theta}}$, where quadratic terms involving second derivatives of $\mathbf{V}$ are approximated by their expectations (Gilmour et al., 2009); this leads to computational

efficiencies for large matrices. The entries are given by

$$\left[\mathcal{I}_A^{\boldsymbol{\theta}}\right]_{st} = \frac{1}{2}\mathbf{y}'\mathbf{Q}\dot{\mathbf{V}}_s\mathbf{Q}\dot{\mathbf{V}}_t\mathbf{Q}\mathbf{y}, \tag{3.34}$$

for $s, t = 1, ..., r$. An estimate of the covariance matrix of $\hat{\boldsymbol{\theta}}$ is given by the inverse of one of these information matrices; thus, define $\mathbf{C}(\hat{\boldsymbol{\theta}}) = \mathcal{I}^{-1}$, where $\mathcal{I}$ indicates $\mathcal{I}_O^{\boldsymbol{\theta}}$, $\mathcal{I}_E^{\boldsymbol{\theta}}$, or $\mathcal{I}_A^{\boldsymbol{\theta}}$.

In order for $\mathcal{I}^{-1}$ to provide a valid estimate of $\text{Cov}(\hat{\boldsymbol{\theta}})$, the model must be parameterized in terms of variances and covariances such as τ_0^2 and τ_{20}, rather than, for instance, log-standard deviations and correlations. Because some software implementations use alternative parameterizations, it may be necessary to convert the information matrix from one parameterization to another. Suppose that the software uses a parameterization $\boldsymbol{\psi} = g(\boldsymbol{\theta})$, where $g()$ is a one-to-one function with inverse $h()$. Write the Jacobian matrix of the inverse as

$$\nabla_{\boldsymbol{\psi}} h = \left[\frac{\partial h_s(\boldsymbol{\psi})}{\partial \psi_t}\right]_{s,t=1,...,r}. \tag{3.35}$$

If software reports the inverse of the information matrix for $\boldsymbol{\psi}$ (whether observed, expected, or average), along with RML estimates $\hat{\boldsymbol{\psi}}$, the approximate covariance matrix for $\hat{\boldsymbol{\theta}}$ can be calculated as

$$\mathbf{C}(\hat{\boldsymbol{\theta}}) = (\nabla_{\boldsymbol{\psi}} h)\left[\mathcal{I}^{\psi}\right]^{-1}(\nabla_{\boldsymbol{\psi}} h)'. \tag{3.36}$$

3.3.3. Boundary estimates in RML

An important characteristic of both FML and RML estimation is that they can produce estimates that lie on a boundary of the parameter space. For example, the variance of the

random intercepts in Model TR1 may be estimated as 0; or in Model TR2, the estimate of the covariance τ_{10} can imply a correlation of exactly one between the random effects η_{0i} and η_{1i}. For purposes of maximizing the (restricted) likelihood, the estimation algorithm may use a parameterization that maps each parameter to the real line, in which case boundary maximums will manifest as convergence problems where the numerical algorithm could run endlessly without reaching the maximum. Both in practice and in order to execute realistic simulations, one needs some method for handling these boundary estimates.

Faced with a boundary estimate (such as a variance component estimate of zero), the applied researcher will in practice take one of several possible courses. One approach would be to simply accept the boundary estimates. Another approach would be to use the boundary estimate for purposes of calculating the RML effect size estimator (3.37), but substitute a subjectively chosen value of the parameter for purposes of evaluating the degrees of freedom ν. This is approach is similar to one suggested by Longford (2000). Alternatively, one might re-specify the model with a more constrained parameterization, such as fixing the correlation between random effects in Model TR2 to a subjectively chosen value. An even more expedient tactic would be to reduce the number of random effects in the model, for instance by moving from Model TR2 to TR1, conditional on receiving an RML estimate of τ_1^2 equal to zero. Finally, if one uses an algorithm constructed to maintain estimates within the parameter space, one could stop the algorithm after a given number of iterations, even if it has not converged. This last approach would yield estimates close to–but not exactly on–a boundary of the parameter space. Though estimates generated by such a strategy would not actually maximize the restricted likelihood and also have certain conceptual flaws, they are accessible and practical from the point

of view of the applied researcher. In the interest of evaluating RML estimation as it will likely be used in practice, this final approach is used in the simulation studies reported in the following section.

3.3.4. Effect size estimation

An initial estimate of the effect size can be formed by substituting the RML estimates $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\theta}}$ in place of the corresponding parameters in (3.26), letting

$$\hat{\delta}_{AB} = \frac{\mathbf{p}'\hat{\boldsymbol{\gamma}}}{\sqrt{\mathbf{r}'\hat{\boldsymbol{\theta}}}}. \tag{3.37}$$

I will refer to this as the RML estimate. Without further adjustment, the RML estimate is approximately unbiased if the number of cases m is sufficiently large. However, for sample sizes typically found in multiple baseline designs, $\hat{\delta}_{AB}$ may nonetheless exhibit substantial bias even if $\hat{\boldsymbol{\gamma}}$ is exactly unbiased and $\hat{\boldsymbol{\theta}}$ approximately unbiased.

The bias of $\hat{\delta}_{AB}$ is analogous to the small-sample bias of the Cohens d statistic from a between-subjects experiment, which can be corrected using methods described by Hedges (1981); the corrected effect size estimate is sometimes referred to as Hedges g. The exact distribution theory used in Hedges g statistic is not available for the present problem, due to the presence of nuisance parameters among the variance components. Still, one can approximate the sampling distribution of $\hat{\delta}_{AB}$ by a Student-t distribution, thereby obtaining an approximate small-sample bias correction and an approximate expression for the variance of the effect size estimate.

In developing the approximation, I will treat $\mathbf{r}'\hat{\boldsymbol{\theta}}$ as an unbiased estimate of $\mathbf{r}'\boldsymbol{\theta}$, even though this is only approximately true as m grows larger. For fairly simple specifications

such as Model MB1, the magnitude of the bias is small; in general the bias will depend on the degree of imbalance in the data and the number of random effects in the model. In another approach to approximation, one could also adjust the RML estimate $\mathbf{r}'\hat{\boldsymbol{\theta}}$ using an approximation to its bias. I provide details regarding this more elaborate approach in Appendix A. However, based on initial simulation studies reported in Section 3.4, I have found that such further bias adjustment makes little difference for estimating the effect size.

I now describe an approximation to the distribution of the effect size estimator. Define the constant

$$\kappa = \sqrt{\frac{\mathbf{p}'\mathbf{C}(\hat{\boldsymbol{\gamma}})\mathbf{p}}{\mathbf{r}'\hat{\boldsymbol{\theta}}}}. \tag{3.38}$$

From a theorem given in Hedges (2007), it then follows that the distribution of $\hat{\delta}_{AB}/\kappa$ can be approximated by a non-central t distribution with ν degrees of freedom and non-centrality parameter δ_{AB}/κ, where

$$\nu = \frac{2(\mathbf{r}'\hat{\boldsymbol{\theta}})^2}{\mathbf{r}'\mathbf{C}(\hat{\boldsymbol{\theta}})\mathbf{r}}. \tag{3.39}$$

It follows further that a bias-corrected effect size estimator is given by

$$g_{AB} = J(\nu) \times \hat{\delta}_{AB}, \tag{3.40}$$

where $J(x) = 1 - 3/(4x - 1)$, and that g_{AB} has approximate variance

$$\text{Var}(g_{AB}) \approx J(\nu)^2 \left[\frac{\nu\kappa^2}{\nu - 2} + \delta_{AB}^2 \left(\frac{\nu}{\nu - 2} - \frac{1}{J(\nu)^2}\right)\right]. \tag{3.41}$$

Substituting g_{AB} for δ_{AB} produces an estimate of the variance of g_{AB}.

3.3.5. A reparameterization

A particular re-parameterization of the variance components provides insight into why the effect size estimate is approximately t-distributed and leads to simplified expressions for $\hat{\delta}_{AB}$, κ, and ν. Parameterize the variance components by the vector $\boldsymbol{\psi} = g(\boldsymbol{\theta})$, with first entry $\psi_1 = \mathbf{r}'\boldsymbol{\theta}$ and remaining entries $\boldsymbol{\psi}_*$ that create a one-to-one mapping from $\boldsymbol{\theta}$; let $h = g^{-1}$. With this parameterization, the effect size (3.26) is then simply $\delta_{AB} = \mathbf{p}'\boldsymbol{\gamma}/\sqrt{\psi_1}$ and the covariance matrix of $\mathbf{y}$ can be written as $\mathbf{V} = \psi_1 \mathbf{W}(\boldsymbol{\psi}_*)$.

Let $\hat{\boldsymbol{\psi}}$ denote the RML estimate of $\boldsymbol{\psi}$, $\hat{\mathbf{W}} = \mathbf{W}(\hat{\boldsymbol{\psi}})$, and

$$\mathbf{R} = \hat{\mathbf{W}}^{-1} - \hat{\mathbf{W}}^{-1}\mathbf{X}\left[\mathbf{X}'\hat{\mathbf{W}}^{-1}\mathbf{X}\right]^{-1}\mathbf{X}'\hat{\mathbf{W}}^{-1}.$$

Also define $\dot{\mathbf{W}}_s = \partial\mathbf{W}/\partial\psi_s|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}$ and $\ddot{\mathbf{W}}_{st} = \partial^2\mathbf{W}/\partial\psi_s\partial\psi_t|_{\boldsymbol{\psi}=\hat{\boldsymbol{\psi}}}$ for $s, t = 2, ..., r$. The first derivative of the log restricted likelihood (3.31) with respect to θ is then

$$\frac{\partial l_R(\boldsymbol{\psi}|\mathbf{y})}{\partial\psi_1} = \frac{1}{2\psi_1^2}\mathbf{y}'\mathbf{R}\mathbf{y} - \frac{N-p}{2\psi_1}.$$

It follows that the RML estimator of ψ_1 is

$$\hat{\psi}_1 = \frac{\mathbf{y}'\mathbf{R}\mathbf{y}}{N-p}, \tag{3.42}$$

from which it is clear that $\hat{\psi}_1$ is a quadratic form in $\mathbf{y}$, conditional on the remaining parameter vector $\boldsymbol{\psi}_*$.

Furthermore, simpler expressions are available for κ and for ν under this parameterization. Since $\mathbf{C}(\hat{\boldsymbol{\gamma}}) = \hat{\psi}_1 \left(\mathbf{X}'\hat{\mathbf{W}}^{-1}\mathbf{X}\right)^{-1}$, it follows that $\kappa = \sqrt{\mathbf{p}' \left(\mathbf{X}'\hat{\mathbf{W}}^{-1}\mathbf{X}\right)^{-1} \mathbf{p}}$. Also, $\nu = 2\hat{\psi}_1^2 / C(\hat{\psi}_1)$, where the denominator of this expression is the diagonal entry of the inverse information matrix corresponding to ψ_1.

So long as $\mathbf{C}(\hat{\boldsymbol{\psi}})$ is calculated using the expected or average information matrix, even further simplification of ν is possible. Letting $\mathbf{k}_E = \left[\text{tr}\left(\mathbf{R}\dot{\mathbf{W}}_2\right), ..., \text{tr}\left(\mathbf{R}\dot{\mathbf{W}}_r\right)\right]'$ and $\mathbf{L}_E$ be a matrix with $(s-1, t-1)^{th}$ entry $\text{tr}\left(\mathbf{R}\dot{\mathbf{W}}_s\mathbf{R}\dot{\mathbf{W}}_t\right)$ for $s, t = 2, ..., r$, the expected information matrix can be written

$$\mathcal{I}_E^{\boldsymbol{\psi}} = \begin{bmatrix} \frac{N-p}{2\hat{\psi}_1^2} & \mathbf{k}_E'/(2\hat{\psi}) \\ \mathbf{k}_E/(2\hat{\psi}) & \mathbf{L}_E/2 \end{bmatrix}.$$

the first diagonal entry of the inverse information matrix is therefore

$$\left[\mathcal{I}_E^{\boldsymbol{\psi}}\right]_{1,1}^{-1} = \frac{2\hat{\psi}_1^2}{N - p - \mathbf{k}_E'\mathbf{L}_E^{-1}\mathbf{k}_E},$$

by which it follows that $\nu = N - p - \mathbf{k}_E'\mathbf{L}_E^{-1}\mathbf{k}_E$. The same holds if the average information matrix is used, replacing $\mathbf{k}_E$ with $\mathbf{k}_A = [\mathbf{y}'\mathbf{R}\dot{\mathbf{W}}_1\mathbf{R}\mathbf{y}, ..., \mathbf{y}'\mathbf{R}\dot{\mathbf{W}}_{r-1}\mathbf{R}\mathbf{y}]'$ and $\mathbf{L}_E$ with the matrix $\mathbf{L}_A$ having $(s,t)^{th}$ entry $\mathbf{y}'\mathbf{R}\dot{\mathbf{W}}_s\mathbf{R}\dot{\mathbf{W}}_t\mathbf{R}\mathbf{y}$. From either expression, the degrees of freedom are equal to the total number of observations, minus the number of fixed effects, minus a penalty term depending on the design matrix $\mathbf{X}$, the correlation matrix $\mathbf{W}$, and the derivatives of $\mathbf{W}$ with respect to $\boldsymbol{\psi}_*$.

Note that the value of ν obtained from the re-parameterized model is the same as that from the original model. Recall that $\boldsymbol{\psi} = g(\boldsymbol{\theta})$, so $\partial g_1(\boldsymbol{\theta})/\partial\boldsymbol{\theta} = \mathbf{r}'$. From (3.36), it can be

seen that

$$\left(\mathcal{I}^{\psi}\right)^{-1} = \left(\nabla_{\boldsymbol{\theta}} g\right)\left(\mathcal{I}^{\boldsymbol{\theta}}\right)^{-1}\left(\nabla_{\boldsymbol{\theta}} g\right)', \tag{3.43}$$

by which it follows that $C(\hat{\psi}_1) = \left[\mathcal{I}^{\psi}\right]^{-1}_{1,1} = \mathbf{r}'\left[\mathcal{I}^{\boldsymbol{\theta}}\right]^{-1}\mathbf{r} = \mathbf{r}'\mathbf{C}(\hat{\boldsymbol{\theta}})\mathbf{r}$.

3.3.6. Estimating non-linear models

My development thus far has been for models that are linear in the parameters. Here I describe an expedient but imperfect approach for estimating Models TR3 and TR4. Both of these models involve a single non-linear parameter ω, conditional upon which the mean specification is linear in $\boldsymbol{\beta}$. Therefore, either one can be estimated by maximizing the profile likelihood in ω, then estimating the remaining parameters conditional on the profile-maximizing value of ω. Procedurally, this involves the following:

(1) For a fixed value of $\omega \in [0, 1)$, calculate the covariate $X^{\omega}_{ij} = (1-\omega)\sum_{k=1}^{j}\omega^{(j-k)}T_{ik}$ for $j = 1, ..., n_i$ and $i = 1, ..., m$. The case-level model given in (3.18) can then be written as $Y_{ij}\left(\tilde{T}_{ij}\right) = \beta_{0i} + \beta_{1i}X^{\omega}_{ij} + \epsilon_{ij}$, which is linear in β_{0i}, β_{1i}.

(2) Using the group-level assumptions of TR3 or TR4, estimate the model with covariate X^{ω} via RML, producing parameter estimates $\hat{\boldsymbol{\theta}}(\omega)$. Note the maximized profile log-likelihood $l_p(\omega)$.

(3) Repeat these steps for varying values of ω. Let $\hat{\omega}$ be the value that maximizes the profile likelihood $l_p(\omega)$.

(4) Set $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\hat{\omega})$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}, \hat{\omega})$. Use these estimates and the approximate covariance matrices $\mathbf{C}(\hat{\boldsymbol{\beta}})$ and $\mathbf{C}(\hat{\boldsymbol{\theta}})$, which are functions of $\hat{\boldsymbol{\theta}}$ and $X^{\hat{\omega}}$, to calculate the adjusted RML effect size estimate g_{AB} and its variance.

(5) It is also advisable to conduct a sensitivity analysis that assesses the impact of varying values of ω on the estimated effect size.

In this profiling approach, choosing a value of ω is directly analogous to choosing a model from a set of possible specifications that could involve polynomial terms for certain covariates. Just as the standard errors of parameter estimates from hierarchical linear models are conditional upon the model specification, so too the profiling approach ignores the uncertainty in the estimate of ω for purposes of assessing the uncertainty of the remaining parameter estimates. The effect size estimate depends on the estimate of ω directly through the term $(1 - \omega^{B-A})$ in (3.21), or indirectly through the treatment effect estimate $\hat{\gamma}_{10}$ if $B - A$ is large. Thus, ignoring the uncertainty of $\hat{\omega}$ will be more reasonable when $\hat{\omega}$ is close to zero, so that it has little effect on the uncertainty of the effect size numerator; it is certainly less defensible when ω is large. In any case, the profiling approach is intended only as a stop-gap until a more thorough analysis of estimation procedures for non-linear models can be carried out. I present it because the benefits of examining a wider class of models for treatment reversal designs would seem to outweigh the drawbacks of flawed estimation and inference techniques.

3.4. Small-sample performance

The estimation methods that I have proposed involve approximating the distribution of the RML effect size $\hat{\delta}_{AB}$ by a non-central t distribution. The small-sample performance of the adjusted estimator g_{AB} depends on the quality of this approximation, which may in turn depend on the particular data-generating model and the design of the study. In

this section, I report results of several small simulation studies examining the operating characteristics of the adjusted estimator under varying designs and data-generating models.

RML estimation methods are applicable in principle to a vast range of models and designs, but the scope of the simulations here is necessarily much more limited. I address two main questions. First, HPS proposed effect size estimators for Models MB1 and TR1 and found that they are nearly unbiased even in designs with small numbers of cases and relatively few measurement occasions; I therefore evaluate the bias and precision of the adjusted RML estimator under these same models, against the benchmark of the HPS estimator. Second, this chapter has introduced several models for multiple baseline and treatment reversal designs to which the HPS estimator is not immediately applicable. I examine the performance of the adjusted RML estimator in two such models that have additional random effects: TR2, which has varying treatment effects, and MB4, which has baseline trends that vary across cases. For each of these models, I focus on the bias of the effect size estimator and its associated variance estimator.

In all of the simulations described below, I used the `lme` function from the `R` package `nlme` (Pinheiro et al., 2012) to obtain RML estimates. This function uses a log-Cholesky parameterization for the random effects covariance matrix, which has an unrestricted parameter space (Pinheiro & Bates, 1996). I allowed the maximization routine in `lme` to run for at most 50 iterations, and accepted the resulting values even if they had not converged to a maximum. I provide further details regarding convergence when describing the simulation results.

3.4.1. Comparison of adjusted RML estimators and HPS estimator

The first simulation study compared the performance of the adjusted RML estimator to that of the estimator proposed by HPS, using bias and root mean-squared error as the criteria. For simulating Model MB1, I used a multiple baseline model in which treatment assignment times are spread as evenly as possible across the range of measurement occasions while maintaining at least 3 measurement occasions within each phase. For simulating Model TR1, I used a simple AB design with baseline and treatment phases of equal length, following Hedges et al. (2012b).

Table 3.1 reports the design of the simulation study used in conjunction with Models MB1 and TR1. The design consists of a $9 \times 5 \times 2 \times 2 \times 2$ factorial and follows the broad outlines of the simulations reported by HPS, but with fewer levels for certain parameters. The auto-correlation ϕ was varied between -0.3 and 0.5 (HPS used -0.9 to 0.9), as this range seems most plausible in applications to single-case research. The total variance was fixed to $\tau_0^2 + \sigma^2 = 1$ while the within-case reliability $\rho = \tau_0^2/(\tau_0^2 + \sigma^2)$ was varied between 0.0 and 0.8. The number of cases and number of measurement occasions was limited to the two smallest levels considered by HPS, because the bias of their estimator was negligible for larger sample sizes. I set the fixed effects equal to $\gamma_{00} = 0, \gamma_{10} = 1$, so that the effect size parameter δ_{AB} is equal to one. HPS observed that the bias of their estimator was proportional to the effect size parameter. On the assumption that the adjusted RML estimators behave similarly, I interpret the simulated biases as proportions of the effect size parameter (e.g., bias of less than 2%).

For each combination of parameter levels, I generated 25,000 datasets. I then calculated three effect size estimates based on each simulated dataset: the estimator described

Table 3.1. Simulation design for Models MB1 and TR1

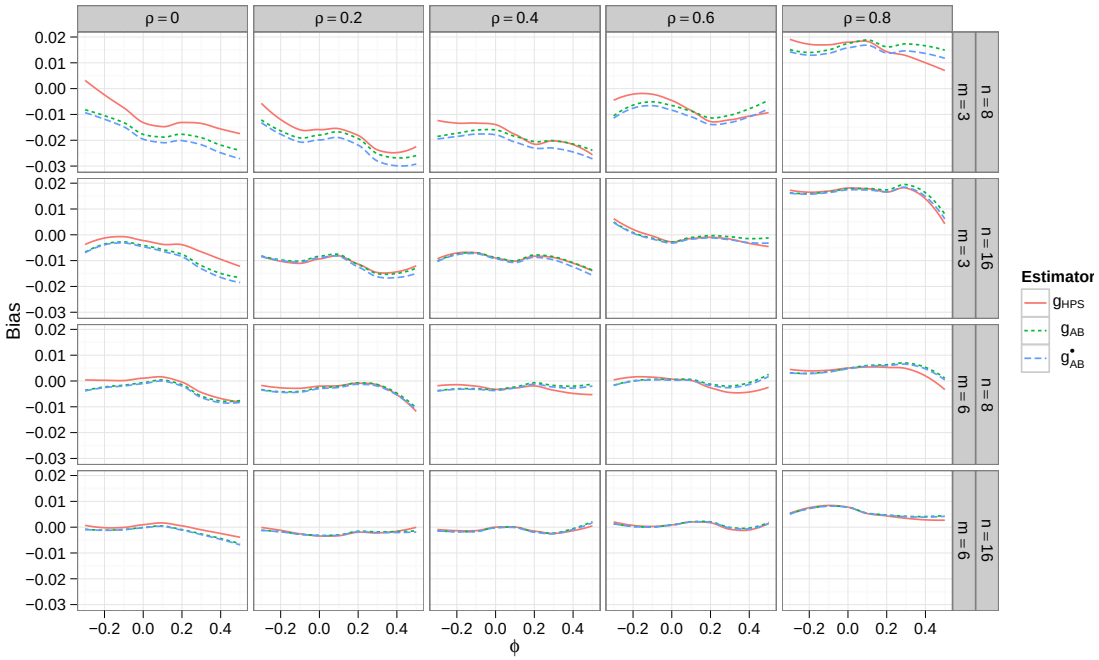
Parameter	Definition	Levels	Min.	Step	Max.
ϕ	Autocorrelation	9	-0.3	0.1	0.5
ρ	Within-case reliability	5	0.0	0.2	0.8
m	Number of cases	2	3	3	6
n	Measurement occasions	2	8	8	16
Design	Multiple baseline / AB	2	-	-	-

in Hedges et al. (2012a), which I refer to as g_{HPS}, the adjusted RML estimator g_{AB}, as given in (3.40), and a twice-adjusted estimator g^*_{AB} that involves an additional term for the bias of $\mathbf{r}'\boldsymbol{\theta}$, described in Appendix A. For the latter two estimators, I used the expected information matrix to evaluate ν and the bias-correction constant ξ.[4] Since each of the three estimators was calculated based on the same simulated data, their sampling distributions are inter-correlated and so differences between the estimators have very low Monte Carlo error.

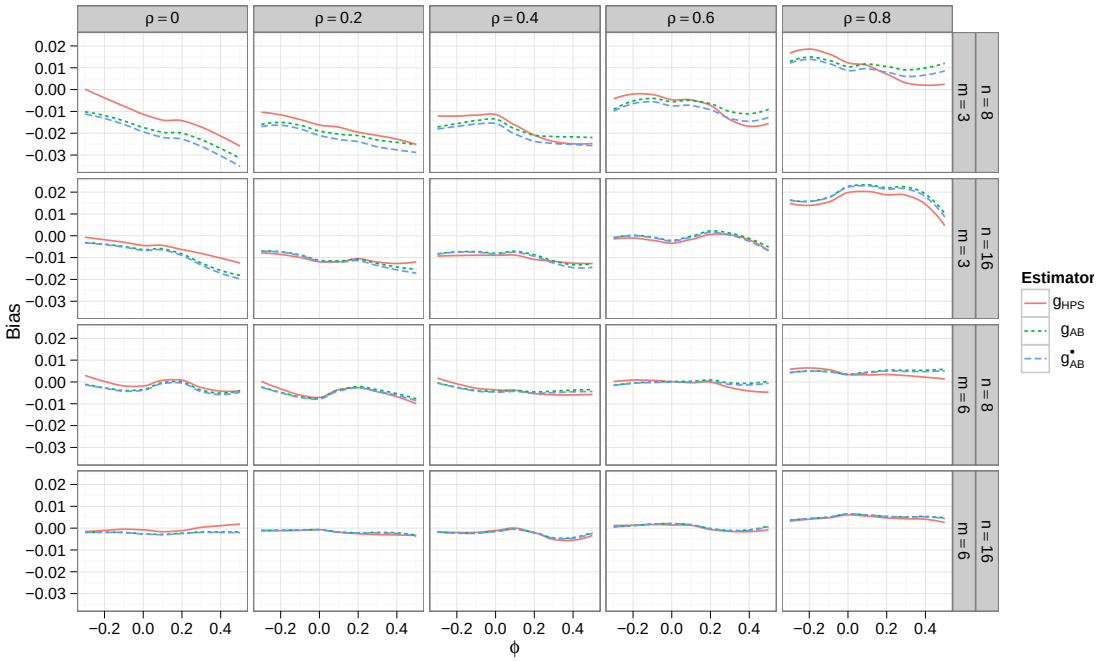
Figures 3.2a and 3.2b plot the bias of the three effect size estimators for the AB design and the multiple baseline design, respectively, across varying levels of the parameters.[5] Overall, the biases of the adjusted RML estimators are quite small and comparable to the bias of the HPS estimator. With the smallest sample size considerd ($m = 3, n = 8$), g_{AB} has a slightly larger bias than g_{HPS} when ρ is low, but a comparable bias when ρ is larger; still, the bias of g_{AB} is never greater than 3% in absolute magnitude. Also with the smallest sample size considered, the expected value of the twice-adjusted estimator g^*_{AB}

[4]The automatic output of the `lme` function does not use the inverse expected information matrix for estimating the variance of the variance components, but rather uses a numerical approximation to the Hessian of the log-likelihood. I wrote a separate function to calculate the expected information matrix from supplied parameter estimates.

[5]In each figure, the rows of the lattice correspond to different values of the sample size m and n, while the columns of the lattice correspond to varying values of ρ; the x-axis of each panel corresponds to ϕ, and different colors and line-types correspond to each estimator.



(a) AB design



(b) Multiple Baseline design

Figure 3.2. Bias of effect size estimators for (a) Model MB1 with a multiple baseline design and (b) Model TR1 with an AB design. Point-wise Monte Carlo standard error < 0.005.

Table 3.2. Average root mean-squared error of effect size estimators under Models MB1 and TR1

Design	m	n	g_{HPS}	g_{AB}	g^*_{AB}
AB	3	8	0.2607	0.2468	0.2457
		16	0.1742	0.1660	0.1657
	6	8	0.1051	0.1017	0.1016
		16	0.0693	0.0668	0.0668
Multiple baseline	3	8	0.2652	0.2429	0.2417
		16	0.1958	0.1742	0.1739
	6	8	0.1045	0.0998	0.0997
		16	0.0771	0.0700	0.0700

is always slightly less than that of g_{AB}; in general though, the value of the bias-correction constant ξ is so small that the two estimators are practically indistinguishable. For larger sample sizes ($m = 6$), all three estimators have biases of less than 1.2%.

Given that all three estimators have comparable biases, it is reasonable to also compare their precision. Table 3.2 reports the average root mean squared error of each estimator, where the average is taken over the levels of the nuisance parameters ϕ and ρ. On average and across designs and sample sizes, g_{AB} has slightly better precision than the HPS estimator and performs practically as well g^*_{AB}. Thus, in the simple models under consideration, g_{AB} provides a viable alternative to the HPS estimator. Further, the more computationally intensive estimator g^*_{AB} does not have any advantage over the simpler estimator g_{AB}, at least for these designs and sample sizes.

In addition to the effect size estimate itself, an estimate of its sampling variance is also needed for meta-analysis. I assessed the performance of proposed variance estimators using the relative variance; for an effect size estimator g with associated variance estimator V_g, the relative variance is defined as the ratio of the expected value of the variance

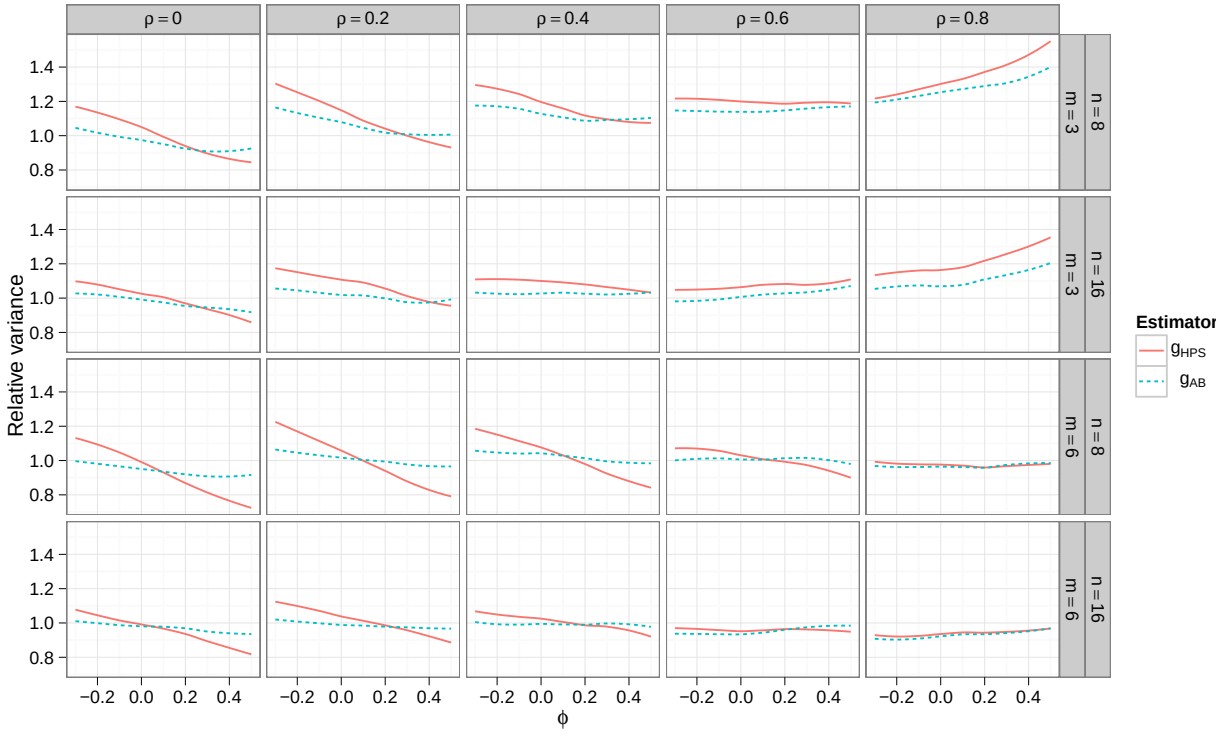


Figure 3.3. Relative variances of effect size estimators for Model MB1 with a multiple baseline design. Point-wise Monte Carlo standard error < 0.015.

estimator $\mathrm{E}(V_g)$ to the true variance of the effect size estimator $\mathrm{Var}(g)$. Relative variances close to one mean that the variance estimator is unbiased.

Figure 3.3 plots the relative variance of g_{AB} and g_{HPS} in the multiple baseline design, and is constructed in the same fashion as Figure 3.2.[6] Results for the AB design are very similar. From the figure, it can be seen that the HPS variance estimator is somewhat inaccurate for smaller values of the within-case reliability ρ, tending to under-estimate the true variance for positive values of the autocorrelation ϕ; this is true even at the larger sample size considered. In contrast, the variance estimator for g_{AB} provides more accurate estimates, with bias that depends less strongly on ρ and ϕ. Together with the

[6]The estimator g^*_{AB} is omitted because its relative variance is indistinguishable from that of g_{AB}.

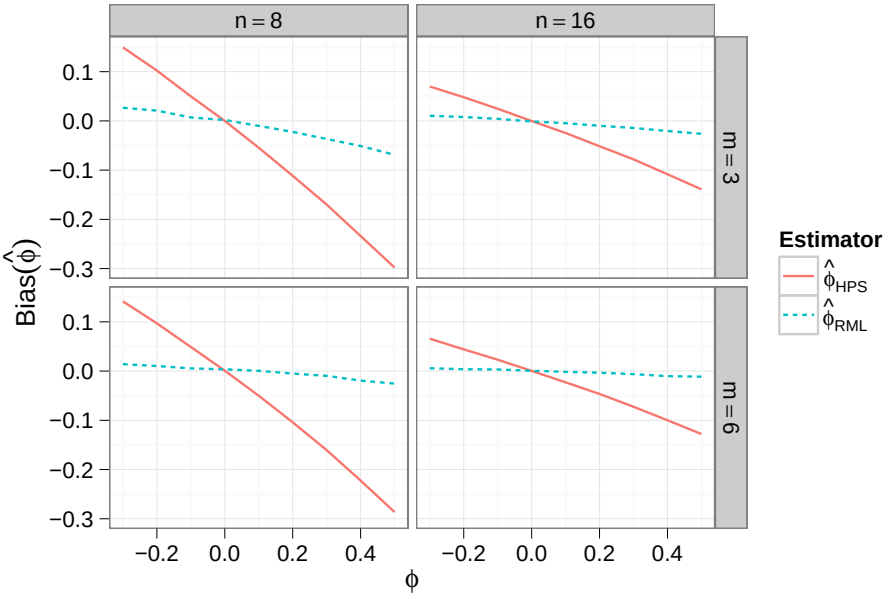


Figure 3.4. Bias of auto-correlation estimators for Model MB1 with a multiple baseline design and $\rho = 0.6$. Point-wise Monte Carlo standard error is less than 0.002.

small biases displayed by g_{AB}, these results suggest that the adjusted RML estimator is a reasonable alternative to the methods described by HPS for estimating effect sizes based on Models MB1 and TR1.

Finally, some incidental results from this first simulation study shed further light the performance of the HPS variance estimation methods. Hedges et al. (2012b) noted that the poor performance of their proposed variance estimator may be due to the fact that it depends strongly on the values ϕ and ρ, which must typically be estimated from the data. For estimating these nuisance parameters, HPS used moment estimators that have poor sampling properties when phase lengths are short, regardless of the number of cases

used.[7] In comparison, the RML estimators of the same nuisance parameters are less biased and more precise. To illustrate this, Figure 3.4 plots the biases of the RML estimator and the estimator used by HPS as a function of the true parameter ϕ, specializing the results to $\rho = 0.6$ and a multiple baseline design. Both estimators are approximately unbiased when $\phi = 0$, but the moment estimator of ϕ has a large, negative bias that is approximately proportional to the true parameter value. Though the bias of the this estimator is mitigated by increasing the number of measurement occasions, it remains constant as the number of cases increases from $m = 3$ to $m = 6$. In light of this, I would speculate that the performance of the variance estimator proposed by HPS might be improved by using different estimators of the nuisance parameters, such as the RML estimators.[8]

3.4.2. Performance of the adjusted RML estimator for Model TR2

The second simulation study examined the operating characteristics of the adjusted RML estimator under Model TR2. Compared to TR1, Model TR2 has one further random effect, and thus two additional variance components: the variance of treatment effects τ_1^2 and the covariance of the treatment effects and baseline levels τ_{10}. Due to the number of between-case variance components, the incidence of non-maximal estimates generated by the RML fitting algorithm is a concern in this model. The simulation design therefore

[7]HPS proposed a Yule-Walker estimator of the auto-correlation, pooled across cases and corrected for bias when $\phi = 0$. For fixed series length n and non-null ϕ, this estimator is inconsistent even as the number of cases m increases.

[8]However, the method used to estimate nuisance parameters also has implications for the small-sample bias of the HPS effect size estimator. HPS found that their proposed nuisance parameter estimators produced estimated degrees of freedom that actually led to smaller bias in g_{HPS} than when the degrees of freedom were calculated based on known values of the nuisance parameters. Such off-setting biases would not necessarily occur if RML estimators of the nuisance parameters were used.

Table 3.3. Simulation design for Model TR2

Parameter	Definition	Levels	Min.	Step	Max.
ϕ	Autocorrelation	5	-0.3	0.2	0.5
ρ	Within-case reliability	5	0.0	0.2	0.8
λ_1	Ratio of variance components τ_1^2/τ_0^2	2	0.1	0.4	0.5
m	Number of cases	4	3	1	6
n	Measurement occasions	2	8	8	16

varied the number of cases m between 3 and 6, as before, but also included the intermediate values of $m = 4$ and $m = 5$ in order to better characterize the relationship between non-maximal estimates and sample size. Again following Hedges et al. (2012b), I used a simple AB design even though longer treatment reversal designs are more common in practice.[9] The total number of measurement occasions was either $n = 8$ or $n = 16$, with baseline and treatment phases of equal length. Table 3.4 summarizes the design of the second simulation study, a $5 \times 5 \times 2 \times 4 \times 2$ factorial.[10]

In the previous simulation study, it was possible to explore the parameter space of Models MB1 and TR1 fairly thoroughly, but the greater number of parameters prohibits as comprehensive a simulation for Model TR2. I limited the simulation design in several ways. First, I parameterized the between-case variance in treatment effects as a proportion of the between-case variation in baseline levels; letting $\lambda_1 = \tau_1^2/\tau_0^2$, I set $\lambda_1 = 0.1$ or $\lambda_1 = 0.5$ to represent moderate and high levels of treatment effect heterogeneity, respectively. Next, I set $\tau_{10} = 0$ because pre-testing indicated that the correlation between

[9]For Model TR2, longer treatment reversal designs such as ABABs are simply further replications of the basic AB pattern, and so do not present any analytic complications. I therefore expect that the performance of the adjusted RML estimators would be similar in in these longer designs, or perhaps slightly better due to the longer total series lengths.

[10]I used just five levels of auto-correlation ϕ (rather than 9 levels as in the first simulation) in order to moderate the total number of factor combinations.

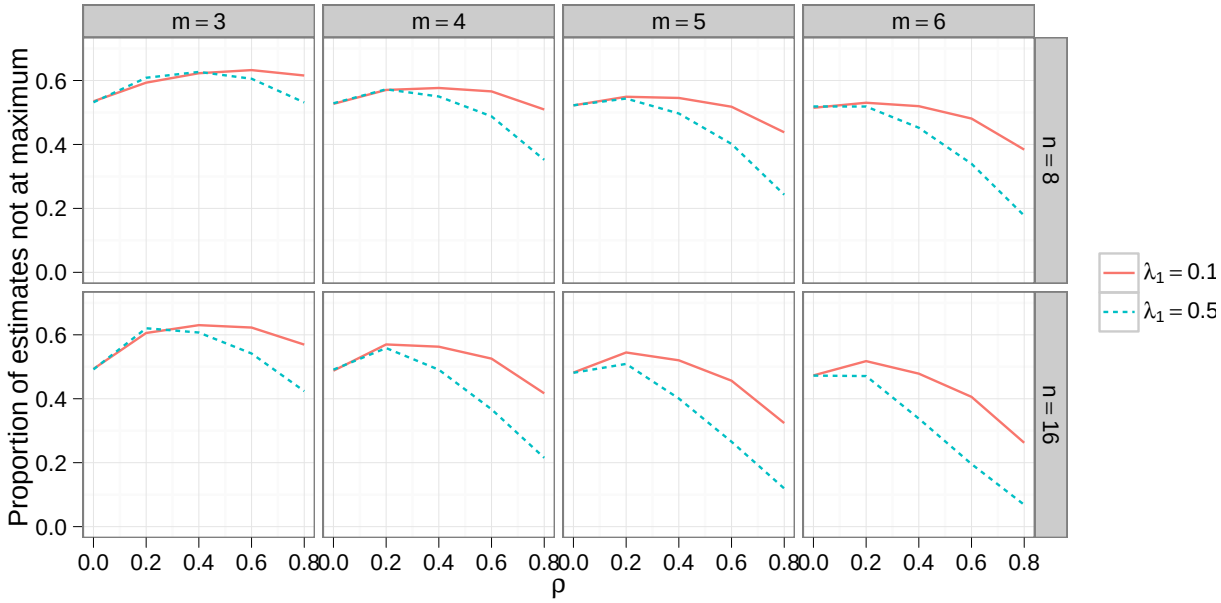


Figure 3.5. Proportion of RML parameter estimates not at maximum for Model TR2, versus ρ, averaging across levels of ϕ. Each line represents a different ratio of variance components λ_2. Point-wise Monte Carlo standard error < 0.002.

random effects had little influence on the bias of the effect size estimates. Finally, I set $\gamma_{00} = 0$, $\gamma_{10} = 1$, and $\tau_0^2 + \sigma^2 = 1$ so that the true effect size parameter $\delta_{AB} = 1$. For each combination of parameter levels, I generated 20,000 datasets and calculated the adjusted RML estimator and the associated variance estimator for each dataset.[11] As in the previous study, I used the expected information matrix to evaluate ν. I present results on the incidence of non-maximal estimates, the bias of the RML effect size estimator, and the relative variance of the RML estimator.

To begin, Figure 3.5 plots the proportion of iterations in which the fitting algorithm did not converge, resulting in estimates that were near to but not strictly at the maximum of the restricted likelihood. The incidence of non-maximal estimates averaged over 50%

[11] I also calculated g_{AB}^*. I omit these results as they are nearly identical to those for g_{AB}.

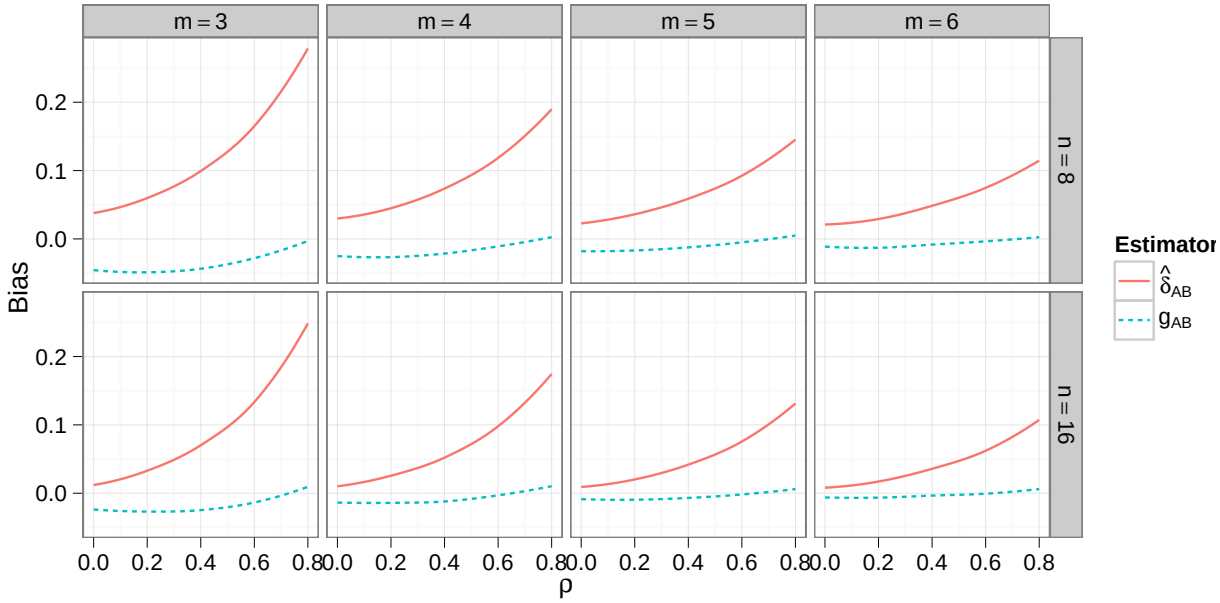


Figure 3.6. Bias of effect size estimators for Model TR2, versus ρ, averaging across levels of ϕ and λ_1. Point-wise Monte Carlo standard error < 0.007.

in designs with only a small number of cases, and remains sizable even when the $m = 6$. The incidence was somewhat lower for designs with more measurement occasions and for larger between-case variance components (i.e., larger ρ or larger λ_1). Taken together, these results suggest that in designs with only a few cases, RML will often produce boundary estimates for a model with two random effects.

Next, Figure 3.6 plots the bias of the adjusted effect size estimator g_{AB} for varying numbers of cases m, measurement occasions n, and within-case reliability ρ. Lacking other effect size estimators comparable to the HPS estimator for Model TR1, I used the unadjusted RML estimator $\hat{\delta}_{AB}$ as a point of comparison; Figure 3.6 plots its bias as well. The bias of g_{AB} is surprisingly small, even at the smallest sample sizes considered. When $m = 3$ and $n = 8$, the absolute bias is less than 7% across all combinations of parameters considered; for $m \geq 5$, the bias is always less than 3%. The bias of g_{AB} is also

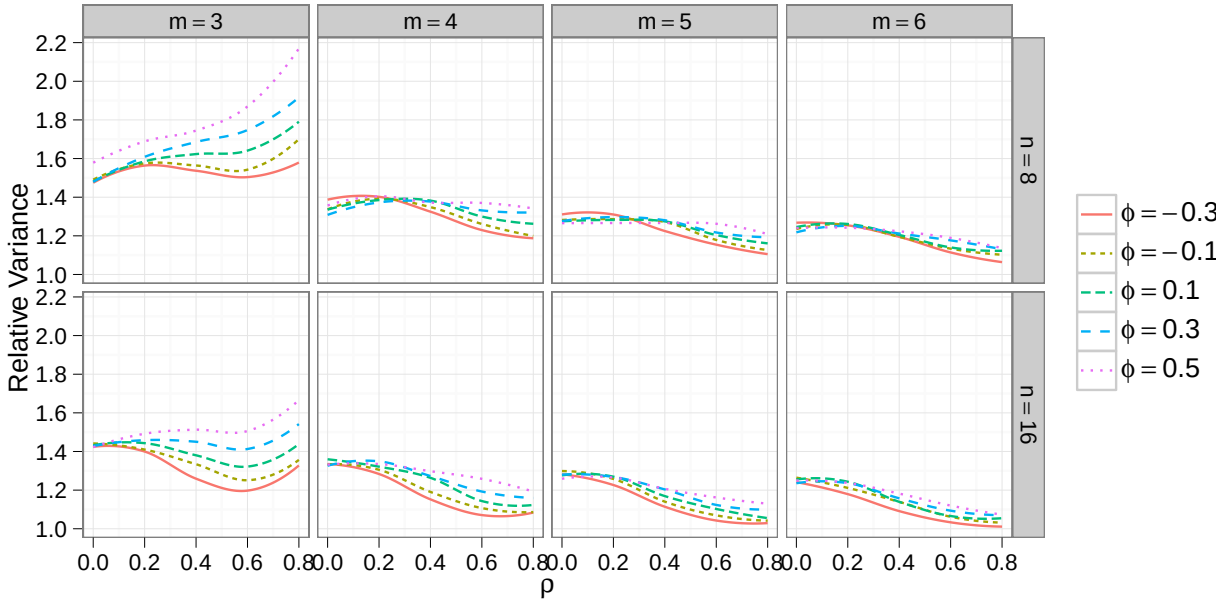


Figure 3.7. Relative variance of g_{AB} for Model TR2, versus ρ, averaging across levels of λ_1. Each line represents a different level of autocorrelation ϕ. Point-wise Monte Carlo standard error < 0.02.

both smaller and less variable than that of $\hat{\delta}_{AB}$, which increases substantially with ρ. In general, the adjusted RML estimator appears to have biases small enough to warrant use in meta-analysis.

Finally, Figure 3.7 displays the relative variance of the adjusted RML estimator versus the within-case reliability ρ, for varying numbers of cases and measurement occasions; separate lines are plotted for different levels of ϕ to illustrate how the relative variance depends on the autocorrelation. Across levels of ϕ, ρ, m, and n, the variance approximation given in (3.41) over-estimates the actual variance of g_{AB}, to a substantial extent when $m = 3$ and to a more moderate extent when m is larger. The bias in this estimator may come from several sources, including from approximating the distribution of $\hat{\delta}_{AB}$ by a t-distribution, from approximating the variance of the RML estimates using the inverse

expected information matrix, and from using imprecise parameter estimates in calculating the information matrix. If the variance approximation were used to apportion weights in the context of a fixed- or random-effects meta-analysis, the weight assigned to the effect size estimate g_{AB} would tend to be understated, relative to the theoretically most efficient weight. This would effectively down-weight evidence from single-case designs, particularly those with very small numbers of cases, in a meta-analysis that included evidence from multiple types of designs. Though more accurate variance estimates would certainly be preferable, use of the proposed variance approximation seems conservative and therefore warranted for meta-analysis.

3.4.3. Performance of the adjusted RML estimator for Model MB4

The third simulation study examined the operating characteristics of the adjusted RML estimator under Model MB4, which allows for baseline trends that vary across cases as well as a change in trend due to treatment that is constant across cases. Compared to MB1, MB4 has two additional parameters in the mean specification (the baseline trend γ_{20} and the trend-by-treatment interaction γ_{30}) and two additional variance parameters (the variance of the baseline slopes τ_2^2 and the covariance of the baseline slopes and levels τ_{20}). Due to the number of between-case variance components, the incidence of non-maximal estimates generated by the RML fitting algorithm is a concern in this model, as it was with Model TR2.

Recall that the effect size parameter in Model MB4 depends on the analyst's choice of times A and B, which characterize a hypothetical between-case randomized experiment by the point of treatment introduction and the point of outcome measurement. For purposes

Table 3.4. Simulation design for Model MB4

Parameter	Definition	Levels	Min.	Step	Max.
ϕ	Autocorrelation	5	-0.3	0.2	0.5
ρ	Within-case reliability	3	0.0	0.2	0.8
λ_2	Ratio of variance components τ_2^2/τ_0^2	2	0.1	0.4	0.5
m	Number of cases	4	3	1	6
n	Measurement occasions	2	8	8	16

of simulation, I took $A = n/2$, and $B = 3n/4$, so that the effect size represents the change due to treatment after $n/4$ measurement occasions, probably a reasonably short time relative to the length of the study. Because Model MB4 includes linear baseline time trends that vary across cases, the choice of centering point for the time trend affects the interpretation of the variance components. I centered time at point $B = 3n/4$, so that τ_0^2 represents the between-case variation in the level of the outcomes at time B and τ_{20} is the covariance between cases' baseline slopes and outcome levels at time B. With $\boldsymbol{\gamma} = (\gamma_{00}, \gamma_{10}, \gamma_{20}, \gamma_{30})'$ and $\boldsymbol{\theta} = (\sigma^2, \phi, \tau_0^2, \tau_{20}, \tau_2^2)'$, the target effect size parameter is then defined by (3.26) with $\mathbf{p} = (0, 1, 0, n/4)'$ and $\mathbf{r} = (1, 0, 1, 0, 0)'$.

Table 3.4 summarizes the design of the third simulation study, which parallels that for Model TR2. To moderate its dimensionality, I again limited the parameter space in several ways. First, I parameterized the between-case variance in baseline slopes as a proportion of the between-case variation in baseline levels; letting $\lambda_2 = \tau_2^2/\tau_0^2$, I set $\lambda_2 = 0.1$ or $\lambda_2 = 0.5$ and used $\tau_{20} = 0$ throughout. Next, I did not vary the fixed effects, instead setting the average baseline outcome level $\gamma_{00} = 0$, the (fixed) change in the level due to treatment $\gamma_{10} = 1$, the average baseline slope $\gamma_{20} = 0$, and the (fixed) increase in slope due to treatment $\gamma_{30} = 0$. Finally, I set $\tau_0^2 + \sigma^2 = 1$ so that the true effect size parameter is $\delta_{AB} = 1$. For each combination of parameter levels, I generated

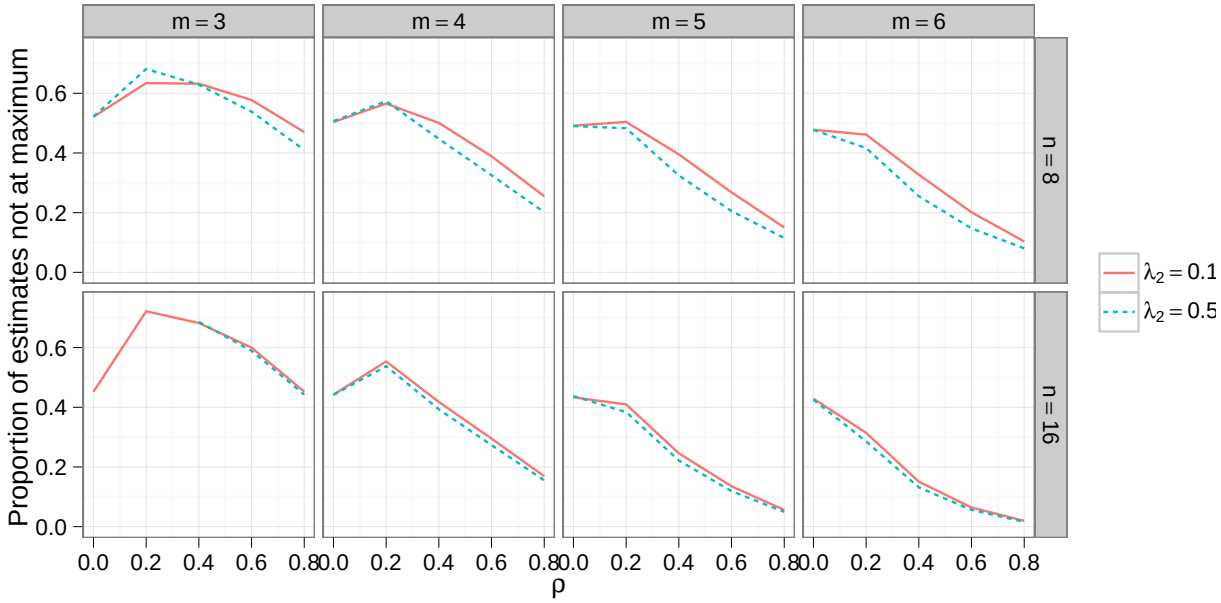


Figure 3.8. Proportion of RML parameter estimates not at maximum for Model MB4, versus ρ, averaging across levels of ϕ. Each line represents a different ratio of variance components λ_2. Point-wise Monte Carlo standard error < 0.002.

20,000 datasets and calculated the adjusted RML estimator for each dataset, using the expected information matrix to evaluate ν.[12] The presentation of results parallels that of the previous simulation study.

First, Figure 3.8 plots the incidence of estimates that do not maximize the restricted likelihood due to convergence issues. Non-maximal estimates awere very common, averaging over 50% when $m = 3$ and remaining quite high even as m increases. Increased between-case variability ρ reduced the incidence of non-maximal estimates, though the relative variability of the baseline trends, controlled by λ_2, had little effect. Compared to the results from Model TR2, the incidence in Model MB4 was slightly lower, on average. Still, non-maximal estimates occurred commonly enough to be of concern, and using a

[12]As in the previous simulations, I also calculated g^*_{AB} but found it to be nearly identical to g_{AB}.

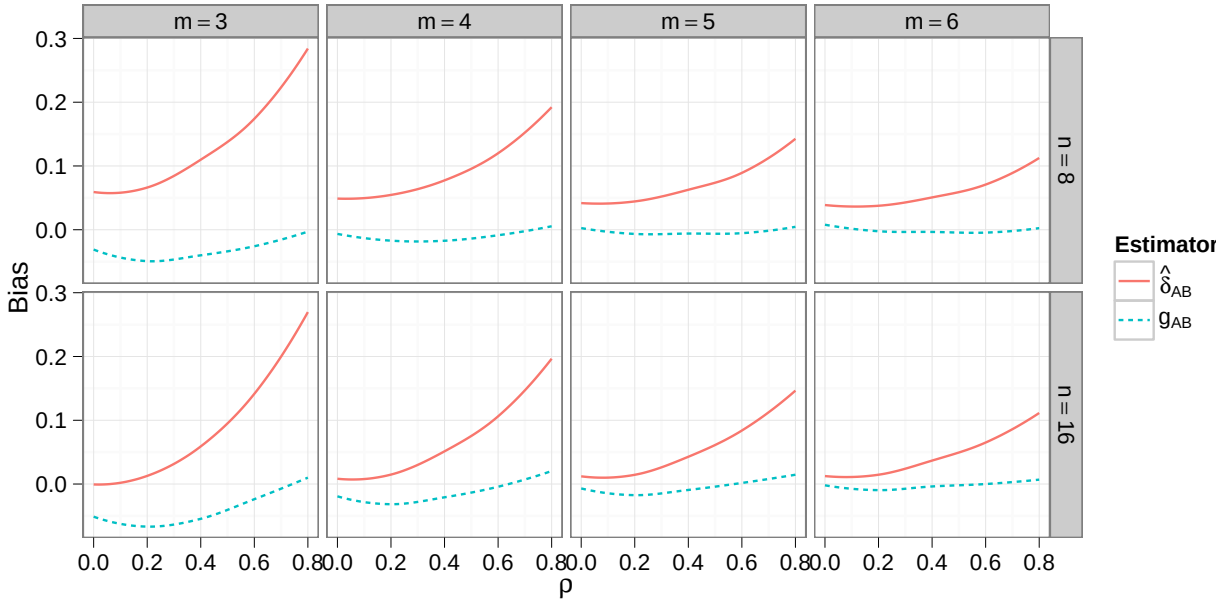


Figure 3.9. Bias of effect size estimators for Model MB4, versus ρ, averaging across levels of ϕ and λ_2. Point-wise Monte Carlo standard error < 0.008.

different method for handling boundary estimates could have some impact on the bias of the adjusted effect size estimator, which I now examine.

Figure 3.9 plots the bias under Model MB4 of the adjusted effect size estimator g_{AB} as well as the unadjusted RML estimator $\hat{\delta}_{AB}$. As with Model TR2, the bias of g_{AB} was small, even at the smallest sample sizes considered. When $m = 3$ and $n = 8$, the absolute bias was less than 7% across all combinations of parameters considered, though it was as large as 11% when $n = 16$. For $m \geq 5$, the bias was never more than 3%. Also as with Model TR2, g_{AB} was substantially less biased than $\hat{\delta}_{AB}$, which had a bias that increases with ρ. As previously, the bias was small enough that g_{AB} should be considered as suitable for use in meta-analysis.

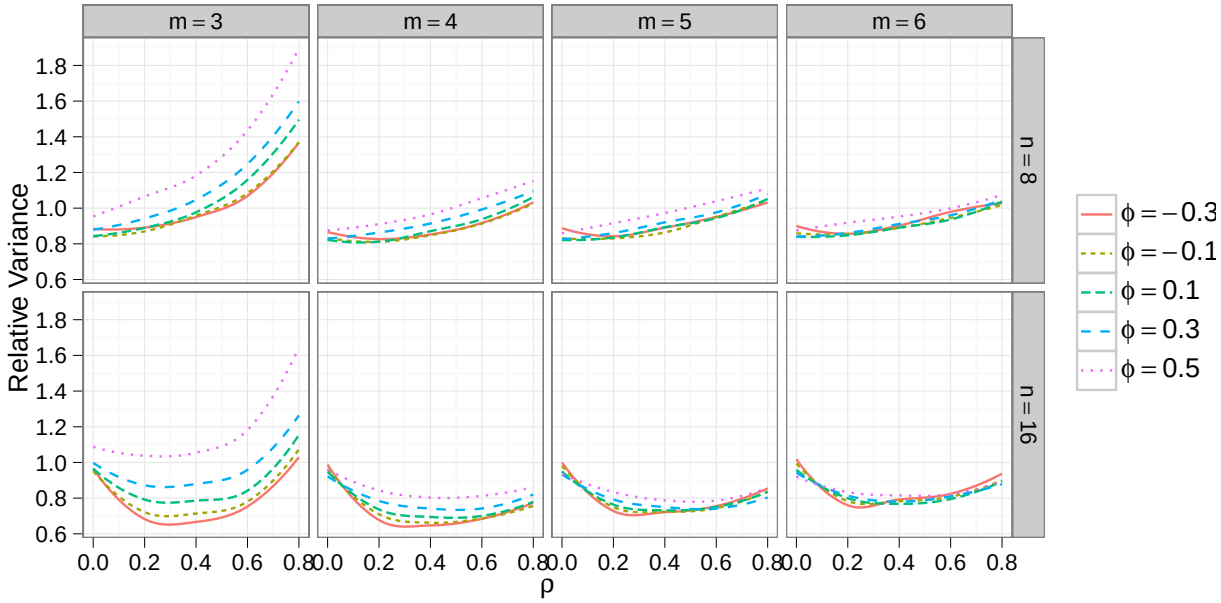


Figure 3.10. Relative variance of g_{AB} for Model TR2, versus ρ, averaging across levels of λ_1. Each line represents a different level of autocorrelation ϕ. Point-wise Monte Carlo standard error < 0.02.

Finally, Figure 3.10 displays the relative variance of the adjusted RML estimator, and is constructed just like Figure 3.7. Unlike in previous models, here the variance approximation tended to have a downward bias, except for when $m = 3$ and $n = 8$. The under-estimation was more pronounced for longer series length. The variance approximation in this model depends on a multiple $n/4$ of the treatment-by-trend interaction γ_{30}, and so may be particularly sensitive to under-statement of the variance of fixed effects. Other methods for estimating the fixed effects covariance matrix, such as those proposed by Kenward and Roger (1997, 2009), could be useful in this instance for obtaining improved estimates of the effect size variance. Better estimates are needed, as use of the current variance approximation for determining fixed- or random-effects meta-analytic weights

will tend to be anti-conservative, assigning too much weight to effect size estimates from Model MB4.

3.5. Discussion

This chapter has outlined a series of models for data from multiple baseline and treatment reversal designs, illustrated how design-comparable d-type effect sizes can be defined under those models, described one method for estimating such effect sizes, and evaluated the small-sample performance of the estimation method using simulations. In this section, I comment further on each of these topics.

3.5.1. Models

Sections 3.1 and 3.2 presented catalogs of hierarchical models for multiple baseline and treatment reversal designs, respectively. I have sought to highlight models that will be useful and interesting for single-case researchers, while also meeting criteria that allow a model to serve as a basis for defining design-comparable effect sizes. All five models for the multiple baseline design used a common case-level regression specification, differing only in whether the case-level regression parameters were allowed to vary across cases, were assumed to be common across cases, or were fixed to zero. Models for the treatment reversal design were similarly inter-related, differing mostly in whether the parameters representing treatment effects were allowed to vary across cases.

The variety of models presented begs the question of how to choose among them. In Chapter 4, I will demonstrate a tentative model selection strategy, based on a combination of formal testing and graphical depiction of model predictions. Beyond this though, I

would speculate that more richly parameterized models will be attractive in the fields that use single-case designs. Given the ideographic orientation of single-case research, it seems fitting to assume that cases vary in any and all respects unless evidence can be presented to the contrary. The alternative–for instance, assuming as a default that treatment effects are homogeneous across cases–would be an uncomfortable stance for researchers accustomed to considering and analyzing each case in its own light, separate and apart from results of other cases even in the same study.[13] Recent criticisms of multilevel modeling approaches for analysis of single-case data have highlighted the disconnect between the knowledge and experience of researchers in the field versus how the statistical models are expressed and applied (Parker & Vannest, 2012). These are very valid concerns, and further work on selection of appropriate models for analysis of single-case data will need to be tied closely and articulately to accumulated scientific knowledge in the fields of application.

3.5.2. Design-comparable effect sizes

The models considered in this chapter permit definition of design-comparable standardized mean differences, and in certain of the models, the operational definition of the target effect size depends on the time-points A and B that define it. For example, in Model MB4, the degree to which the effect size is sensitive to chosen values of A and B will depend on the degree of variability in baseline slopes, which affects the extent of change in total variance over time, and on the size of the treatment-by-time trend interaction, which

[13]Of course, the models proposed in this chapter do in fact entail using information from one case to inform the analysis of another, but in a more tempered fashion: randomly varying parameters are pooled "partially" rather than completely (Gelman & Hill, 2007). Hierarchical models are therefore a natural tool for bridging between ideographic and nomothetic perspectives.

captures the extent to which the average treatment effect changes over time. Though such arbitrary dependence may strike some as a drawback of the proposed effect size, I would argue that it is an inherent consequence of design-comparability. Rather than a problem with the effect size itself, unexpected sensitivity of the effect size may instead be symptom that one's broader modeling assumptions need to be re-evaluated relative to the context of the analysis.

That said, I also acknowledge that design-comparability is not the only or over-riding desideratum of effect sizes for use in analysis and meta-analysis of single case research. Other effect sizes, including ones that are not identified in between-subjects designs, may be as useful or more useful as summaries of treatment effects from single-case studies. If scientific theory or empirical evidence suggests that a different metric better quantifies the underlying phenomenon being studied, its lack of design-comparability should not prevent its consideration.

3.5.3. Estimation and small-sample performance

I have described one method for estimating design-comparable standardized mean differences, based on the results of a conventional and widely-used method for estimating the component parameters of hierarchical linear models. I have focused on restricted maximum likelihood estimation because it is extensible, in that it can be applied to a wide range of different models, as well as accessible, in that it can be implemented using standard software. The simulation studies presented in this chapter provided some initial evidence that the adjusted RML effect size estimator and its approximate variance estimator have reasonably small biases, even in samples with very few cases.

Findings based on the simulation studies are limited in several ways, the most obvious being that they have examined only a fairly small set of models and, for the models with more than a single variance component, only a subset of the full parameter space. Also, in all three simulations, I evaluated the covariance matrix of the variance component parameters using an explicit formula for the expected information matrix of the restricted likelihood. The exact bias of the adjusted RML effect size estimator depends on this covariance (through the degrees of freedom correction), and thus also on the method used to evaluate it. Intuitively, the difference between different information matrices (i.e., expected versus observed versus average) will probably be minor, though it is unclear whether use of numerically-evaluated approximations rather than explicit formulas may lead to greater differences. Software packages differ in how the covariance matrix is estimated, both in terms of the default used and what alternatives are available. When presenting the results of RML model estimation and effect size calculations, it would therefore prudent to also report the software version used to fit the model and the method used to estimate the covariance matrix of the variance component parameters.[14]

It is important to emphasize that I have investigated one very specific and circumscribed aspect of RML estimation in small samples. That RML estimates appear to be useful for forming effect size estimates should not lead one to conclude that RML estimation is the best all-around method for estimating hierarchical models on small samples of independent cases. On the contrary, the simulation studies demonstrated that with small samples, RML will often lead to boundary estimates of certain variance components even when the true parameters are very far from the boundary. Other estimation approaches

[14]However, the exact method used to estimate the covariance matrix can be surprisingly difficult to determine.

may have better properties for general inferential purposes. In particular, fully Bayesian specifications with weakly informative priors (e.g., Gelman, 2006) and related varieties of penalized likelihood methods (e.g., Chung, Rabe-Hesketh, Gelman, Liu, & Dorie, 2013) warrant further consideration, especially for studies with few cases relative to the number of random effects to be estimated.

Several other questions related to estimation methods also need to be investigated. First and foremost, a more principled approach to estimation of the non-linear models TR3 and TR4 is needed. Relevant theory exists (e.g., Davidian & Giltinan, 1995); it remains only to apply that theory in the context of the models that I have proposed. A less pressing but still interesting question is whether the approximate bias correction to the variance components described in Appendix A could be used to improve the estimation of the degrees of freedom ν, with consequences for the accuracy of the effect size variance given in (3.41). I found that the approximate bias of the total variance (that enters the denominator of the effect size) is so small as to be inconsequential. However, the approximate biases of the component parameters (e.g., σ^2 and τ_0^2 in Model MB1) tend to be larger. It could be that the the degrees of freedom estimate could be improved by evaluating the information matrix using bias-corrected estimates of the variance components, rather than using the RML estimates as I have done. Unfortunately I see no alternative route for investigating this hypothesis other than further simulation studies, which would be computationally quite intensive.

Finally, it is worth reflecting on the differences between the adjusted RML estimator and the effect size estimator proposed by HPS for Models MB1 and TR1. Based on simulations, I found that the mean-squared error of the adjusted RML estimator is slightly

lower than that of the HPS estimator, while both have only small, comparable biases. However, this small improvement in precision may come at the expense of robustness to modeling assumptions. I have evaluated these two estimators under a known data-generating model involving a parametric error distribution and a dependence structure (e.g., first-order auto-regression for the within-case errors); from a model-building and model-testing standpoint, these assumptions may be hard to evaluate based on the small samples typically used in single-case studies. The RML estimator may be more severely affected by violations of these assumptions because it uses a fully specified likelihood, whereas the HPS estimator maintains a certain "robustness" because it uses an exactly unbiased moment estimator for the variance in the denominator of the effect size. Unfortunately, as of this writing the HPS estimation method is only available for the most basic models discussed in this chapter. I sketch one possible extension of the HPS approach in Section 7.1.

CHAPTER 4

Design-comparable standardized mean differences: Applications

In this short chapter, I present several examples of the models and estimation methods proposed in Chapter 3. The first example illustrates the calculation of the design-comparable standardized mean difference estimate based on the results of an RML estimation routine. The remaining examples illustrate the process of model fitting and comparison. Several of the examples were also analyzed by Hedges et al. (2012a, 2012b, HPS hereafter), using Model MB1 or TR1; my analysis of these studies includes estimates of the same model by RML for comparison with their estimation methods. Unless otherwise stated, the UnGraph software (Biosoft, 2004; Shadish et al., 2009) was used to extract the data for each example from graphs presented in the original articles. I estimated all of the models below using the `nlme` package in `R` (Pinheiro et al., 2012).

4.1. Saddler, Behforooz, & Asaro (2008)

Saddler, Behforooz, and Asaro (2008) used a multiple baseline across individuals to evaluate the effect of a particular instructional technique on the quality of fourth grade students' writing. The design included $m = 6$ fourth grade students as cases and a total of 10 unique measurement occasions, though none of the cases had complete data. Instead, each case was measured 3 or 4 times in the baseline phase and 3 times in the treatment phase, producing a total of 41 observations (excluding data from a third maintenance phase). Writing quality was measured on a seven-point scale, which I treat as a continuous

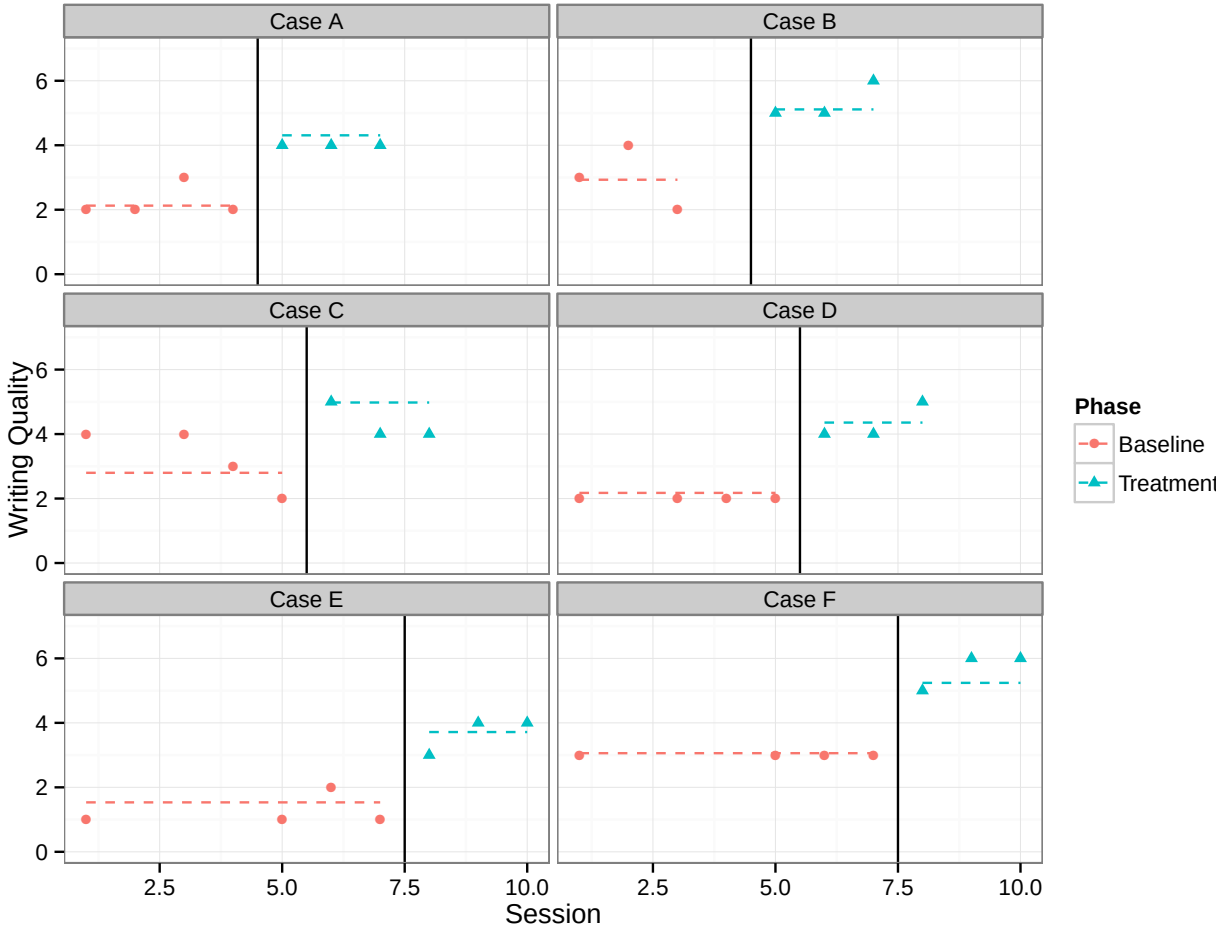


Figure 4.1. Data from Saddler et al. (2008): Writing quality over time for six fourth grade students. Solid vertical lines marks the point of treatment introduction for each student. Dashed lines represent the empirical Bayes estimates of each student's average level of writing quality in each phase, based on Model MB1.

measure. This last assumption is certainly tenuous, particularly because one students baseline scores were often at the lower extreme of the scale. Figure 4.1 plots the data from the study as well as the empirical Bayes estimates from Model MB1.[1]

My analysis assumes that there are no time trends and that the treatment effect consists of a shift in the mean outcome that is constant across cases, as in Model MB1.

[1]For details on empirical Bayes estimates, see Pinheiro and Bates (2000) or Raudenbush and Bryk (2002).

Table 4.1. Model MB1 estimates for Saddler et al. (2008) data

	HPS		RML	
Parameter	Estimate	(s.e.)	Estimate	(s.e.)
Variance components				
Autocorrelation ($\hat{\phi}$)	0.010		0.382	(0.235)
Within-case var. ($\hat{\sigma}^2$)	0.349		0.443	(0.166)
Between-case var. ($\hat{\tau}_0^2$)	0.603		0.438	(0.366)
Total var ($\hat{\tau}_0^2 + \hat{\sigma}^2$)	0.952		0.882	(0.366)
Fixed effects				
Intercept ($\hat{\gamma}_{00}$)			2.437	(0.319)
Treatment ($\hat{\gamma}_{10}$)	2.097	(0.196)	2.181	(0.234)
Effect size				
Unadjusted ($\hat{\delta}_{AB}$)	2.149		2.323	
Adjusted (g_{AB})	1.963	(0.578)	2.169	(0.565)
Degrees of freedom (ν)	8.918		11.603	
Constant (κ)	0.201		0.249	

Table 4.1 reports RML estimates of the fixed effects, variance components, and the effect size, along with associated standard errors. For comparison, Table 4.1 also reports corresponding HPS estimates from Hedges et al. (2012a).

RML estimates of the autocorrelation, within-case variance, and between-case variance are generated by the software. The corresponding standard errors are based on the inverse expected information matrix.[2] The estimate of the total variance is obtained directly by summing estimates of its components, while its standard error is obtained by summing the approximate covariance matrix of its components. Based on the variance component estimates, the fixed effects estimates and corresponding standard errors are generated by the software according to (3.29) and (3.30), respectively. The remaining estimates are calculated according to the formulas given in Section 3.3.4. From expression (3.37), I

[2]I wrote a function to calculate the expected information matrix from supplied parameter estimates, rather than relying on the numerical approximation generated by the software.

find the unadjusted effect size $\hat{\delta}_{AB} = \hat{\gamma}_{10}/\sqrt{\hat{\tau}_0^2 + \hat{\sigma}^2} = 2.181/\sqrt{0.882} = 2.323$ standard deviation (s.d.) units. Using expression (3.38), I calculate $\kappa = \sqrt{\text{Var}(\hat{\gamma}_{10})}/\sqrt{\hat{\tau}_0^2 + \hat{\sigma}^2} = 0.234/\sqrt{0.882} = 0.249$; from (3.39), I calculate

$$\nu = \frac{2\left(\hat{\tau}_0^2 + \hat{\sigma}^2\right)^2}{\text{Var}\left(\hat{\tau}_0^2 + \hat{\sigma}^2\right)} = \frac{2 \times 0.882^2}{0.346^2} = 11.603.$$

Based on the calculated values of ν and $\hat{\delta}_{AB}$, I find the adjusted effect size estimate to be $g_{AB} = J(11.603) \times 2.323 = 2.169$ s.d. Based on ν, κ, and g_{AB}, I use expression (3.41) to calculate the approximate variance of g_{AB}:

$$V\left(g_{AB}\right) = J(11.603)^2 \left[\frac{11.603 \times 0.249^2}{11.603 - 2} + 2.169^2 \left(\frac{11.603}{11.603 - 2} - \frac{1}{J(11.603)^2}\right)\right] = 0.319,$$

which corresponds to a standard error of 0.565 s.d. The adjusted effect size estimate describes an average treatment effect that is in the same metric as the standardized mean difference from a between-subjects study. Since it is design-comparable, g_{AB} can be synthesized along with effect sizes from other studies, including those from between-subjects designs and other multiple baseline designs.

The estimate is based on the same statistical model as used in Hedges et al. (2012a); only the estimation method differs. Overall, the RML estimate of the effect size is fairly similar to the HPS estimate: the point estimate is about 10% larger, while the standard error is nearly identical. The RML estimates of the nuisance parameters differ more substantially from the HPS estimates. The RML estimated autocorrelation is substantially higher than the HPS estimate, but the latter is based on a moment estimator that is known to be biased towards zero and that does not perform well when data are missing.

The ratio of the between-case variance to the total variance is somewhat lower when based on RML ($\rho = 0.497$) than when based on the HPS method ($\rho = 0.633$). In combination with the estimated autocorrelation, this leads to slightly larger degrees of freedom.

4.2. Laski, Charlop, & Schreibman (1988)

Laski, Charlop, and Schreibman (1988) used a multiple baseline across individuals to evaluate the effect of a training program for parents on the speech production of their autistic children. The design included $m = 8$ children between the ages of 8 and 10 years.[3] Cases were measured between 4 and 11 times in the baseline phase and between 7 and 11 times in the treatment phase, with a maximum of 20 consecutive measurement occasions in all. Speech production was measured using a partial interval recording technique, calculated as 100% times the number of 10-second intervals during which the child verbalized, divided by 60 intervals per session. Figure 4.2 displays a plot of the data from each case.

My initial analysis assumes that there are no time trends and that the treatment effect consists of a shift in the mean outcome that is constant across cases, as in Model MB1. Table 4.2 reports RML estimates of the fixed effects, variance components, and the effect size, along with associated standard errors. For comparison, Table 4.2 also reports corresponding HPS estimates from Hedges et al. (2012a).

RML estimates of the autocorrelation, within-case variance, and between-case variance are generated by the software, as described in the previous example. The remaining estimates are calculated according to the formulas given in Section 3.3.4, again as described

[3]Child 3 was measured separately with each parent, but for simplicity I include only the measurements taken with his mother.

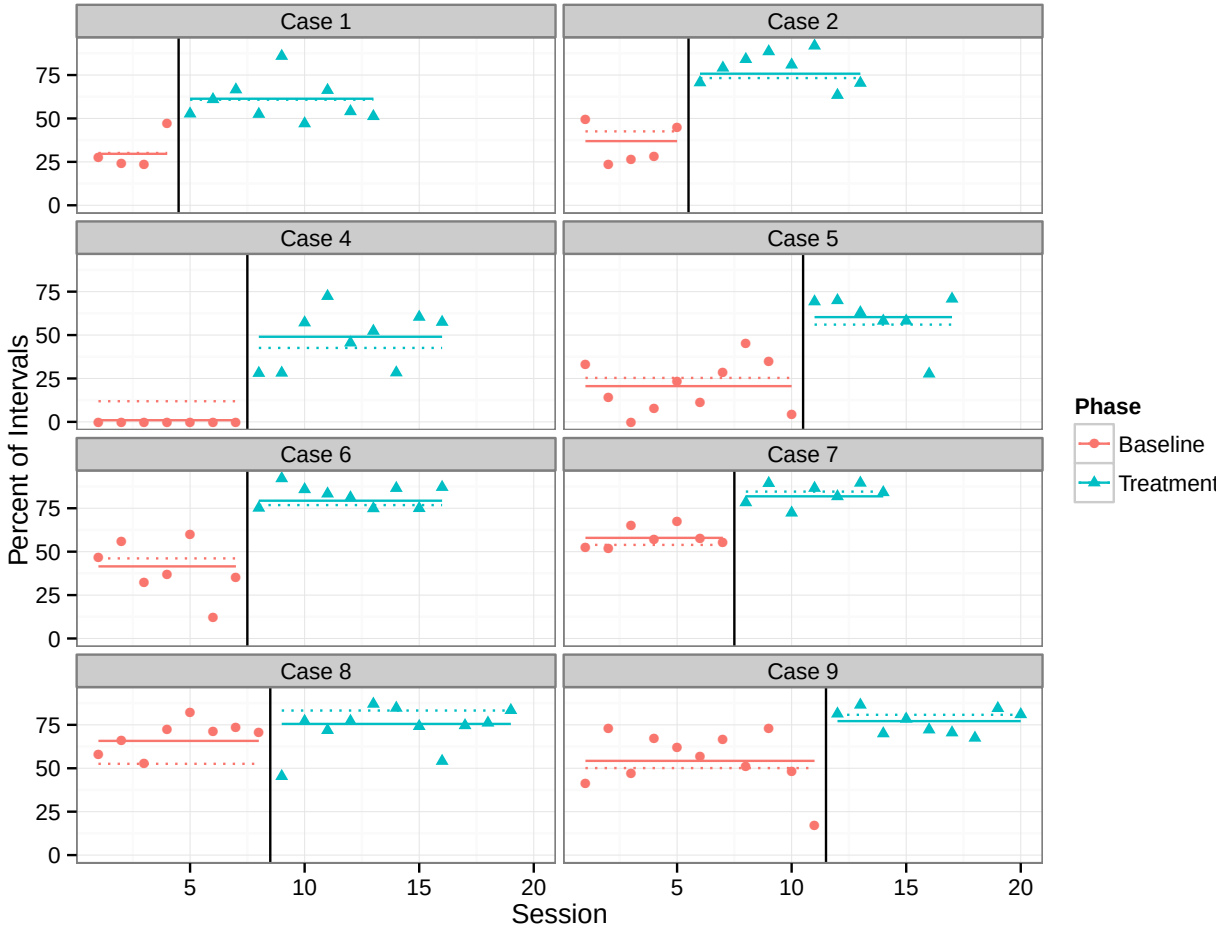


Figure 4.2. Data from Laski et al. (1988): Percentage of intervals with vocalization by session. Solid vertical lines mark the point of treatment introduction for each child. Horizontal lines represent the empirical Bayes estimates of each student's average percentage of intervals with vocalization in each phase, based on Model MB1 (dashed lines) or MB2 (solid lines).

in the previous example. Based on the RML estimate of Model MB1, the treatment increased the percentage of intervals during which children vocalized by an average of 30.7 percentage points, very close to the HPS estimate of 31.8 percentage points. The RML estimate of the total outcome variance (including both between- and within-case variation in the absence of treatment) is 439 squared percentage points, somewhat smaller than the

Table 4.2. Model estimates for Laski et al. (1988) data

	Model MB1				Model MB2	
	HPS		RML		RML	
Parameter	Estimate	(s.e.)	Estimate	(s.e.)	Estimate	(s.e.)
Variance components						
Autocorrelation ($\hat{\phi}$)	0.017		0.253	(0.100)	0.023	(0.107)
Within-case var. ($\hat{\sigma}^2$)	142.555		192.771	(28.266)	142.590	(19.571)
Between-case var. ($\hat{\tau}_0^2$)	323.261		245.950	(142.179)	475.057	(265.658)
Case-treatment cov. ($\hat{\tau}_{10}$)					-250.853	(160.840)
Treatment var. ($\hat{\tau}_1^2$)					175.326	(115.013)
Total var. ($\hat{\tau}_0^2 + \hat{\sigma}^2$)	465.816		438.721	(144.047)	617.647	(266.095)
Fixed effects						
Intercept ($\hat{\gamma}_{00}$)			39.076	(5.990)	38.452	(7.881)
Treatment ($\hat{\gamma}_{10}$)	31.822	(2.212)	30.684	(3.000)	31.592	(5.178)
Effect size						
Unadjusted ($\hat{\delta}_{AB}$)	1.474		1.465		1.271	
Adjusted (g_{AB})	1.388	(0.317)	1.405	(0.286)	1.181	(0.358)
degrees of freedom (ν)	13.100		18.552		10.776	
Constant (κ)	0.102		0.143		0.208	
Log-likelihood			-519.1		-511.0	
Akaike Info. Criterion			1048.3		1036.0	

HPS estimate of 466 squared percentage points. Scaling the treatment effect estimates by the corresponding total variance estimates produces unadjusted effect size estimates that differ by less than 1%. The difference between the adjusted effect size estimates is only slightly larger, due to the larger degrees of freedom in RML. As in the previous example, the RML estimate of the autocorrelation is significantly larger than the HPS estimate, while the ratio of between-case variance to total variance is smaller.

The extensibility of RML estimation makes it possible to examine other models for these data as well. The final column of Table 4.2 reports the results of fitting Model MB2, which allows treatment effects to vary across cases. The RML estimate of the variance in the treatment effect is 175 squared percentage points, suggesting that the effect of the

training program is not homogeneous across cases. More formally, a likelihood ratio test can be used to compare the fit of MB1 against MB2 (see for instance Pinheiro & Bates, 2000). I applied such a test for the hypothesis that $\tau_1^2 = \tau_{10} = 0$. Compared to an equal mixture of χ_1^2 and χ_2^2 distributions, the likelihood ratio statistic of 16.3 is clearly significant ($p < 0.001$), leading me to prefer the more general Model MB2.[4]

Allowing the treatment effects to vary randomly has little effect on the estimates of the average treatment effect across cases. However, the estimates of the variance components change substantially, affecting the denominator of the unadjusted effect size and the degrees of freedom used to adjust it. The total baseline variance is 41% larger in MB2 than in MB1, leading an adjusted effect size estimate that is 16% smaller. The difference between the two models is due to a difference in between-case variance across phases not captured by the assumptions of MB1. The between-case variation in the outcome appears to be substantially larger in the baseline phase than in the treatment phase, but the constant treatment effect in MB1 constrains the between-case variance to be constant across phases. Since the between-case variance is estimated by pooling across both phases, MB1 leads to a smaller estimate than one based on the baseline phase alone.

MB2 allows the between-case variance to change during the treatment phase. In the hypothetical between-subjects experiment, this change would be observed as a difference in variances between treatment and control groups. The expected variance of the control group would be $\text{Var}(Y_{iB}(n)) = \tau_0^2 + \sigma^2$, with RML estimate 617. The expected variance

[4]On reference distributions for likelihood ratio statistics involving variance component restrictions, see Stram and Lee (1994).

of the treatment group would be

$$\text{Var}(Y_{iB}(A)) = \text{Var}(\eta_{0i} + \eta_{1i} + \epsilon_{ij}) = \tau_0^2 + \tau_1^2 + 2\tau_{10} + \sigma^2,$$

with RML estimate 219, less than half as large. Thus, the treatment not only raised the average level of speech production, but also dampened the variation across cases around that level.

4.3. Schutte, Malouff, & Brown (2008)

The third example illustrates Models MB4 and MB5, which include random trends in the baseline phase. Schutte, Malouff, and Brown (2008) evaluated the effect of an emotion-focused therapy program for adults with prolonged fatigue using a multiple baseline across individuals.[5] The design included 13 adults who met clinical criteria for prolonged fatigue. Cases were measured weekly for 2, 5, or 8 weeks in the baseline phase and between 1 and 7 times in the treatment phase, with a maximum of 15 consecutive measurements in all. For each case and measurement occasion, fatigue severity was measured using a self-reported scale that ranged from 1 to 63. I exclude from my analysis data for participant 4 because nearly all of these measurements are at the upper extreme of the scale. Data for the remaining $m = 12$ participants are plotted in Figure 4.3.

Visual inspection and preliminary analysis suggested that it would be necessary to include time trends in any model for these data. In all of the following models, I use the full, piece-wise linear regression specification from (3.2) that alloweds non-zero time trends; the models differ only in whether the case-level regression coefficients are assumed to be

[5]Data for this example were extracted from a table in the original article.

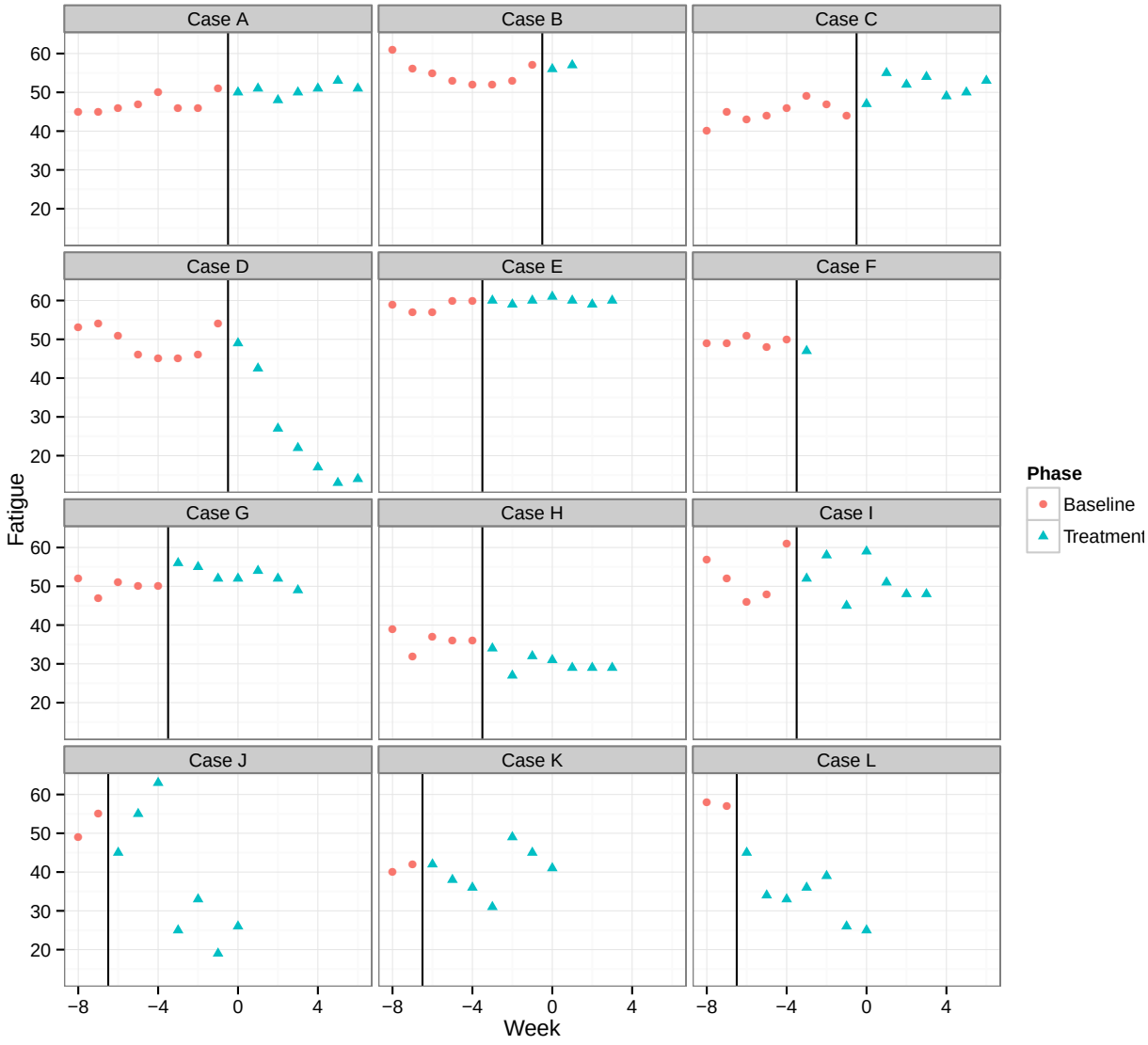


Figure 4.3. Data from Schutte et al. (2008): Fatigue severity over time for adults with prolonged fatigue. The vertical line marks the point of treatment introduction.

constant across individuals or are allowed to vary. In the models under consideration, the effect size parameter depends on the choice of time-points A and B for describing the hypothetical between-subjects design. I use $A = 2$, meaning that in a hypothetical

Table 4.3. Model estimates for Schutte et al. (2008) data

Parameter	Model MB3 RML Estimate	(s.e.)	Model MB4 RML Estimate	(s.e.)	Model MB5 RML Estimate	(s.e.)
Variance components						
Autocorrelation ($\hat{\phi}$)	0.809	(0.023)	0.399	(0.093)	0.240	(0.111)
Within-case var. ($\hat{\sigma}^2$)	99.004	(6.199)	29.394	(4.161)	22.540	(3.353)
Between-case var. ($\hat{\tau}_0^2$)	14.766	(27.157)	95.713	(46.793)	38.301	(33.000)
Case-trend cov. ($\hat{\tau}_{20}$)			11.209	(6.359)	0.384	(3.402)
Trend var. ($\hat{\tau}_2^2$)			1.994	(1.068)	0.147	(0.542)
Case-trend×trt. cov. ($\hat{\tau}_{30}$)					1.741	(7.417)
Trend-trend×trt. cov. ($\hat{\tau}_{32}$)					0.666	(0.962)
Trend×trt. var. ($\hat{\tau}_3^2$)					3.009	(2.700)
Total var. ($\hat{\tau}_0^2 + \hat{\sigma}^2$)	113.770	(27.140)	125.107	(46.756)	60.841	(32.745)
Fixed effects						
Intercept ($\hat{\gamma}_{00}$)	52.932	(4.422)	50.292	(4.074)	50.526	(2.822)
Treatment ($\hat{\gamma}_{10}$)	0.489	(0.622)	0.203	(0.616)	0.219	(0.364)
Trend ($\hat{\gamma}_{20}$)	-1.373	(1.972)	-0.542	(1.752)	0.027	(1.600)
Trend × Trt. ($\hat{\gamma}_{30}$)	-1.896	(0.938)	-1.632	(0.656)	-1.671	(0.741)
Trt. effect after 7 weeks ($\mathbf{p}'\hat{\boldsymbol{\gamma}}$)	-14.646	(6.340)	-11.966	(4.609)	-11.672	(5.184)
Effect size						
Unadjusted ($\hat{\delta}_{AB}$)	-1.373		-1.070		-1.496	
Adjusted (g_{AB})	-1.344	(0.621)	-1.013	(0.469)	-1.328	(0.825)
degrees of freedom (ν)	35.145		14.320		6.904	
Constant (κ)	0.594		0.412		0.665	
Log-likelihood	-435.1		-429.0		-424.5	
Akaike Info. Criterion	884.2		876.0		873.0	

between-subjects experiment, the treatment would be introduced after the second measurement occasion. I also use $B = 9$, meaning that the effect size parameter measures the effect of $B - A = 7$ weeks of treatment, the maximum length observed in the data. Note that for some cases, estimating the effect of seven weeks of treatment involves considerable extrapolation past the observed treatment data. To simplify the calculations, I center the weekly trend at $C = 9$ weeks, so that the case-level intercepts correspond to the average level of the outcome after 9 weeks, in the absence of treatment.

I examine three different models for these multiple baseline data. An initial model assumes that the baseline time trends, the initial treatment effects, and the treatment-by-trend interaction are all constant across cases, but allows the baseline intercept (the average level of fatigue at week 9) to vary across cases. These assumptions correspond to Model MB3. Table 4.3 reports RML estimates of the variance components, the fixed effects, and the effect size for this and following models. Also, Figure 4.4 plots predicted trends in the baseline and treatment phases for each case, generated using the empirical Bayes estimates of the random effects. Based on these trend lines and on formal comparisons to the other models under consideration, it is apparent that MB3 provides a poor fit; the assumption that baseline and treatment trends are constant across cases does not adequately describe these data. Consequently, I do not interpret the model estimates any further.

Next, I consider allowing the baseline time trend to vary randomly across cases, which corresponds to Model MB4. RML estimates are reported in the second column of Table 4.3; predicted trend lines based on this model are plotted in Figure 4.4 using short, dotted lines. A likelihood ratio test comparing MB3 to MB4 rejects the simpler model ($p = 0.001$); visual inspection of the predicted trends also suggests an improved fit.

Based on the RML estimates from Model 4, the intervention has a very small immediate effect, raising participants fatigue scores by $\hat{\gamma}_{10} = 0.2$ scale points, followed by decreases of $\hat{\gamma}_{30} = 1.6$ scale points per week. Combined, the treatment effect after seven weeks of intervention is -12.0 scale points. This treatment effect estimate is in the same units as the outcome measure; to convert it into an effect size, estimates of the model's variance components are needed. The RML estimate of within-case variance is $\hat{\sigma}^2 = 29.4$

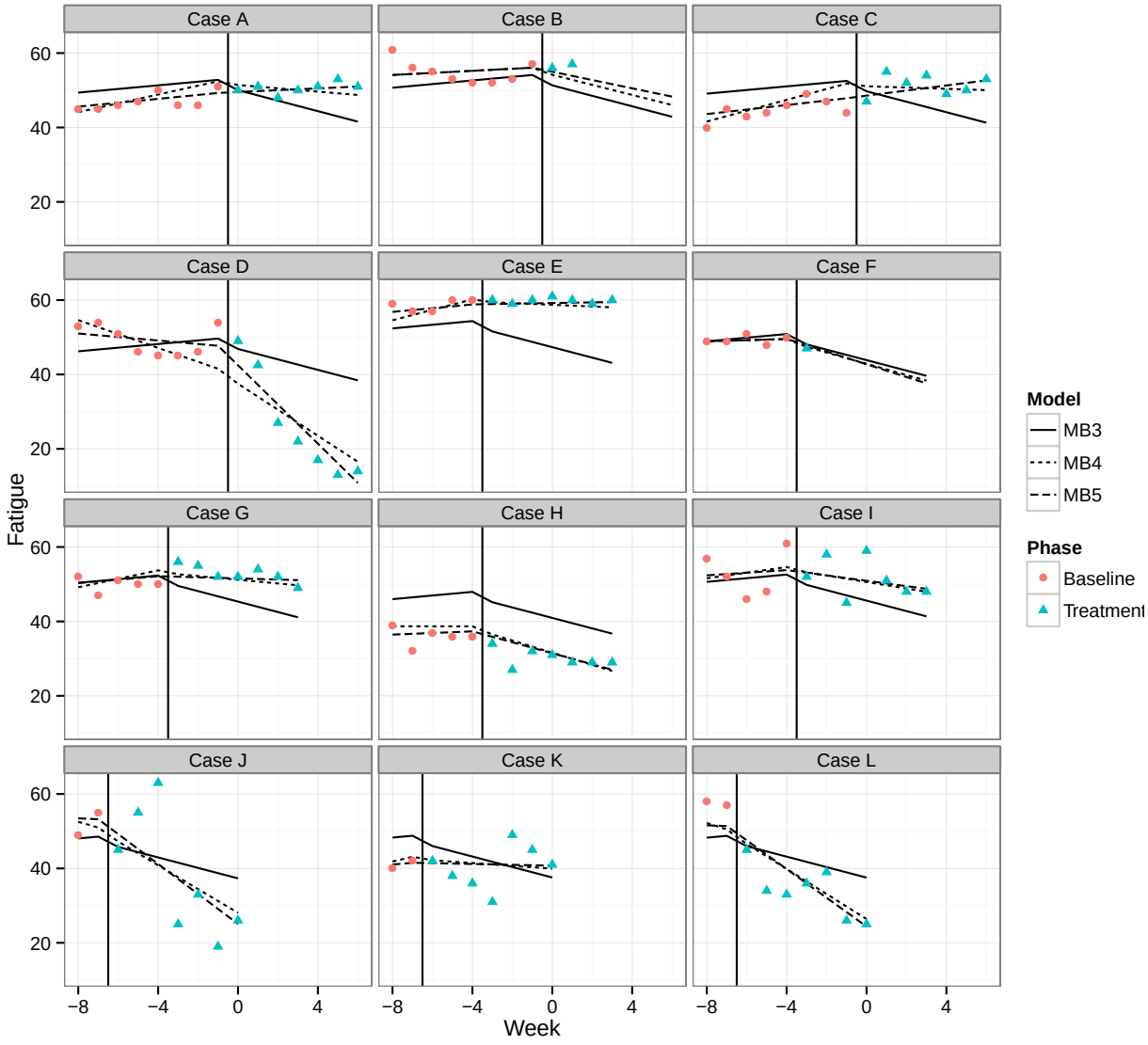


Figure 4.4. Empirical Bayes estimates of case-level trends for Schutte et al. (2008). Trends are extended through 7 weeks of treatment, to illustrate the implicit extrapolation in the target effect size.

squared scale points, assumed to be constant across measurement occasions. The between-case baseline variation is much larger: $\hat{\tau}_0^2 = 95.7$ squared scale points. Recall that the between-case variation is specific to a point in time, because MB4 assumes that the baseline trends vary across cases. Since I have centered time $C = 9$ weeks, τ_0^2 represents the

between-case variation in the average level of the outcome at week 9, in the absence of treatment. Therefore, the total variation at week 9 is $\tau_0^2 + \sigma^2$, the square-root of which goes into the denominator of the effect size. I estimate the unadjusted effect size as $\hat{\delta}_{AB} = (\hat{\gamma}_{10} + 7\hat{\gamma}_{30}) / \sqrt{\hat{\tau}_0^2 + \hat{\sigma}^2} = -1.07$ s.d. After using the estimated degrees of freedom $\nu = 14.3$ to make a small-sample correction, the adjusted effect size is $g_{AB} = -1.01$ s.d., with a standard error of 0.47. According to MB4 estimates, the average baseline trend is slightly negative ($\hat{\gamma}_{20} = -0.54$), though there is variation across cases ($\hat{\tau}_2^2 = 1.99$). Also, steeper trends are highly associated with higher average outcomes at week 9, with a correlation of $\hat{\tau}_{20}/(\hat{\tau}_0\hat{\tau}_2) = 0.81$.

I consider one further model for these data. MB5 assumes that baseline intercepts, baseline trends, and treatment-by-trend interactions all vary across cases, implying a covariance matrix for the random effects that has six parameters. With only twelve cases, use of RML estimation might be questionable if my goal were to draw inferences about the structure of the case-level random effects. However, the simulation evidence presented in the previous chapter tentatively suggests that for purposes of effect size estimation, RML may be a reasonable strategy even with such a limited sample of cases.

The final column of Table 4.3 reports RML estimates for MB5. RML estimation of this model leads to an estimated correlation of one between the random effects of trend and treatment-by-trend interaction (i.e., $\tau_{32}/(\tau_3\tau_2) = 1$). Using an equal mixture of χ_2^2 and χ_3^2 as a reference distribution, the likelihood ratio test for MB5 versus MB4 is statistically significant ($p = 0.021$), though this asymptotic test may provide poor guidance with such a small sample of cases.

Though the estimated fixed effects based on MB5 are very similar to those from MB4, the estimated variance components and effect size are quite different. In particular, the between-case variation is only $\hat{\tau}_0^2 = 38.3$ squared scale points, compared to 95.7 under MB4, and the within-case variation is $\hat{\sigma}^2 = 22.5$ squared scale points, compared to 29.4 under MB4. The unadjusted effect size estimate is $\hat{\delta}_{AB} = -1.50$ s.d.; after the degrees-of-freedom adjustment, $g_{AB} = -1.33$. These estimates are much larger than the corresponding estimates from MB4 due mostly to the much smaller estimate of the between-case variation at week 9, which is only partially tempered by the reduced degrees of freedom.

MB5 allows the effect of the treatment on the slope of the outcome series to vary across cases, unlike in MB4. The RML estimate of the treatment-by-trend variance is $\hat{\tau}_3^2 = 3.01$, corresponding to a standard deviation of 1.73 scale points per week. One can get a sense of the extent of variation in treatment effects by comparing the effect of a 7 week treatment course across cases. Using the empirical Bayes estimates of individual random effects, the individual effects after 7 weeks of treatment range from -33.53 to 0.57 scale points; for case C, MB5 predicts that the treatment actually leads to slightly increased fatigue.

4.4. Anglesea, Hoch, & Taylor (2008)

The fourth example comes from a treatment reversal design and illustrates Models TR1 and TR2. This example was also analyzed in Hedges et al. (2012b), where it was chosen to demonstrate their calculations on a design with a very small number of cases.

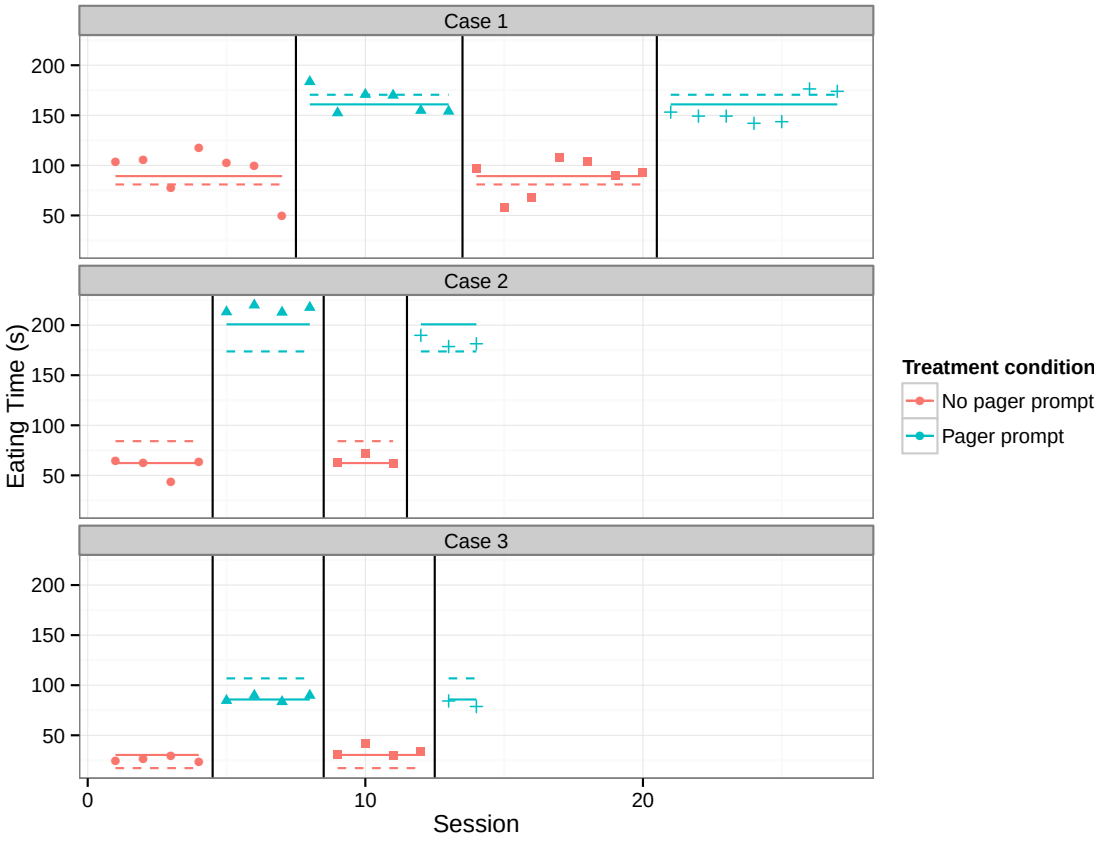


Figure 4.5. Data from Anglesea et al. (2008): Eating time in seconds by session. Solid vertical lines denote changes between treatment and no-treatment phases. Horizontal lines represent the empirical Bayes estimates of each child's average eating time, based on Model TR1 (dashed lines) or TR2 (solid lines).

Anglesea, Hoch, and Taylor (2008) used an ABAB design to evaluate the effect of using an electronic pager prompt to moderate the rapid eating of children with autism. The design included $m = 3$ teenage males. Cases were measured between 2 and 7 times in each of the four phases. The primary outcome measure was a latency measure: the amount of time (in seconds) taken to consume a target food during lunch time. Figure 4.5 displays a plot of the data from each case.

The HPS estimation methods rely on the assumptions of Model TR1, which posits that cases vary in the level of the outcome but that the treatment has a constant, additive effect for all cases. To facilitate comparison of estimation methods, Table 4.4 reports the HPS estimates of TR1 in column 1 and the RML estimates, including the adjusted RML effect size estimate, in column 2. The two methods produce substantially different effect size estimates: based on HPS, the effect is estimated to be 1.15 s.d., while based on RML, the effect is estimated as 1.491 s.d., nearly 30% larger. The difference between the two estimates stems largely from a difference in the degrees of freedom used for bias correction; with such small values of ν, bias correction makes a substantial difference. The difference in the degrees of freedom stems in turn from differing variance component estimates: the HPS method leads to a very large within-case reliability $\hat{\tau}_0^2/(\hat{\tau}_0^2 + \hat{\sigma}^2)$ of 0.92 while the RML estimate is 0.73.

Given that the two estimation methods are based on the same model, it may seem odd that they produce such different variance component estimates. I offer two observations to shed light on the discrepancy. First, the HPS method estimates the within-case variance σ^2 using within-case, within-phase sample variances, which do not constrain the treatment effect to be constant across either cases or phases; in contrast, the RML method relies on those constraints, which inflates the estimated within-case variance (as well as the auto-correlation estimate). Second, the HPS method uses an estimate of the total variance that is pooled across phases, an approach that implicitly assumes homogeneity of variance. The HPS and RML estimates of the total variance are therefore much less incongruous than respective estimates of the components. Taken as a whole, the dissimilarity between

Table 4.4. Model estimates for Anglesea et al. (2008) data

	Model TR1				Model TR2	
	HPS		RML		RML	
Parameter	Estimate	(s.e.)	Estimate	(s.e.)	Estimate	(s.e.)
Variance components						
Autocorrelation ($\hat{\phi}$)	0.176		0.498	(0.143)	0.200	(0.158)
Within-case var. ($\hat{\sigma}^2$)	198.4		569.5	(164.2)	218.8	(48.8)
Between-case var. ($\hat{\tau}_0^2$)	2149.5		1509.6	(1600.6)	900.3	(933.1)
Case-treatment cov. ($\hat{\tau}_{10}$)					274.2	(996.3)
Treatment var. ($\hat{\tau}_1^2$)					2000.9	(2063.2)
Total var. ($\hat{\tau}_0^2 + \hat{\sigma}^2$)	2347.8		2079.1	(1604.2)	1119.1	(933.9)
Fixed effects						
Intercept ($\hat{\gamma}_{00}$)			60.762	(23.340)	60.633	(17.636)
Treatment ($\hat{\gamma}_{10}$)	86.870		89.595	(7.181)	88.503	(26.224)
Effect size						
Unadjusted ($\hat{\delta}_{AB}$)	1.793		1.965		2.645	
Adjusted (g_{AB})	1.150	(1.562)	1.491	(0.988)	1.889	(1.858)
degrees of freedom (ν)	2.340		3.359		2.872	
Constant (κ)	0.091		0.157		0.784	
Log-likelihood			-241.5		-227.3	
Akaike Info. Criterion			493.1		468.7	

the results of the two estimation methods highlights the inadequacy of Model TR1 as a description of the data.

Table 4.4 also reports the RML fit of Model TR2, which allows the treatment effect to vary across cases. The less-constrained model leads to a greatly improved fit, whether judging by visual inspection of Figure 4.5 or using a formal criterion.[6] The fixed effect estimates of TR2 remain nearly identical to those from the RML fit of TR1, but the variance component estimates change drastically. The between-case baseline variance estimate $\hat{\tau}_0^2$ shrinks from 1509 to 900 while the within-case variance $\hat{\sigma}^2$ shrinks from 570

[6]A likelihood ratio test of TR1 versus TR2 rejects the more constrained model ($p < 0.001$).

to 219.[7] This leads to an estimate of the total variance in the absence of treatment that is only 54% of the estimate from TR1, and consequently an effect size estimate that is 27% larger than the RML estimate of TR1 (64% larger than the HPS estimate). Note as well that the effect size estimate has a very large standard error; heuristically, this is because allowing the between-case variation to change by treatment condition leaves even less data available to estimate the total variation across cases in the absence of treatment.

Of the two models and three sets of estimates that I have presented, I prefer the RML estimate of TR2 because it appropriately captures the variability in the treatment effect across cases. On average, the effect of the pager prompt is to increase the time taken to consume the target foods by 1.9 s.d. (with a standard error of 1.9 s.d. units). The pager prompt also appears to have a highly variable effect across cases, since the estimated standard deviation of the individual treatment effects is $\hat{\tau}_1 = 44.7$ seconds, or $\hat{\tau}_1/\sqrt{\hat{\tau}_0^2 + \hat{\sigma}^2} = 1.3$ s.d. units.

4.5. Lambert, Cartledge, Heward, & Lo (2006)

The final example comes from a treatment reversal design, chosen to explore the non-linear models TR3 and TR4. Because comparison of estimation methods is not the focus, HPS estimates of Model TR1 are omitted even though this example was also analyzed in Hedges et al. (2012b).

[7]Note that the RML estimate of the within-case variance from TR2 is much closer to the HPS estimate. This is because HPS estimates σ^2 using within-case, within-phase sample variances, a method that allows for between-case heterogeneity as in Model TR2. The RML estimate of TR2 remains somewhat larger than the HPS estimate because the latter also partitions out variation between replications of the same treatment condition on each case.

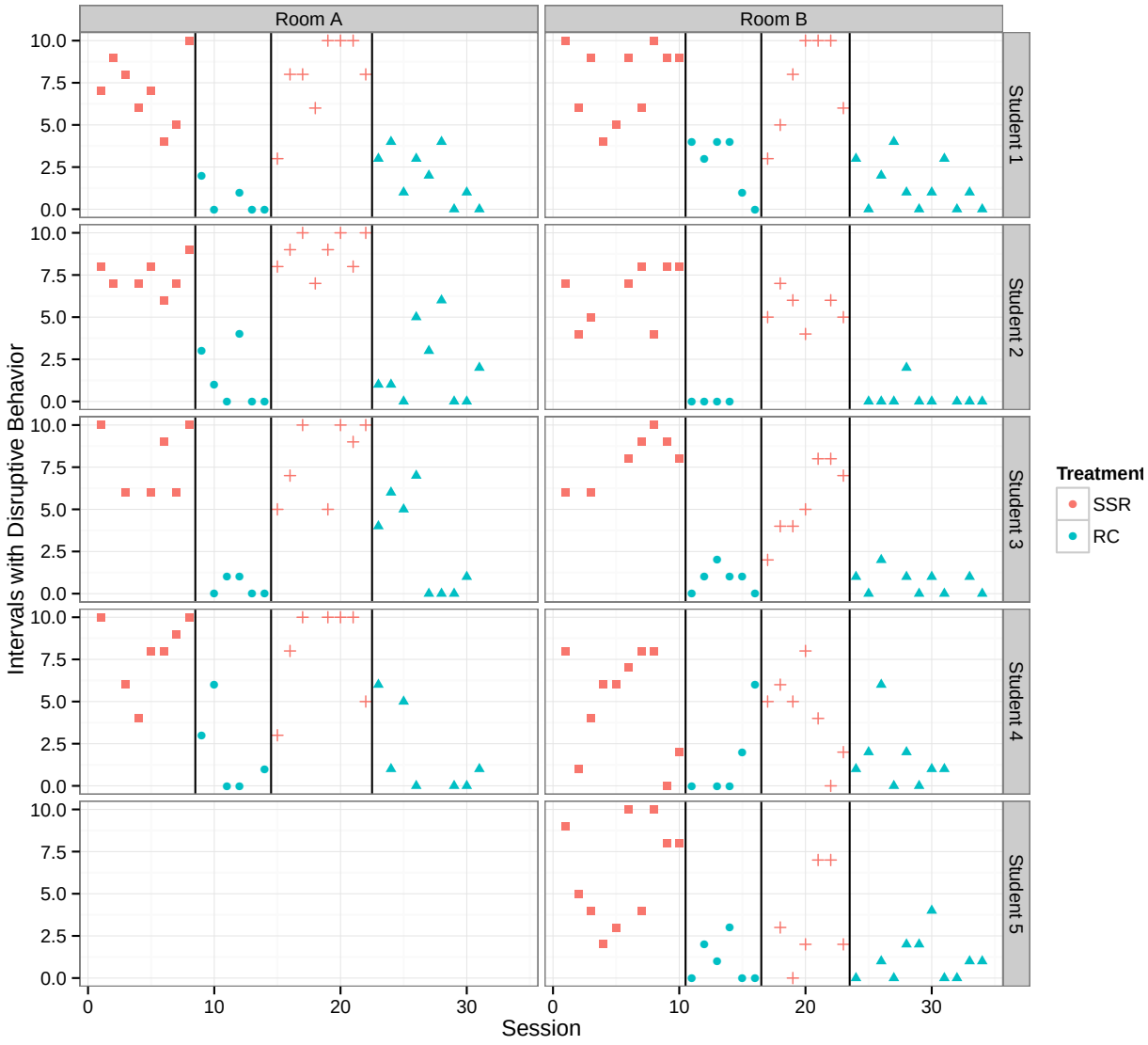


Figure 4.6. Data from Lambert et al. (2006): Number of intervals with disruptive behavior by session, for 9 students across two classrooms. Teachers used conventional single-student responding during "SSR" phases and response cards during "RC" phases. Solid vertical lines denote changes between treatment and no-treatment phases.

Lambert et al. (2006) evaluated the effect of a particular classroom teaching technique–using response cards during question-and-answer sessions versus more conventional single-student responding–on the disruptive behavior of fourth grade students. Using an ABAB

design and a partial interval recording technique, the investigators measured the level of disruptive behavior among nine target students across two classrooms. I cited this study in Section 2.3.4 as an example of an intervention at a higher level of assignment (the classroom, versus the student). It is difficult to properly account for this nested structure because the study took place in just two classrooms; consider that a (hypothetical) randomized trial conducted on the same sample would have had only one independent unit in each treatment condition. For present purposes, I neglect this aspect of the study design, instead treating each student as an independent case. Figure 4.6 plots the data from each of the nine students.

I begin by fitting the linear models TR1 and TR2 to these data. Table 4.5 reports the RML estimates of the model parameters and the adjusted RML effect size estimates. Based on TR1, using response cards rather than single-student responding reduces student disruption by about 5.4 intervals per session, per student (i.e., over 50% of the possible intervals) for an adjusted standardized mean difference of -2.4 s.d. TR2 fits only slightly better than TR1.[8] The less constrained TR2 suggests that there is a small degree of treatment heterogeneity across cases, but also produces correlation of -1 between the treatment effect and the average level of disruption during single-student responding, a rather suspect estimate.

Models TR1 and TR2 assume that the level of each student's disruptive behavior is constant in the absence of intervention, and that the treatment effect has an immediate, transient effect on the level of disruption. Visual inspection of Figure 4.6 suggests that the treatment effect might not be fully transient (note in particular the curvature evident

[8]Using an even mixture of χ^2_1 and χ^2_2 distributions, a likelihood ratio test of TR1 versus TR2 is only marginally significant ($p = 0.050$).

Table 4.5. Model estimates for Lambert et al. (2006) data

	Model TR1 RML		Model TR2 RML		Model TR4 RML, $\omega = 0.294$	
Parameter	Estimate	(s.e.)	Estimate	(s.e.)	Estimate	(s.e.)
Variance components						
Autocorrelation ($\hat{\phi}$)	0.267	(0.064)	0.243	(0.067)	0.212	(0.067)
Within-case var. ($\hat{\sigma}^2$)	4.528	(0.445)	4.343	(0.434)	4.054	(0.399)
Between-case var. ($\hat{\tau}_0^2$)	0.473	(0.371)	1.220	(0.829)	1.525	(0.968)
Case-treatment cov. ($\hat{\tau}_{10}$)			-0.905	(0.723)	-1.317	(0.930)
Treatment var. ($\hat{\tau}_1^2$)			0.671	(0.752)	1.138	(1.015)
Total var. ($\hat{\tau}_0^2 + \hat{\sigma}^2$)	5.001	(0.547)	5.563	(0.907)	5.578	(1.025)
Fixed effects						
Intercept ($\hat{\gamma}_{00}$)	6.796	(0.320)	6.809	(0.427)	7.021	(0.462)
Treatment ($\hat{\gamma}_{10}$)	-5.396	(0.309)	-5.413	(0.405)	-5.990	(0.471)
Effect size						
Unadjusted ($\hat{\delta}_{AB}$)	-2.413		-2.295		-2.536	
Adjusted (g_{AB})	-2.402	(0.191)	-2.272	(0.255)	-2.504	(0.308)
degrees of freedom (ν)	167.130		75.320		59.295	
Constant (κ)	0.138		0.172		0.199	
Log-likelihood	-569.7		-567.1		-560.6	
Akaike Info. Criterion	1149.3		1148.2		1135.2	

in the second single-student responding phase). I now explore whether models TR3 and TR4, which allow for more gradual forms of transience, can better capture the shape of the mean outcome process across phases.

Following the method described in Section 3.3.6, I fit the non-linear models by profiling in the decay parameter ω. Figure 4.7 displays the profile log-likelihood as a function of ω for models TR3 and TR4. The log-likelihood of TR4 is maximized at $\hat{\omega} = 0.294$, reaching a value of -560.6 (compared to -564.8 for TR3 at the same value of ω). The final column of Table 4.5 reports the RML fit of TR4, conditional on the estimated $\hat{\omega}$. The estimated treatment effect is somewhat larger ($\hat{\gamma}_{10} = -6.0$) than in the other models, while the total variance estimate is comparable to that of TR2. As with TR2, the correlation between

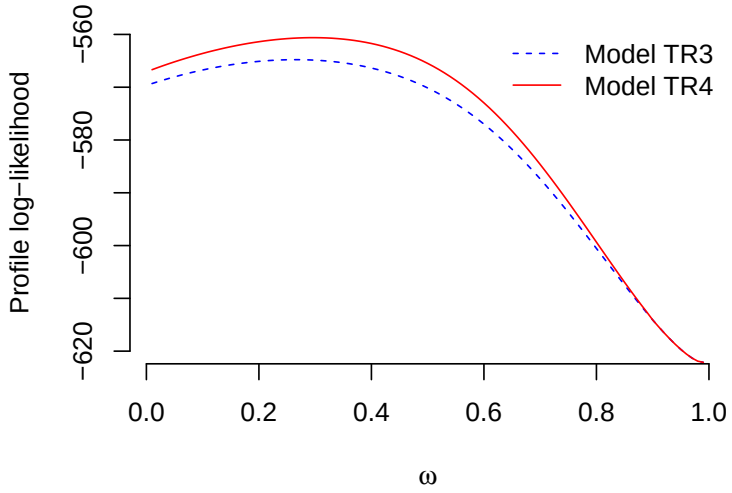


Figure 4.7. Profile log-likelihood in ω for Models TR3 and TR4, based on Lambert et al. (2006) data.

the treatment effect and the level of the outcome in the absence of treatment is estimated as -1.

In the non-linear TR4, the magnitude of the design-comparable standardized mean difference depends on the pattern of treatment assignment. To provide the greatest comparability with TR1 and TR2, I consider the effect size $\gamma_{10}/\sqrt{\tau_0^2 + \sigma^2}$, which can be interpreted as the equilibrium effect of continually using the treatment.[9] Based on the estimates for TR4, the equilibrium effect size is estimated as $g_{AB} = -2.5$ s.d., slightly larger than the estimates from the simpler, linear models, and with a slightly larger standard error.

As a sensitivity analysis, I calculated an approximate confidence interval for ω using the log-profile likelihood and examined the effect size estimate for values within this

[9]Note also that with the estimated decay parameter of $\hat{\omega} = 0.294$, the effect of using response cards for four or more days is substantively equal to the equilibrium effect.

interval.[10] I found that that the effect size estimate ranged from -2.40 to -2.59 s.d., with larger values of ω leading to more negative effect size estimates in an approximately linear relation over the range of values considered. Thus, the effect size estimate is relatively insensitive to the value of ω considered, which provides some small justification for the profiling method used to estimate it.

Figure 4.8 plots the empirical Bayes estimates of the average level of disruptiveness for each student in each phase of the study. The non-linear model TR4 appears to capture some of the curvature in the outcomes that is evident just after changes in treatment condition, leading me to prefer it over the linear TR1 or TR2. Actually though, none of the models considered provides a very good description of the data because the outcome measure has a constrained range of [0,10] and much of the data are near the extremes of that range. A sound model for this study will need to better account for the properties of the outcome measurements, a challenge to which I turn in later chapters.

4.6. Discussion

I have presented five examples demonstrating proposed models and estimation methods for a design-comparable standardized mean difference effect size, using data from a selection of single-case designs. I conclude this chapter by offering several thoughts stimulated by these applications.

First, in several of the examples I compared the results of fitting the same model via RML versus using the HPS methods. In most cases, the two methods produced

[10]For a fixed value of ω, -2 times the profile likelihood of the RML estimator $\hat{\omega}$ is asymptotically χ^2_1-distributed (Bickel & Doksum, 2007, p. 395). An approximate $1-\alpha$-level confidence interval can therefore be defined defined as the set of values of ω with profile likelihood greater than $l_p(\hat{\omega}) - \frac{1}{2}\chi^2_1(1-\alpha)$, where $l_p(\omega)$ is the profile likelihood at ω and $\chi^2_1(p)$ is the p^{th} quantile of the χ^2 distribution with one degree of freedom. In this example, a 95% confidence interval is given by $\{\omega : l_p(\omega) > 560.6 - 1.9\} = (0.15, 0.43)$.

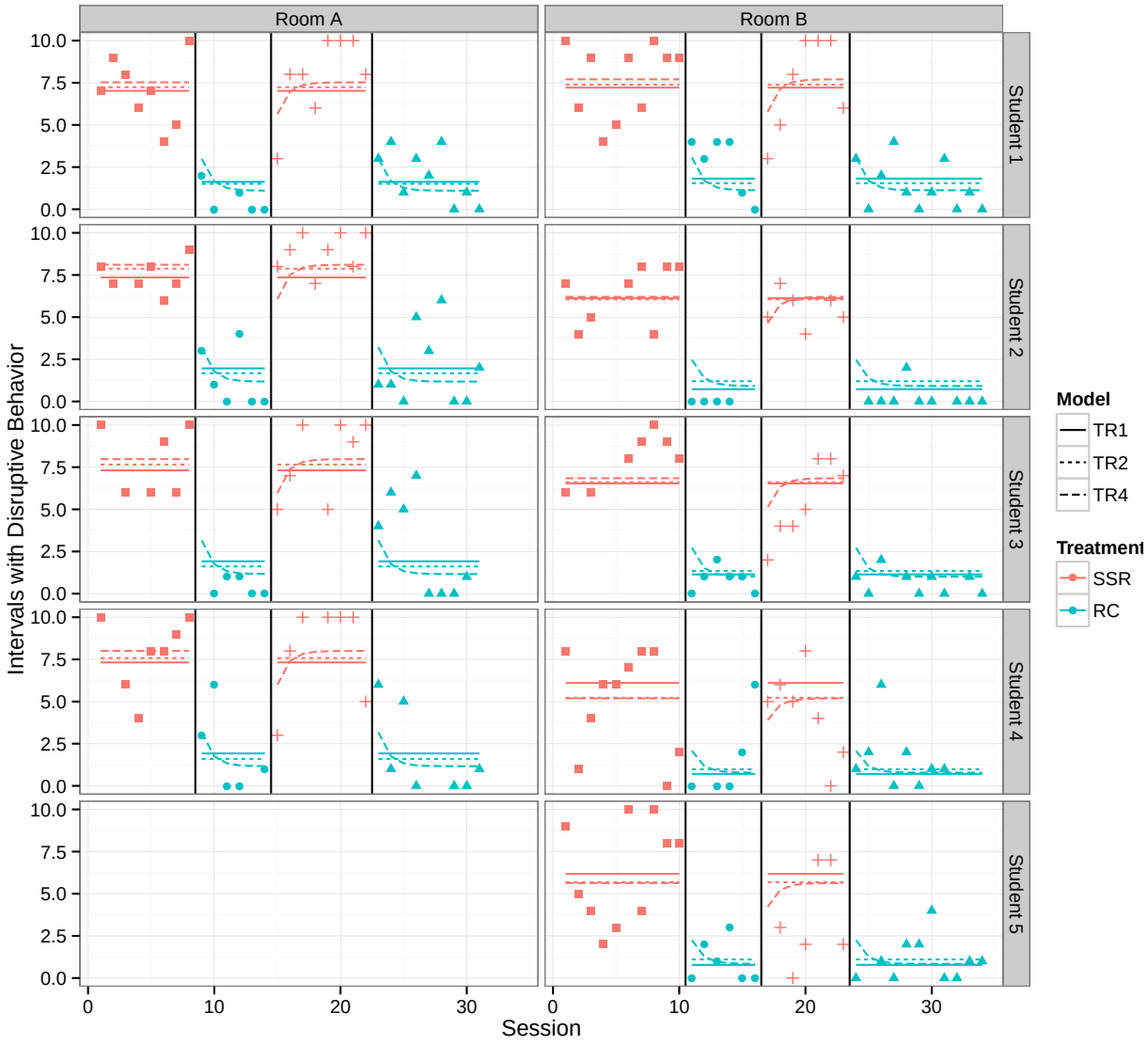


Figure 4.8. Empirical Bayes estimates for Lambert et al. (2006) data, based on Models TR1, TR2, and TR4.

very similar effect size estimates when based on the same underlying model. However the discrepancy seen in the Anglesea et al. (2008) example demonstrates a drawback of the nuisance parameter estimators ued by HPS. Specifically, the HPS estimator of the within-case variance in models MB1 and TR1 uses assumptions that are not internally

consistent with the assumptions behind their estimator of the total variance, leading to an odd estimator for the within-case reliability ρ. This is a relatively minor issue, given that the HPS effect size estimator is designed to be robust to mis-estimation of the nuisance parameters, but it made a difference in a study where the degrees of freedom were very small. The problem could be mitigated by using RML estimates of the nuisance parameters while otherwise following the HPS method for effect size estimation. More importantly, in none of the examples was the simplest model–the only one considered by HPS–preferred over the alternatives, which again highlights the importance having estimation methods that are extensible.

Second, the applications in this chapter were based on a rather rudimentary model selection strategy. I have proceeded by relying on graphical representations of the data, augmented by empirical Bayes estimates of the case-level trends. Such graphs, which depict model predictions as well as the raw data, may be a particularly important and valuable tool in this context due to the long and established tradition of visual analysis in single-case research.[11] I have also reported likelihood ratio tests (Stram & Lee, 1994) to support my model selections, but it is important to keep in mind that these tests rely on asymptotic approximations that may be a poor when the data contain only a few independent cases (Crainiceanu, Ruppert, & Vogelsang, 2003). Furthermore, I noted in Chapter 3 that there single-case researchers may have a strong *a priori* preference for more richly parameterized models, in which case likelihood ratio tests would not be immediately relevant.

[11]Empirical Bayes predictions can also be a useful tool for practitioners who use single-case research in clinical settings, insofar as they can provide improved estimates of treatment effectiveness for each individual case (Zucker et al., 1997).

Finally, throughout this chapter I have tried to demonstrate useful practices in terms of graphical displays and reporting model estimates. However, due to the focus on estimation of design-comparable effect sizes, my presentation has omitted relevant details regarding model checking. Future work will need to identify good tools for model assessment, explain their use, and develop guidance around what details should be presented in journal articles that report primary analyses of single-case studies. This will be important if hierarchical modeling of single-case data is to become a relevant and practical analytic strategy.

CHAPTER 5

Measurement-comparable effect sizes for free-operant behavior

A desirable characteristic of any effect size measure is that its magnitude should not depend on the operational details of how the outcome construct was measured. Effect sizes that have this property are *measurement comparable* (see Section 1.1.2). Without measurement comparability, it becomes very difficult to draw meaningful inferences from averages across and comparisons between effect size estimates because true variation in magnitude is confounded by differences between measurement scales.

Previously proposed effect sizes for quantifying the results of single-case studies have addressed measurement comparability in one of two ways. First, some have argued that standardizing outcomes based on within-case sample variances makes the resulting effect sizes comparable (e.g., Hershberger et al., 1999; Swaminathan et al., 2010; Van den Noortgate & Onghena, 2003b). Second, others have proposed effect size metrics on a 0-to-100% scale, arguing that a uniform scale permits direct comparison (e.g., Parker, Vannest, & Davis, 2011; Scruggs & Mastropieri, 2012). For the most part, proposed effect sizes have been described in terms of computational formulas, unmotivated by any particular statistical model. Claims of measurement comparability have therefore gone unscrutinized.

In this chapter, I propose several effect size measures that are motivated by a particular stochastic model, which permits their measurement comparability (or lack thereof) to be established in precise and formal terms. In contrast to previous chapters, I set aside

issues of design comparability, focusing instead on effect sizes for quantifying treatment effects at the level of the individual case. I do so because individual-level effect sizes are more intuitively interpretable and more closely aligned with the dominant conception of treatment effects in single-case research.

Rather than attempting to model the comparability of any and all outcome measurement operations used in single-case research, I limit the scope of this investigation to one particular class: direct observations of behavior in free-operant contexts. Free-operant contexts are defined by a setting or time-frame in which behaviors are free to occur at any time, without prompting or restriction by the investigator. This scope limitation is appropriate for three reasons. First, as discussed in Section 1.2.2, it is very common for single-case studies to use outcomes based on direct observation of free-operant behavior; thus, even methods specialized to this outcome domain will still be widely applicable. Second, empirical meta-analyses of single-case research often employ similar scope limitations in defining a search strategy (e.g., Gage, Lewis, & Stichter, 2012; Hart & Banda, 2009; Shogren, Faggella-Luby, Bae, & Wehmeyer, 2004) or draw similar distinctions between outcome domains at the analysis stage (e.g., Machalicek et al., 2008). Finally, a measurement comparability model that is applicable across multiple outcome domains will almost certainly involve stronger and more tenuous assumptions than a model for a single domain. Models for other outcome domains and for cross-domain comparisons remain a topic for future work.

Within the domain of directly observed, free-operant behavior, several different procedures are commonly used to record measurements. As noted in Section 1.2.2, four of the most common procedures are event counting, continuous recording, momentary time

sampling, and interval recording.[1] Using *event counting* (also known as frequency recording or the tally method), the observer notes the start of each occurrence of a behavior, either by recording the time of each occurrence or simply by tallying the number of occurrences. Often, the data from an observation session are summarized by the number of occurrences per fixed unit of time. Using *continuous recording* (also known as duration recording), the observer notes the beginning and end of each instance of a behavior. This technique is aided by the use of electronic recording equipment for noting times and behavior codes, though it can also be done simply by starting and stopping a timer. Often, the data from an observation session are summarized by the proportion of session time during which a behavior is observed. Using *momentary time sampling* (also known as time-sample recording), the observer notes whether a behavior is or is not occurring at each of a set of fixed moments in time, which are typically equally spaced over the course of a session. Often, the data from an observation session are summarized by the proportion of moments at which the behavior is observed.

Finally, interval recording techniques have a long history within behavioral analysis and go by several different names, including interval sampling, one-zero sampling, modified frequencies, Hansen sampling, or simply time-sampling (Mann et al., 1991). Two variants of the technique are partial interval recording and whole interval recording. In both variants, an observation session is divided into short time intervals, sometimes with a short break in between each interval to allow time for recording observations.[2] In *partial interval recording*, each interval receives a score of one if the behavior occurs at any point

[1]Terminology varies somewhat across authors. My presentation follows the terminology and the main outline of Ayres and Gast (2010).

[2]For instance, a 20 minute session may be divided into 60 intervals, each with 15 seconds for observation and 5 seconds for recording.

during the interval, and otherwise receives a score of zero. In the less common variant of *whole interval recording*, an interval is scored as a one if the behavior occurs for the entire interval, and is otherwise scored as a zero. Both variants are typically summarized by the proportion of intervals receiving scores of one (equivalently, the mean score across the intervals).

The properties of the main direct observation recording procedures have long been subject to scrutiny and debate. Much of the debate has centered on the theoretical interpretation and practical utility of interval recording methods (J. Altmann, 1974; A. Harrop, Daniels, & Foulkes, 1990; Mann et al., 1991). The sensitivity of results to variation in recording methods has been studied through simulations (e.g., A. Harrop & Daniels, 1986; Rapp, Colby-Dirksen, Michalski, Carroll, & Lindenberg, 2008) and through empirical examples (e.g., Alvero, Struss, & Rappaport, 2007; Bornstein, 2002; Gardenier, MacDonald, & Green, 2004; Gunter, Venn, Patrick, Miller, & Kelly, 2003; Murphy & Goodall, 1980; J. Powell, Martindale, & Kulp, 1975; Rapp et al., 2007), but rarely through formal modeling. The most relevant exception is Rogosa and Ghandour (1991), who used alternating renewal process models to study the psychometrics of behavioral observation procedures.

In practice, various operational procedures might be applied to measure very similar constructs. For example, Shogren et al. (2004) conducted a systematic review of single-case studies examining the effects of providing choice-making opportunities on the problem behavior of disabled children.[3] They identified 13 studies meeting search criteria, including a total of 32 individual cases. The primary outcome for each case was problem behavior, but the procedures used to measure problem behavior varied across cases and studies.

[3]The authors synthesized the results of identified studies using percentage of non-overlapping data (PND) and percentage of zero data (PZD) as effect size metrics.

Table 5.1 reports the study design and the number of individual cases measured using each recording procedure for each study included in the review. For the majority of studies and cases, problem behavior was measured using interval recording (following a partial interval recording procedure in all but one case). The next most commonly used method was continuous recording, applied in one study with five cases.[4] This systematic review, which attempted to synthesize studies using heterogeneous operational procedures for measuring a common construct, motivates the in-depth consideration of measurement comparability presented in this chapter.

In examining the comparability of different measurement procedures, I follow an approach similar to Rogosa and Ghandour (1991), using a stochastic model for the free-operant behavior that is observed over the course of an observation session, or what is sometimes called the "behavior stream" (e.g., Hartmann & Wood, 1990; Schoenfeld, 1972). Because the model for the behavior stream has to do with measurements made and recorded during the course of a single observation session, I call it the *within-session* model. In addition to the within-session model, a *between-session* model is needed to describe changes in behavior across subsequent observation sessions and phases of the design. This chapter considers the simplest possible such model, assuming that the behavior stream process is stable within a given phase, and therefore that repeated measurements

[4]Four cases were measured using procedures that I have categorized as "other." All of these studies measured individuals' problem behavior as they performed a set of fixed task steps that varied from case to case, ranging from clerical tasks in a school setting to self-care or house-keeping in a residential setting. In these cases, outcomes were reported as the proportion of task steps during which the individual displayed problem behavior.

Table 5.1. Studies in Shogren et al. (2004) meta-analysis

		Number of participants by procedure				
Study	Design	Event counting	Continuous recording	Momentary time sampling	Interval recording	Other
Dyer et al. (1990)	ABAB				3	
Dunlap et al. (1994)	ABAB				3	
Bambara et al. (1995)	ABAB					1
Seybert et al. (1996)	ABAB				3	
S. Powell and Nelson (1997)	ABAB			1		
Moes (1998)	AB/BA				4	
Dibley and Lim (1999)	ABAB					1
Frea et al. (2001)	Multiple Baseline	1				
Jolivette et al. (2001)	BAB				3	
Kern et al. (2001)	ABAB	1			2	
Peterson et al. (2001)	Functional Assessment				1	
Cole and Levinson (2002)	ABAB					2
Romaniuk et al. (2002)	ABAB	1	5			
Total		3	5	1	19	4

are identically distributed and uncorrelated. Chapter 6 considers more elaborate between-session models that allow deterministic time trends and serially dependent variation in behavior.

The remainder of this chapter is organized as follows. In Section 5.1, I describe and analyze a within-session model of the major measurement procedures for direct behavioral observation in free-operant contexts. In Section 5.2, I define several different effect size measures for quantifying changes in free-operant behavior and note the relationships among the different measures. In Section 5.3, I propose simple moment estimators that can be applied when behavior is stable within phases, repeated measurements are uncorrelated, and the outcome is a direct measure of a behavioral parameter. In Section 5.4, I sketch several methods for use with partial interval recording data, which is not a direct measure. Section 5.5 presents an in-depth application of the methods to the systematic review by Shogren et al. (2004). Section 5.6, notes limitations, discusses the choice between alternative effect sizes, and concludes.

5.1. Within-session models for behavior stream, recorded, and reported data

In this section, I present a model that puts the different observation procedures on a common basis. I do this by relating the data generated by different procedures to behavior stream, that is, the sequence of behaviors that occur over the course of an observation session. I then propose a stochastic model for the behavior stream and examine its implications for the data generated by each procedure.

It is important to bear in mind that this section is focused entirely on describing *measurements*, meaning the process by which a single datum is generated. Previous

chapters have referred to data Y_{ij} for case $i = 1, ..., m$ and measurement occasions $j = 1, ..., n_i$; this section is a model for a single measurement Y. I will refer to Y as a *reported* datum, because typically it is only this observation that is reported in a single case study (often in the form of a single-case graph), and only this information that will be available for secondary meta-analysis. In addition to the reported data, I will also refer to *recorded* data, by which I mean measurements that an observer makes during the course of an observation session. I reference the recorded data as an intermediate device for describing the mechanics of different recording procedures and the properties of the reported data.

The behavior stream model considered here, known as an equilibrium alternating renewal process, is a fairly general model that encompasses a wide variety of special cases. Even in the general formulation, its implications are strong enough to characterize the expected value of each type of reported datum. It is this attribute of each procedure that I interpret as the measurand, or target of quantification. In other words, a procedure measures its mean outcome over realizations of the stochastic model, and variation around that mean is simply measurement error.

I proceed by first formalizing the relationships among the behavior stream data, recorded data, and reported datum for each measurement procedure; I then describe the assumptions of the equilibrium alternating renewal process model and discuss the measurement procedures in light of the model.

5.1.1. Behavior stream, recorded, and reported data

I now describe the data generated by a sequence of behavioral events that occur over a single session, using a structure similar to that used in Rogosa and Ghandour (1991).

Over the course of a given session, assume that behavioral events occur sequentially and can be numbered $u = 1, 2, 3, ...$ Let D_u denote the duration of event $u = 1, 2, 3, ...$; let $D_0 = 0$; let E_u denote the length of time between the end of event u and the beginning of event $u + 1$, or what I will call the *interim time*; let E_0 denote the length of time until the first event, with $E_0 = 0$ if event 1 is occurring at the beginning of the observation period. The quantities $\{D_0, E_0, D_1, E_1, D_2, E_2, ...\}$ are the underlying (latent) data that describe the behavior stream from a given session. Further define the counting process

$$N(t) = \sum_{w=1}^{\infty} I\left[\sum_{v=0}^{w-1}(D_v + E_v) \leq t\right], \tag{5.1}$$

where $I()$ denotes the indicator function, so that $N(t)$ is the number of events that have begun by time t. Finally, define

$$Y(t) = \sum_{w=1}^{\infty} I\left[0 \leq t - \sum_{v=0}^{w-1}(D_v + E_v) < D_w\right] = I\left[\sum_{v=0}^{N(t)}(D_v + E_v) - t > D_{N(t)}\right], \tag{5.2}$$

so that $Y(t) = 1$ indicates that an event is occurring at time t and $Y(t) = 0$ indicates that an event is not occuring at time t.

Based on the underlying behavior stream, the recorded data (and thence the reported datum) are derived according to one of the observation procedures. Any of several different recording procedures could in principle be applied to the same behavior stream, generating a reported datum. I will examine five types of recording procedures. In order to differentiate among the reported data produced by each procedure, I will denote the reported datum from an event counting procedure as Y^E, that from a continuous recording procedure as Y^C, that from a momentary time sampling procedure as Y^M, that

Table 5.2. Notation and design parameters for five recording procedures

Quantity	Event counting	Continuous recording	Momentary time sampling	Partial interval recording	Whole interval recording
Reported datum	Y^E	Y^C	Y^M	Y^P	Y^W
Session length	L	L	L	L	L
Number of intervals			K	K	K
Active interval length				l	l

from partial interval recording as Y^P, and that from whole interval recording as Y^W. Throughout, I denote the length of the observation session by L. Several of the recording procedures involve additional design parameters; these are summarized in Table 5.2 and defined in the course of describing each procedure.

Using event counting, the observer notes the beginning of each new event, producing recorded data $\{D_0 + E_0, D_1 + E_1, D_2 + E_2, ...\}$. The reported datum from this procedure is simply number of times that the behavior begins over the course of the session, so that $Y^E = N(L)$.

Using a continuous recording procedure, the observer notes the beginning and end of each behavioral event. If each time is noted, the recorded data can then be used to completely reconstruct the behavior stream data $\{D_0, E_0, D_1, E_1, D_2, E_2, ...\}$. The reported datum from a continuous recording procedure is the proportion of session time during which the behavior occurs:

$$Y^C = \frac{1}{L} \int_0^{L^-} Y(t)dt.$$

Using a momentary time sampling procedure, the observer notes the presence or absence of a behavior at each of K times within a session, equally spaced at intervals of L/K. The recorded data are described by a sequence of binary indicator variables $X_1, ..., X_K$,

where $X_k = Y(kL/K)$. The reported datum is then the proportion of moments at which the behavior is observed, $Y^M = \sum_{k=1}^{K} X_k/K$.

In partial interval recording, an observer first divides the session into K intervals, each of length L/K. The first l time units of each interval are devoted to observation, while the remainder $L/K - l$ is used for recording or resting; I call l the *active interval length*. During a given interval, the observer counts a behavior as present if it occurs at any point during the active interval; indicators of the presence or absence of behavior during each of the K intervals constitutes the recorded data. Let $U_k = 1$ if the behavior occurs at any point during the k^{th} interval, $U_k = 0$ otherwise; formally,

$$U_k = I\left(0 < \int_0^{l^-} Y\left(l + (k-1)L/K\right) dt\right), \tag{5.3}$$

for $k = 1, ..., K$. The reported datum from partial interval recording is typically the proportion of intervals during which the behavior is observed at any point: $Y^P = \sum_{k=1}^{K} U_k/K$.

Whole interval recording is structured in the same way as partial interval recording: the session is divided into K intervals of equal length L/K, with the first l time units of each interval devoted to observation. Only the rule for scoring each interval is different: in whole interval recording, the observer counts a behavior as present only if it occurs for the entire duration of the active part of the interval. Formally, let

$$W_k = I\left[\int_0^{l^-} Y\left((k-1+t)L/K\right) dt = r\right].$$

The sequence $W_1, ..., W_K$ constitutes the recorded data. Note that whole interval recording is equivalent to partial interval recording applied to the absence of a behavior rather than

its presence. The reported datum from whole interval recording is the proportion of intervals during which the behavior is present for the duration: $Y^W = \sum_{k=1}^{K} W_k/K$.

So far, I have shown the relationships between the behavior stream data during the course of a session and the reported datum generated by different recording procedures. These relationships are entirely deterministic. In order to interpret the measurements as random variables, I will treat the underlying behavior stream as a realization of a stochastic process, then derive the properties of the recorded data based on that process.

5.1.2. Equilibrium alternating renewal process

It is useful to examine the properties of the recording procedures under a stochastic model that invokes only fairly weak assumptions, because the properties so established will apply in general and regardless of whatever further, stronger assumptions might be considered. One such model is the equilibrium alternating renewal process (ARP). An equilibrium alternating renewal process involves the following assumptions:

(1) First, interim times are assumed to be identically distributed random quantities with distribution function F_E, survivor function $\tilde{F}_E = 1 - F_E$, and mean $\mathrm{E}(E_1) = \lambda$, where $0 < \lambda < \infty$.

(2) Likewise, event durations are assumed to be identically distributed random quantities with distribution function F_D, survivor function $\tilde{F}_D = 1 - F_E$, and mean $\mathrm{E}(D_1) = \mu$, where $0 < \mu < \infty$.

(3) Further, all interim times and event durations are assumed to be mutually independent.

(4) Finally, the process is assumed to be aperiodic and in equilibrium, so that $\Pr(Y(0) = 1) = \mu/(\mu + \lambda)$.

In addition to the two mean parameters μ and λ, two other quantities will also of interest: the rate of event occurrence, known as the incidence $\zeta = 1/(\lambda + \mu)$, and the event prevalence $\phi = \mu/(\mu + \lambda)$. The incidence is the inverse of the average time between successive events. In an equilibrium process, event prevalence is both the probability that an event is occurring at any specific moment in time and the overall proportion of time that an event is occurring.

Note that the ARP model parameterizes only the first moments of the event durations and interim times, and so can encompass a very wide variety of parametric distributions. For example, event durations and interim times might each be log-normally distributed, gamma-distributed, or exponentially distributed. The ARP applies even if events have constant, positive duration, so long as the distribution of interim times is such that the aperiodicity assumption is satisfied.

The assumptions of the ARP are sufficient to characterize the first moment of the reported datum from each type of recording procedure. These are summarized in Table 5.3. The expectation of an event counting datum is equal to the product of the session length and the incidence of the behavior, a fact established by Blackwell's Renewal Theorem (Kulkarni, 2010, p. 360). The expectation of a continuous recording datum is equal to the prevalence of the behavior (Cox, 1962, p. 101), as might be intuitively suspected since continuous recording is calculated as the sample prevalence of the realized behavior stream. The expectation of a momentary time sampling datum is also equal to the prevalence of the behavior, due to the assumption that the process is in equilibrium. Each of

Table 5.3. Expectations of reported data under an alternating renewal process

Recording procedure	E(Y)
Event counting	ζL
Continuous recording	ϕ
Momentary time sampling	ϕ
Partial interval recording	$\phi + \zeta \int_0^{l^-} \tilde{F}_E(x)dx$
Whole interval recording	$\phi - \zeta \int_0^{l^-} \tilde{F}_D(x)dx$

these three recording procedures yields a reported datum that is a direct measurement of a parameter of the ARP; I therefore refer to them as *direct measurement procedures*.

It is apparent from Table 5.3 that partial and whole interval recording procedures produce measurements that have no immediate interpretation in terms of the parameters of the ARP; thus, they are not direct measurement procedures.[5] Instead, the expectation of an interval recording datum depends on both the prevalence and incidence of the behavior, as well as on the interval length and the distribution of interm times or event durations. I provide derivations of the expectations in Appendix B.1.[6] Although many studies have investigated the sensitivity and accuracy of interval recording procedures using simulations or empirical examples, there is little statistical guidance regarding the

[5]It has long been recognized that interval recording data measures neither prevalence nor incidence. See for instance J. Altmann (1974) for a discussion of the origins and arguments regarding interval recording methods.

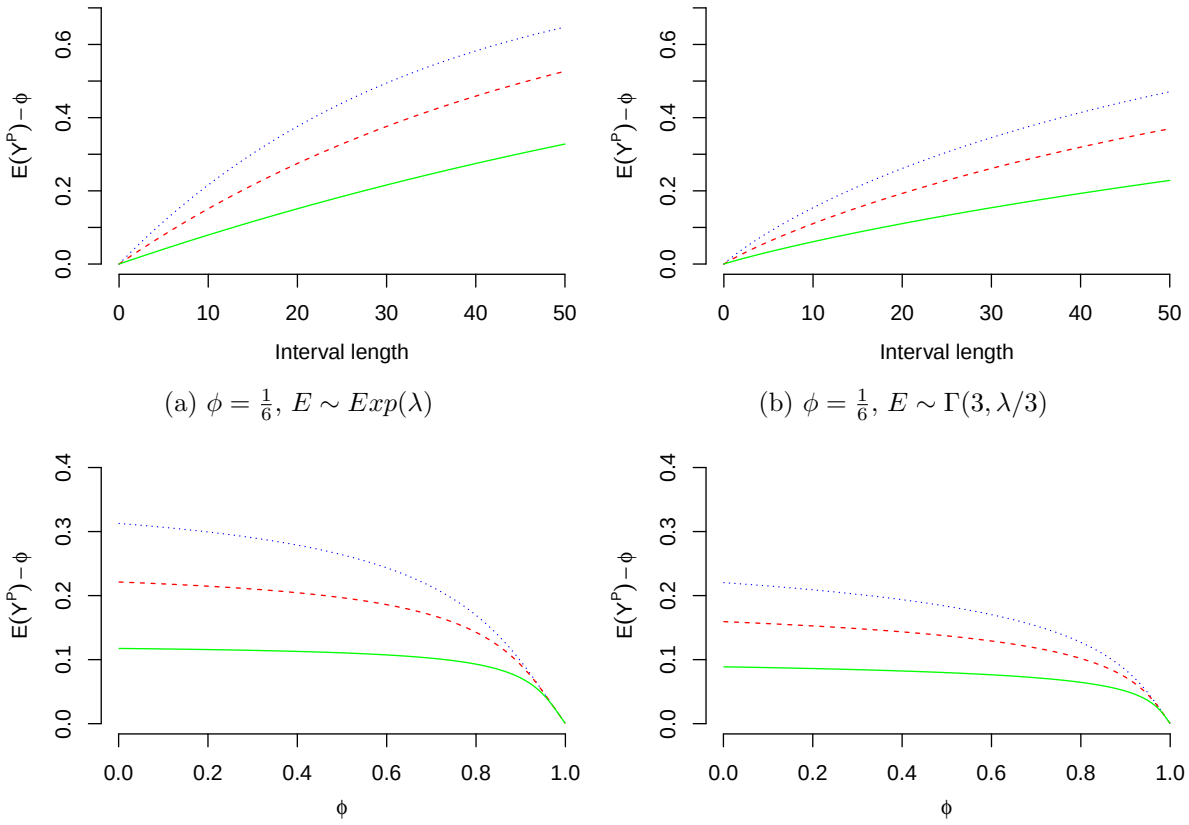
[6]To my knowledge, exact expressions for the expectations of the interval counting procedures have not previously been given in a general form, though two previous analyses should be noted. First, Kraemer (1979) observed that, fixing a realization of the behavior stream, the reported datum from a partial interval recording procedure is an approximately linear combination of the sample prevalence and sample incidence. Second, Rogosa and Ghandour (1991) provided the expectation of a reported partial interval recording datum for the special case of an alternating poisson process, in which both D_1 and E_1 follow exponential distributions.

interpretation of this type of data. It is therefore worth dwelling briefly on its properties. I focus on partial interval recording because it is the more common procedure.

As a measure of behavioral prevalence, the expectation of a partial interval recording datum has bias that depends on the length of the active interval l, the incidence ζ, and the distribution of the interim times F_E.[7] All of these dependencies create complications for the interpretation of partial interval data and are problematic with respect to measurement comparability. Consider first that the bias depends on the incidence and on the chosen interval length. Figure 5.1a plots the bias of Y^P as a function of the interval length for interim times following an exponential distribution, holding prevalence fixed at $\phi = \frac{1}{6}$ and varying incidence $\zeta = \frac{1}{40}, \frac{1}{60}, \frac{1}{120}$. It can be seen that the bias increases with interval length and with incidence, and that for longer interval lengths, the bias is increasingly affected by the incidence. This sensitivity implies that data collected using 15-second partial interval recording is not directly comparable with data collected using 10-second or 30-second intervals. Partial interval recording is thus an operationally sensitive measurement procedure.

Next, consider that the bias depends on the interim time distribution F_E. Figure 5.1b replicates 5.1a, but uses a gamma distribution with shape parameter 3 rather than an exponential distribution; compared to the figure on the left, the same qualitative relationships hold but with differing magnitude. Bias that depends on the entire distribution of interim times (rather than just on the average) is particularly troublesome because it will

[7]Note however that the bias does not depend on the distribution of event durations F_D except through the mean event duration μ.



(a) $\phi = \frac{1}{6}$, $E \sim Exp(\lambda)$

(b) $\phi = \frac{1}{6}$, $E \sim \Gamma(3, \lambda/3)$

(c) $l = 15$, $E \sim Exp(\lambda)$

(d) $l = 15$, $E \sim \Gamma(3, \lambda/3)$

Figure 5.1. Bias of partial interval recording datum $E\left(Y^P\right) - \phi$ as a function of active interval length l, prevalence ϕ, incidence ζ, and interim time distribution F_E. Dotted blue lines correspond to $\zeta = \frac{1}{40}$; dashed red lines correspond to $\zeta = \frac{1}{60}$; solid green lines correspond to $\zeta = \frac{1}{120}$.

often be difficult to obtain information about F_E. Even a primary researcher who has personally collected the observations might find it difficult to characterize this distribution, to say nothing of a secondary meta-analyst.

In a typical application of partial interval recording, a fixed interval length will be used to collected data for multiple observation sessions. Figures 5.1(c) and (d) plot the bias of Y^P for a constant active interval length $l = 15$ while varying ϕ, ζ, and F_E. For

the smallest value of the incidence $\zeta = \frac{1}{120}$, the bias is relatively insensitive to ϕ except when the prevalence is near 1; for larger values of the incidence, the bias is sensitive over wider ranges of ϕ. These relationships hold regardless of whether interim times follow an exponential or a gamma distribution, and suggest a strategy for bounding the magnitude of the bias.

The bias of Y^P as a measure of prevalence can be bounded under certain assumptions about the event durations and interim times. Suppose that bounds for the average event duration can be established using prior experience or other data, so that $\mu_L^* \leq \mu \leq \mu_U^*$ for known μ_L^*, μ_U^*. Suppose further that at most $p^* \times 100\%$ of interim times last less than l, so that $F_E(l) \leq p^*$.[8] It follows that the proportionate bias of Y^P is bounded by

$$\frac{(1-p^*)l}{\mu_U^* + (1-p^*)l} \leq \frac{\mathrm{E}\left(Y^P\right) - \phi}{\mathrm{E}\left(Y^P\right)} \leq \frac{l}{\mu_L^* + l}. \tag{5.4}$$

A derivation can be found in Appendix B.2. Note that the assumption involving p^* is only relevant if an upper bound for μ can be established; otherwise, the lower bound on the bias reduces to 0. Conversely, if no informative bound on $F_E(l)$ can be established, then $p^* = 1$ and the lower bound on the bias reduces to 0, regardless of any assumption about μ_U^*. Still, so long as a lower bound μ_L^* is given, an informative upper bound on the bias of Y^P can be established.

Partial interval recording data is also sometimes interpreted as measuring incidence rather than prevalence, particularly when event durations are known to be very short. A bound on its bias as a measure of incidence can be constructed using an argument

[8]Making an assumption about the proportion of interim times lasting less than a fixed amount of time strikes me as more reasonable and feasible than making an assumption about the entire parametric form of F_E; clearly, however, assumptions such as these will need to be evaluated by researchers with subject-matter expertise, in the context of applications.

similar to the above. Suppose that an upper bound for the average event duration can be established, so that $\mu \leq \mu_U^*$; also suppose that $F_E(l) \leq p^*$ for known p^*. The proportionate bias of Y^P as a measure of incidence is then bouned by

$$\frac{(1-p^*)l - 1}{(1-p^*)l} \leq \frac{\mathrm{E}\left(Y^P\right) - \zeta}{\mathrm{E}\left(Y^P\right)} \leq \frac{\mu_U^* + l - 1}{\mu_U^* + l}, \tag{5.5}$$

See Appendix B.2 for a derivation. I revisit these bounds in Section 5.3, when considering estimation strategies.

5.2. Case-level effect size parameters

The ARP model provides a basis for understanding the relationships among measurements generated by different observation procedures and consequently for establishing the measurement-comparability of different effect sizes. This is because effect sizes defined in terms of the parameters of the ARP can be interpreted in terms of the data from *any* of the measurement procedures, rather than being contingent on the procedure. In this section, I propose several such effect sizes for measuring behavioral changes and describe the relationships among them.

Before turning to the effect sizes, the ARP model needs to be elaborated in order to allow for changes in behavior from measurement occasion to measurement occasion.[9] The model from the previous section described the behavior stream observed on an single measurement occasion, leading to a single reported measurement. Now suppose that interest is in comparing the behavior of an individual under different treatment conditions.

[9]The ARP model itself deals with observation over time, but over the relatively brief spans of time during which behaviors are observed. In single-case studies, observation sessions are typically a matter of minutes, while the time between measurement occasions may be on the order of hours or days; an expanded model is needed to describe this session-to-session time scale.

In this and following sections, I focus on a very simple model in which the behavior stream process is stable under a given treatment condition.

Consider a study in which a total of n outcome measurements are made using measurement procedure r, where $r \in \{E, C, M, P, W\}$. The reported data are then Y_j^r, for $j = 1, ..., n$.[10] Let $T_j = 1$ if the case is in a treatment phase at time j, with $T_j = 0$ otherwise. I assume that for session j, the behavior of the case follows an alternating renewal process and that the reported data are generated by applying measurement procedure r to independent samples from that process. I assume that the parameters of the alternating renewal process are constant within each treatment phase; thus let μ^0 denote the average event duration and λ^0 the average interim time in the baseline phase or phases and let μ^1, λ^1 denote the average event duration and average interim time, respectively, in the treatment phase or phases. These assumptions can be summarized as

$$\left(Y_j^r | T_j = t\right) \overset{\text{iid}}{\sim} M_r\left(ARP\left[\mu^t, \lambda^t\right]\right), \tag{5.6}$$

where iid indicates that measurements on successive occasions within a phase are independent and identically distributed, $M_r()$ denotes the application of measurement procedure r to a behavior stream, and $ARP(\mu, \lambda)$ indicates that the behavior stream is generated by an equilibrium alternating renewal process with mean duration μ and mean interim time λ.

[10] In contrast to previous chapters, the focus is on models for an individual case. For ease of notation, I suppress the i subscript denoting the case.

Under model (5.6), measurement-comparable effect sizes involve comparisons between (μ^0, λ^0) and (μ^1, λ^1). A number of such comparisons are worth considering; I will describe five: the log-duration ratio, the log-interim ratio, the log-incidence ratio, the log-prevalence ratio, and the log-prevalence odds ratio. I focus on logged ratios for two reasons. First and perhaps most crucially, published single-case studies often describe results using measures of percentage change (J. M. Campbell & Herzinger, 2010).[11] Log-ratios are very closely related to proportionate changes, and thus have the advantage of aligning to a certain extent with how applied researchers already think. Second, log-ratios are useful in a purely technical sense to the extent that they conform to the scales of the quantities being measured and can be defined without range restriction. Most of the effect size metrics that I will describe range from negative infinity to positive infinity, with zero corresponding to no change; this allows certain problems with estimation and meta-analysis to be avoided.

5.2.1. Log-duration ratio and log-interim ratio

In the abstract, one of the most informative way to quantify a change in behavior would be to use separate contrasts between each component of the ARP. For instance, the case-level *log-duration ratio*, defined as $\omega^\delta = \ln(\mu^1/\mu^0)$, measures the proportionate change in the average event duration; the case-level *interim ratio*, defined as $\omega^\lambda = \ln(\lambda^0/\lambda^1)$ measures the proportionate change in the average interim time. These two effect sizes would be particularly useful in intervention contexts if the experimenters' goal is to affect change in one dimension of the behavior but not the other, or to evaluate detailed hypotheses

[11] As noted in Section 1.3.2, several percentage-change metrics have been proposed as effect sizes for use with single-case research, though none of these have supporting statistical methodology.

about the mechanism of an intervention. Unfortunately, these effect sizes might only be of hypothetical interest, because none of the observation procedures under consideration yield direct measurements of the separate components.

5.2.2. Log-incidence ratio

The case-level *log-incidence ratio* is defined as $\omega^{\zeta} = \ln(\zeta^1/\zeta^0) = \ln(\mu^0 + \lambda^0) - \ln(\mu^1 + \lambda^1)$. It measures the proportionate change in a behavior's incidence; if observation sessions are of equal length, then the incidence ratio is also equivalent to the log of the proportionate change in the expected number of behaviors during a session. Considering that event counting directly measures incidence and is a very commonly used procedure, the log-incidence ratio should be a very useful effect size for describing changes in behavior. However, as a summary metric it has the disadvantage of not being sensitive to behavioral prevalence. Rather, an observed decrease in incidence could be the result of either an increase in average interim time or an increase in average event duration, with very different substantive implications.

5.2.3. Log-prevalence ratio

When a behavior has non-negligible duration, the foremost concern of interventionists will often be its prevalence, or the overall proportion of time that it occurs. One metric for quantifying changes prevalence is the case-level *log-prevalence ratio*, defined as $\omega^{\phi} = \ln(\phi^1/\phi^0) = \ln\left(\frac{\mu^1}{\mu^1+\lambda^1}\right) - \ln\left(\frac{\mu^0}{\mu^0+\lambda^0}\right)$. This effect size is comparatively straight-forward to interpret in terms of proportionate changes, as with the other log-ratio effect sizes. Compared to the log-incidence ratio, the log-prevalence ratio has the advantage of being

sensitive to changes in both event duration and interim time. However, there are two inter-related drawbacks to this effect size. First, since prevalence ranges from 0 to 1, the effect size has a range that depends on the initial level: for a given initial prevalence ϕ^0, the log-prevalence ratio can never be greater than $-\ln(\phi^0)$. Second, the log-prevalence ratio is not symmetric with respect to how behaviors are defined; re-defining prevalence as the proportion of time that behavioral events do not occur will alter the magnitude of the log-prevalence odds ratio, rather than only affecting the sign. As a consequence of the latter drawback, application of the log-prevalence ratio will require establishing conventions as to how behaviors are defined, such as always defining prevalence in terms of negative or undesirable behavior.

5.2.4. Log-prevalence odds ratio

The case-level *log-prevalence odds ratio* is an alternative metric for quantifying changes in prevalence, and is defined as

$$\psi = \ln\left(\frac{\phi^1(1-\phi^0)}{(1-\phi^1)\phi^0}\right) = \ln\left(\mu^1/\lambda^1\right) - \ln\left(\mu^0/\lambda^0\right).$$

This effect size measures proportionate change in the prevalence odds, or the ratio of the average event duration to the average interim time. As a result, this effect size weighs a given proportionate increase in duration as equal to a corresponding proportionate decrease in interim time. In contrast to the log-prevalence ratio, its range is unconstrained by the initial prevalence ratio (instead, it ranges from $-\infty$ to ∞) and it is symmetric with respect to how behaviors are defined. These mathematical advantages come at the cost

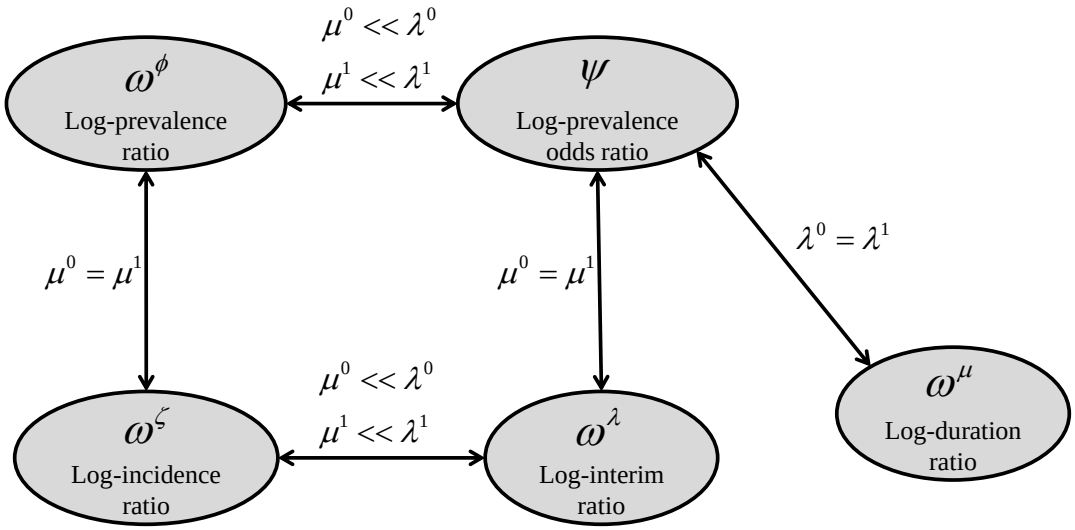


Figure 5.2. Effect sizes for quantifying change in behavior

of lessened intuitive appeal, since researchers and consumers of research may find odds ratios more difficult to interpret than proportionate changes.

5.2.5. Relationships among effect sizes

The five effect sizes that I have described represent different metrics for quantifying change in a behavioral process, as modeled by an ARP. The effect sizes are closely related to one another because they are all defined in terms of ARP parameters. Under certain conditions, the effect sizes also become either approximately or exactly equivalent, in which case it is reasonable to directly compare estimates of those different effect sizes. Understanding the circumstances under which the different effect sizes are measurement-comparable is important for meta-analytic applications, which will often involve combining information across studies that used different outcome measurement procedures.

Figure 5.2 displays the inter-relationships among the effect sizes, with arrows between two effect sizes indicating equality under a given condition. Thus, if event duration is constant across the two points of comparison, $\mu^0 = \mu^1$, then the log-prevalence ratio is equal to the log-incidence ratio and the log-prevalence odds ratio reduces to the log-interim ratio. Similarly, if the average interim time is constant across the points of comparison, $\lambda^0 = \lambda^1$, then the log-prevalence odds ratio is equal to the log-duration ratio. If events have very short average duration relative to the average interim time, and this is true at both points of comparison so that $\mu^0 << \lambda^0$ and $\mu^1 << \lambda^1$, then the log-prevalence ratio will be approximately equal to the log-prevalence odds ratio and the log-incidence ratio will be approximately equal to the log-interim ratio.

5.3. Basic effect size estimators

Thus far, I have described a within-session model for outcome data generated by various measurement procedures, posited a simple between-session model, and defined several different effect size metrics for measuring changes in directly observed behavior. This section presents some basic, easily calculated estimators for these effect sizes, for use with outcome data generated by direct measurement procedures. The following section considers estimators based on interval recording data.

All of the estimators described in this section involve sample means and sample variances calculated within phases. Let n_t denote the number of observations made in treatment condition t, so that

$$n_0 = \sum_{j=1}^{n}(1 - T_j), \qquad n_1 = \sum_{j=1}^{n} T_j.$$

Let $\bar{y}_t^r$ denote the sample mean outcome in treatment condition t, so that

$$\bar{y}_0^r = \frac{1}{n_0}\sum_{j=1}^{n} Y_j^r(1-T_j), \qquad \bar{y}_1^r = \frac{1}{n_1}\sum_{j=1}^{n} Y_j^r T_j.$$

Finally, let s_{rt}^2 denote the sample variance of the outcomes in treatment condition t, so that

$$s_{r0}^2 = \frac{1}{n_0 - 1}\sum_{j=1}^{n_i}\left(Y_j^r - \bar{y}_0^r\right)^2(1-T_j), \qquad s_{r1}^2 = \frac{1}{n_1 - 1}\sum_{j=1}^{n}\left(Y_j^r - \bar{y}_1^r\right)^2 T_j.$$

Several of the effect size estimators described in this section can be viewed as special cases of the log-response ratio, a well-known effect size used for meta-analysis in ecology and other disciplines (Hedges, Gurevitch, & Curtis, 1999). For sample data collected using measurement procedure $r \in \{E, C, M, P, W\}$, define the log-response ratio estimator

$$L^r = \ln\left(\hat{y}_1^r\right) - \ln\left(\hat{y}_0^r\right) \tag{5.7}$$

and variance estimator

$$V_L^r = \frac{s_{r0}^2}{n_0\left(\hat{y}_0^r\right)^2} + \frac{s_{r1}^2}{n_1\left(\hat{y}_1^r\right)^2}, \tag{5.8}$$

where

$$\hat{y}_t^r = \begin{cases} \bar{y}_t^r & \bar{y}_t^r > 0 \\ \hat{y}_t^r = c_t^r & \bar{y}_t^r = 0 \end{cases}$$

for constants c_t^r defined below. Hedges et al. (1999) studied the distribution of the log response ratio under the assumption that the raw data are normally distributed; the exact distribution theory and approximations that they reported are not applicable with the

behavioral observation data under consideration here. Furthermore, in some single-case studies, the within-phase sample sizes n_0, n_1 can be quite small. I therefore consider an alternative estimator, based on a second-degree Taylor series approximation to the bias of L^r:

$$L_2^r = \ln\left(\hat{y}_1^r\right) + \frac{s_{r0}^2}{2n_0\left(\hat{y}_0^r\right)^2}. - \ln\left(\hat{y}_0^r\right) - \frac{s_{r1}^2}{2n_1\left(\hat{y}_1^r\right)^2} \tag{5.9}$$

In the remainder of this section, I describe estimators for the log-incidence ratio, log-prevalence ratio, and log-prevalence odds ratio based on data collected using a direct measurement procedure.

5.3.1. Log-incidence ratio estimators based on event counting

Event counting data measures incidence directly. Assuming that session length is held constant over the duration of the study, the log-response ratio with event counting data, L^E or L_2^E, can therefore be used as estimators for the log-incidence ratio ω^ζ. In simulation studies and the empirical examples discussed in this section, I use the constant $c_t^E = 1/(2n_t)$ to adjust for zero mean outcomes. Based on simulation studies reported in Appendix C.1, the bias-corrected estimator L_2^E should be used in application, particularly for short phase lengths, because it is nearly unbiased and has comparable mean-squared error to the moment estimator L^E.

Example 1. For one case reported by Romaniuk et al. (2002), the investigators used an ABAB design to assess the effect of providing choice between activities (versus a no-choice condition) on the frequency of problem behavior displayed by a child in a kindergarten setting. The child's problem behavior was measured using event counting

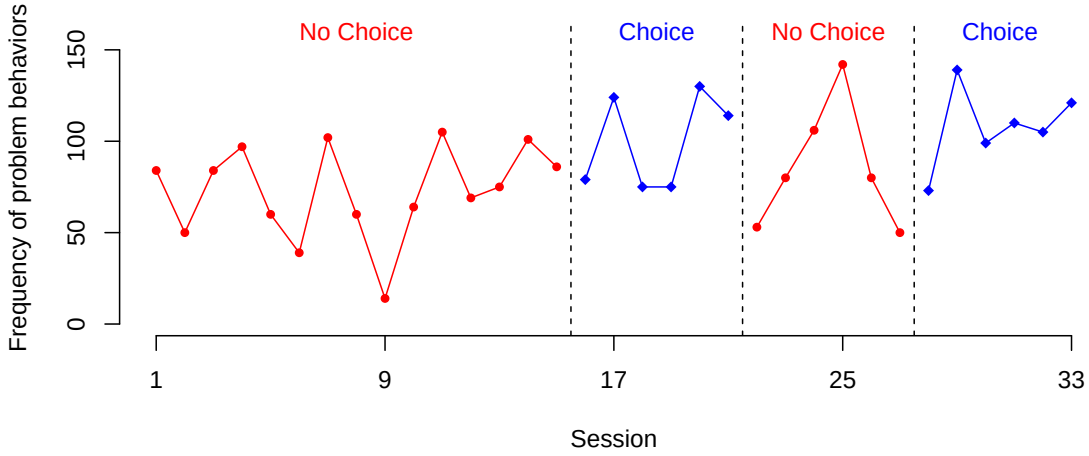


Figure 5.3. Frequency of problem behavior by session for one case from Romaniuk et al. (2002).

with an observation session $L = 5$ minutes in length. Figure 5.3 displays a graph of the data from this case. Based on these data, an estimate of the log-incidence ratio can be calculated under the assumptions given in (5.6). Pooling across both phases in the no-choice (baseline) condition, the mean frequency of problem behavior is $\bar{y}_0^E = 76$ and the sample variance is $s_{E0}^2 = 795$, based on $n_0 = 21$ observations. Pooling across both phases within the choice (treatment) condition, the mean frequency of problem behavior is $\bar{y}_1^E = 104$ and the sample variance is $s_{E1}^2 = 547$, based on $n_1 = 12$ observations. Inserting these summary statistics into (5.9) and (5.8), the estimated log-incidence ratio is $L_2^E = 0.31$ with an approximate standard error of $\sqrt{V_L^E} = 0.10$. This corresponds to an increase in the frequency of problem behavior of $\left[\exp(L_2^E) - 1\right] \times 100\% = 36\%$ when the child is allowed to choose between activities. An approximate 95% confidence interval for the percentage increase in problem behavior can be constructed as $\left[\exp\left(L_2^E \pm 2\sqrt{V_L^E}\right) - 1\right] \times 100\% = [10\%, 67\%]$.

5.3.2. Log-prevalence ratio estimators based on continuous recording or momentary time sampling

Continuous recording and momentary time sampling both produce direct measures of behavioral prevalence. The log-response ratios based on these types of data are therefore estimates of the log-prevalence ratio ω^{ϕ}. For continuous recording data, the constant for adjusting sample mean outcomes of zero should ideally depend on the incidence of the behavior because outcomes of zero are less probable for higher incidence. Given a prior estimate of the incidence ζ, one can use $c_t^C = 1/(2\zeta L n_t)$. For momentary time sampling data, one can use $c_t^M = 1/(2K n_t)$, where K is the number of intervals per observation session. As with event counting data, the bias-corrected estimators L_2^C and L_2^M should be used in application because they have lower bias and approximately equal mean-squared error compared to the moment estimators (see Appendix C.1 for details).

Example 2. For five cases reported by Romaniuk et al. (2002), the investigators measured children's problem behavior using continuous recording with an observation session $L = 5$ minutes in length. Treatment reversal designs with either three or five reversals were used to assess the effect of providing choice between activities (versus a no-choice condition) on the prevalence of problem behavior. Figure 5.4 displays graphs of the data from these cases. A key research question in this study had to do with whether the treatment was differentially effective for children whose problem behavior was maintained by escape versus by attention. Prior assessment (through functional analysis) identified three cases with escape-maintained problem behavior and three cases with attention-maintained problem behavior.

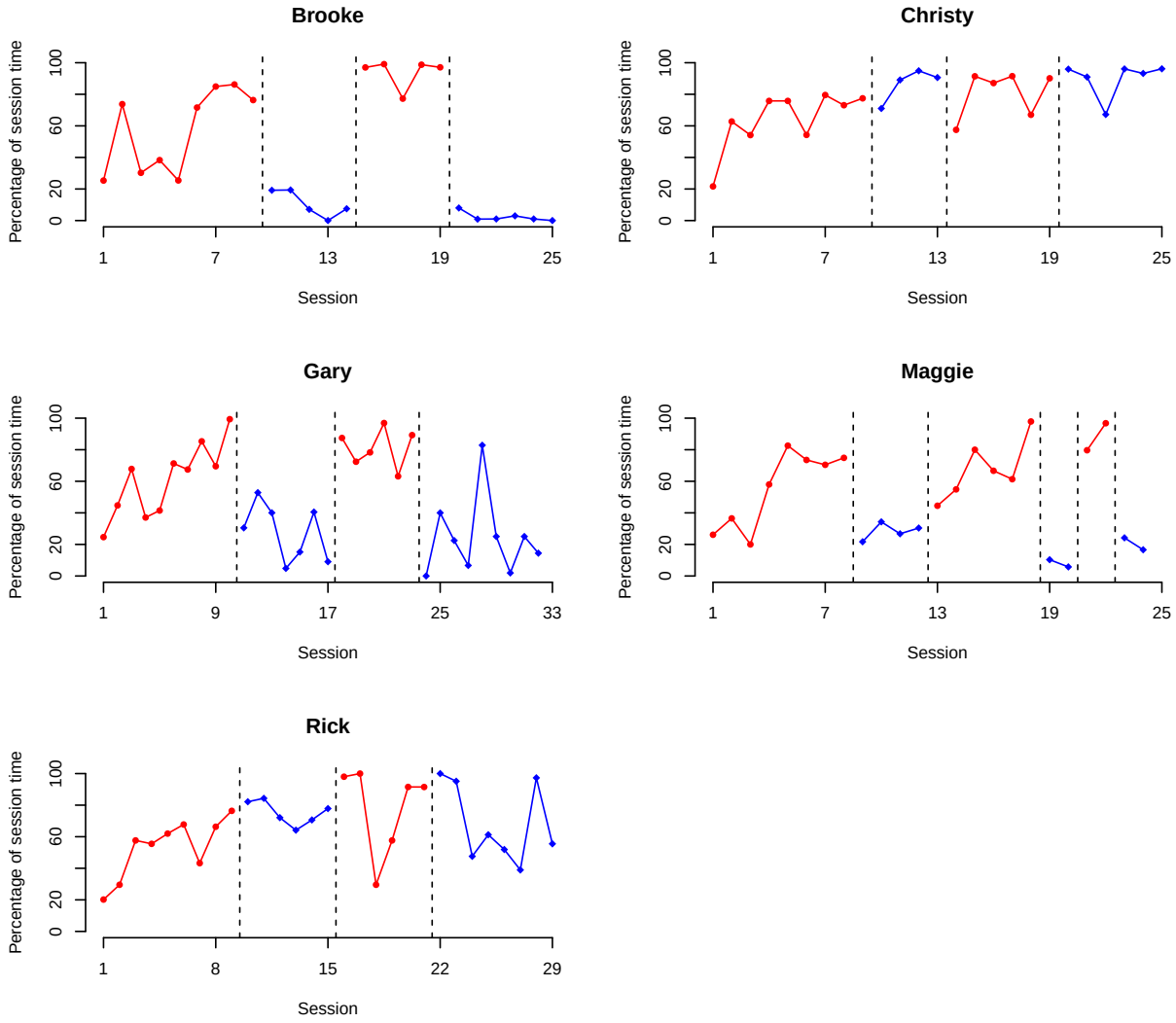


Figure 5.4. Prevalence of problem behavior by session for five cases from Romaniuk et al. (2002). Red circles indicate no-choice condition; blue diamonds indicate choice condition.

Table 5.4 reports summary statistics by treatment condition (pooled across phases) for each of these five cases and also reports estimated log-prevalence ratios L_2^C and standard errors $\sqrt{V_L^C}$, calculated according to (5.9) and (5.8), respectively. For the three cases with escape-maintained behavior, the estimated log-prevalence ratios are large and negative, ranging from -0.96 to -2.39; for these cases, providing choice-making opportunites greatly

reduces the prevalence of problem behaviors. A fixed-effects meta-analysis of the three cases, reported in the penultimate row of the table, provides a succinct summary of findings.[12] The average log-prevalence ratio for cases with escape-maintained behavior is -1.22, with an approximate 95% confidence interval of [-1.48, -0.95] corresponding to a reduction of between 61% and 77%.

The sixth row of Table 5.4 includes summary statistics and effect size estimates for Riley, whose behavior was measured using event counting rather than continuous recording; these estimates were discussed in Example 1. Assuming that the average duration of Riley's problem behaviors was unaffected by the intervention ($\mu^0 = \mu^1$), the calculated log-incidence ratio can be understood as on a comparable scale to the log-prevalence ratios calculated for the cases measured by continuous recording. Under this assumption, the estimated log-prevalence ratios for the three cases with attention-maintained behavior are moderately positive, ranging from 0.12 to 0.31. Based on a fixed-effects meta-analysis (reported in the final row of the table), the average log-prevalence ratio for these cases is 0.23, with an approximate 95% confidence interval of [0.11,0.34] corresponding to increases in problem behavior of 13% to 40%.

5.3.3. Log-prevalence odds ratio estimators based on continuous recording or momentary time sampling

The log-prevalence odds ratio ψ offers an alternative metric for measuring differences or changes in prevalence. Unlike the log-prevalence ratio, this metric can range from negative to positive infinity, regardless of the baseline level of prevalence; it treats a proportionate

[12] I used the `metafor` package in R (Viechtbauer, 2010) for calculation of the fixed-effects estimates.

Table 5.4. Effect size estimates for Romaniuk et al. (2002)

,

Case	Outcome measure	Behavior function	No Choice			Choice						
			$\bar{y}_0$	s_0^2	n_0	$\bar{y}_1$	s_1^2	n_1	L_2	$\sqrt{V^L}$	$\hat{\psi}_2$	$\sqrt{V_\psi}$
Brooke	C	escape	70	792	14	6	52	11	-2.39	0.37	-3.50	0.52
Gary	C	escape	69	479	16	26	480	16	-0.96	0.23	-1.81	0.38
Maggie	C	escape	64	531	16	21	96	8	-1.09	0.19	-1.86	0.32
Christy	C	attention	71	349	15	89	112	10	0.22	0.08	1.13	0.40
Rick	C	attention	63	644	15	71	368	14	0.12	0.13	0.37	0.38
Riley	E	attention	76	795	21	104	547	12	0.31	0.10	0.31	0.10
FE meta-analysis		escape							-1.22	0.13	-2.14	0.22
		attention							0.23	0.06	0.36	0.10

decrease in average event duration as equivalent to the same proportionate increase in average interim time. For continuous recording data or momentary time sampling data, and under the assumptions of the stable phase model from (5.6), a basic moment estimator for ψ is given by

$$\hat{\psi}^r = \text{logit}\left(\tilde{y}_1^r\right) - \text{logit}\left(\tilde{y}_0^r\right), \tag{5.10}$$

where $\text{logit}(x) = \ln(x) - \ln(1-x)$ and

$$\tilde{y}_t^r = \begin{cases} c_t^r & \bar{y}_t^r = 0 \\ \bar{y}_t^r & 0 < \bar{y}_t^r < 1 \\ 1 - c_t^r & \bar{y}_t^r = 1 \end{cases}$$

for $t = 0, 1$ and $r \in \{C, M\}$. The constants for correcting sample means of zero or one are identical to those used with the log-prevalence ratio: for continuous recording $c_t^C = 1/(2\zeta L n_t)$ and for momentary time sampling, $c_t^M = 1/(2K n_t)$. An estimator of the

approximate variance of the effect size estimate is given by

$$V_\psi^r = \frac{s_{r0}^2}{n_0 \left(\tilde{y}_0^r\right)^2 \left(1 - \tilde{y}_0^r\right)^2} + \frac{s_{r1}^2}{n_1 \left(\tilde{y}_1^r\right)^2 \left(1 - \tilde{y}_1^r\right)^2}, \tag{5.11}$$

for $r \in \{C, M\}$. A bias-corrected estimator is given by

$$\hat{\psi}_2^r = \text{logit}\left(\tilde{y}_1^r\right) - \frac{s_{r1}^2(2\tilde{y}_1^r - 1)}{2n_1 \left(\tilde{y}_1^r\right)^2 \left(1 - \tilde{y}_1^r\right)^2} - \text{logit}\left(\tilde{y}_0^r\right) + \frac{s_{r0}^2(2\tilde{y}_0^r - 1)}{2n_0 \left(\tilde{y}_0^r\right)^2 \left(1 - \tilde{y}_0^r\right)^2} \tag{5.12}$$

for $r \in \{C, M\}$. Based on simulation results reported in Appendix C.1, the bias-corrected estimators should be used rather than the moment estimators when the number of observations per phase is small.

Example 2 (continued). Table 5.4 reports log-prevalence odds ratios estimates $\hat{\psi}_2^C$ and associated standard errors $\sqrt{V_\psi^C}$ for the five cases from Romaniuk et al. (2002) that were measured using continuous recording. For four of the five cases, the relative magnitudes of the log-prevalence odds ratio estimates are comparable to the log-prevalence estimates. The exception is Christy, whose log-prevalence odds ratio is much larger than the other two cases with attention-maintained problem behavior, even while her log-prevalence ratio is intermediate between those of the other two cases. This discrepancy is due to the fact that the two effect size metrics are less comparable for $\phi > 0.5$. Christy's average prevalence is fairly high in both the no-choice ($\bar{y}_0 = 71\%$) and the choice conditions ($\bar{y}_1 = 89\%$), leading to a divergence between the magnitude of the two metrics.

Table 5.4 also includes an effect size estimate for the one case measured using event counting (Riley); this effect size is comparable to the estimated log-prevalence odds ratios under the assumptions that the treatment does not alter average event duration ($\mu^0 =$

μ^1) and that the average event durations are very short.[13] Based on fixed-effects meta-analyses, cases with escape-maintained problem behavior had an average log-prevalence odds ratio of -2.14 (approximate 95% confidence interval: $[-2.58, -1.71]$) and cases with attention-maintained problem behavior had an average log-prevalence odds ratio of 0.36 (approximate 95% confidence interval: $[0.17, 0.55]$). In this example, the log-prevalence odds ratio and the log-prevalence ratio lead to much the same substantive conclusions and display similar levels of residual heterogeneity. Pooling across levels of the moderator and retaining the fixed-effect specification, the Q statistic based on log-prevalence odds ratio estimates is 12.23, only slightly smaller than the comparable Q statistic of 12.90 based on log-prevalence ratio estimates. It is therefore very difficult to determine which metric is preferable on an empirical basis alone.

5.4. Estimators based on interval recording

Interval recording data measures of neither prevalence nor incidence. Consequently, the response ratio based on interval recording data estimates no directly interpretable parameter. For instance, consider partial interval recording and denote

$$\pi_t^P = \mathrm{E}\left(\bar{y}_t^P\right) = \phi^t + \zeta^t \int_0^{l^-} \tilde{F}_E(t; \lambda^t) dt.$$

To a very close approximation, the bias-corrected response ratio estimator based on partial interval recording has expectation $E\left(L_2^P\right) \approx \ln\left(\pi_1^P\right) - \ln\left(\pi_0^P\right)$ (see Appendix C.1). Still, interval recording data can provide some information about interpretable effect size parameters under certain assumptions about the behavior being measured. In this section, I

[13]The investigators noted that this case displayed short-duration problem behaviors, and in fact this was the motivation for using event counting rather than continuous recording (Romaniuk et al., 2002, p. 351).

consider several strategies for estimating bounds on the log-incidence ratio, log-prevalence ratio, log-prevalence odds ratio, and log-interim ratio, based on different identifying assumptions. I focus on partial interval recording because it is much more commonly used; formally similar assumptions and analysis strategies can be applied for whole interval recording.

5.4.1. Log-incidence ratio

Suppose that average event durations in each treatment condition are shorter than some known value $\mu_U^* \geq \mu^0, \mu^1$ established based on prior experience. Also suppose that the interim time between behavioral events is rarely less than the active interval length, so that $F_E(l|\lambda^t) < p^* < 1$ for known, small p^*, $t = 0, 1$. Letting

$$z^\zeta = \ln(\mu_U^* + l) - \ln(1 - p^*) - \ln(l),$$

it follows from (5.5) that

$$\ln\left(\pi_1^P\right) - \ln\left(\pi_0^P\right) - z^\zeta \leq \omega^\zeta \leq \ln\left(\pi_1^P\right) - \ln\left(\pi_0^P\right) + z^\zeta. \tag{5.13}$$

Thus, bounds for the incidence ratio can be estimated from partial interval recording data by $L_2^P \pm z^\zeta$, because L_2^P is an unbiased estimator for $\ln\left(\pi_1^P\right) - \ln\left(\pi_0^P\right)$. Note that the variance of these bounds estimators can be estimated by V_L^P because z^ζ is a fixed quantity.

Example 3. Dunlap et al. (1994) used a treatment reversal design to evaluate the effect of providing choice between academic activities on the disruptive behavior of three elementary school students with emotional and behavioral disorders. The investigators

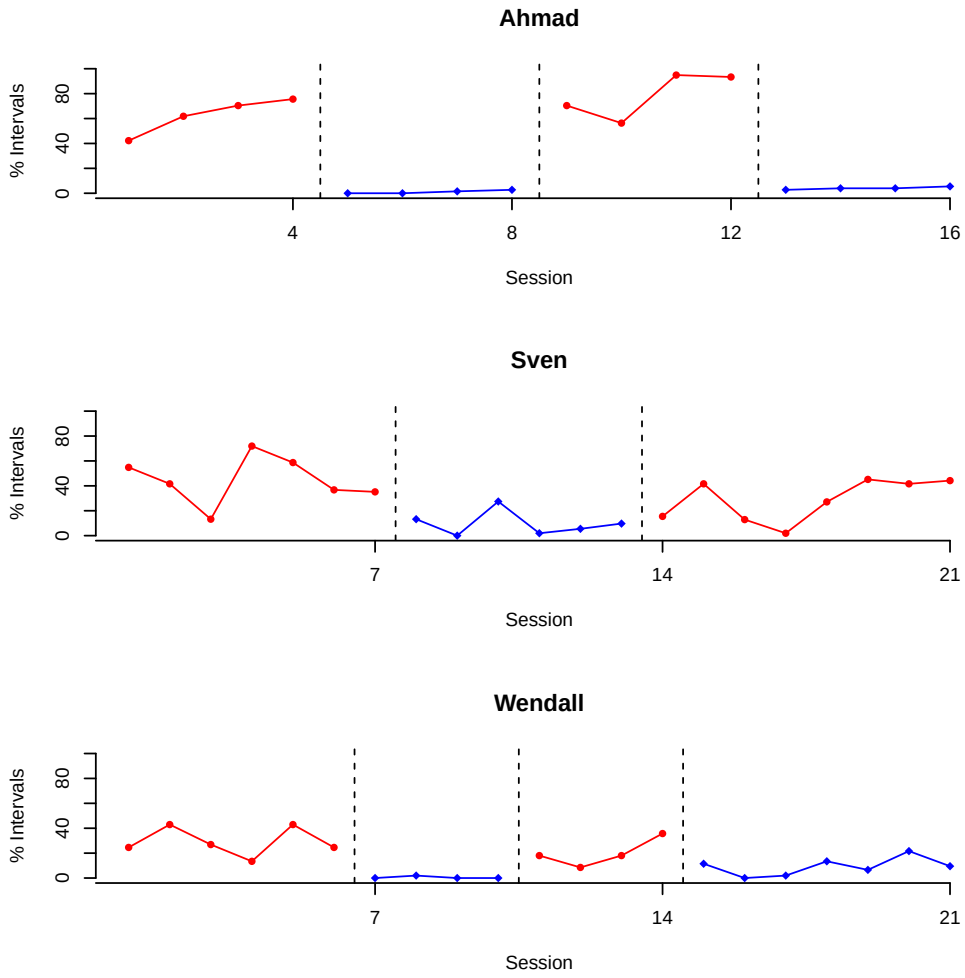


Figure 5.5. Percentage of partial intervals with disruptive behavior for three cases from Dunlap et al. (1994). Red circles indicate no-choice condition; blue diamonds indicate choice condition.

used partial interval recording to measure disruptive behavior; for two cases (Sven and Ahmad), measurements were based on an active interval length of $l = 10$ seconds and 5 seconds for recording, while for the third case (Wendell), measurements were based on an active interval length of $l = 15$ seconds with no time for recording. Observation sessions lasted $L = 15$ minutes, implying that each reported datum was based on $K = 60$ intervals. Figure 5.5 plots the partial interval data for each of the three cases.

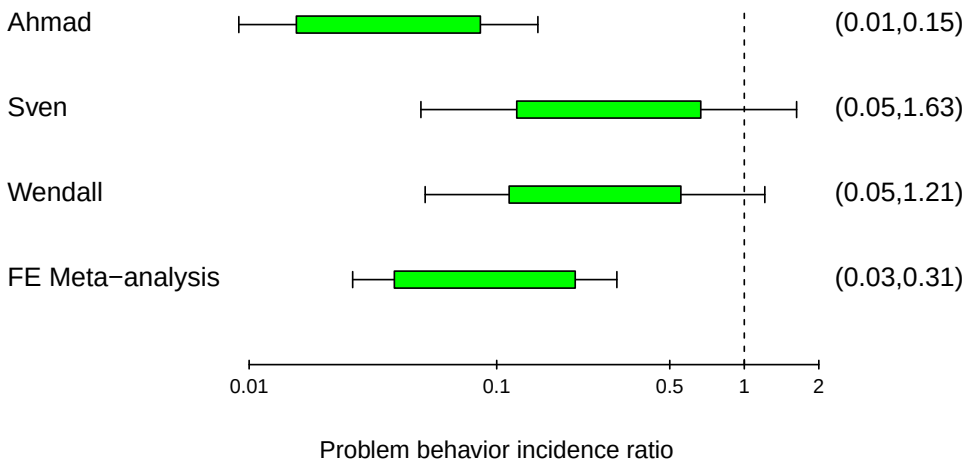


Figure 5.6. Forest plot of bounds for the incidence ratio, based on Dunlap et al. (1994) data.

Based on descriptions of the types of disruptive behaviors exhibited by the three students, it may be reasonable to assume that the average duration of each behavioral event was fairly short, and therefore that primary interest is in the incidence of disruptive behavior. To bound the log-incidence ratio, I assume that $\mu^* < 10$ seconds. Since the active interval length varies across cases, I make separate assumptions about p^* for each case: for Sven and Ahmad I assume that at most $p^* = 15\%$ of interim times are less than $l = 10$ seconds, implying that $z^\zeta = 0.86$; for Wendell, I assume that at most $p^* = 25\%$ of interim times are less than $l = 15$ seconds, implying that $z^\zeta = 0.80$. Based on these assumptions, I estimate bounds for the log-incidence ratio using $L_2^P \pm z^\zeta$. I calculate an approximate 95% confidence interval for the bounds using $L_2^P \pm \left(z^\zeta + 2\sqrt{V_L^P}\right)$.

Figure 5.6 depicts the estimated incidence ratio bounds and approximate confidence intervals for each case in the study, using a graphic known as a forest plot; note that the horizontal axis is on a log-scale.[14] The estimated bounds are below 1 for all three cases,

[14]Forest plots are a commonly used graphical tool for representing effect sizes in meta-analysis. Figure 5.6 uses a modified version of the conventional forest plot to represent bounds estimates, rather than

suggesting that the treatment may have reduced the incidence of disruptive behavior. However, accounting for both sampling uncertainty and the identification-related uncertainty, it is possible that the treatment may have had zero effect on Sven and Ahmad's problem behavior. The final row of the forest plot reports a fixed-effects meta-analysis of the bounds, which yields an estimate of the bounds on the mean incidence ratio across the cases in the study. Given that the 95% confidence interval for the average incidence ratio is (0.03, 0.31), it is possible to conclude that, on average, the treatment reduces the incidence of disruptive behavior by more than 69%, and possibly as much as 97%. Even based on rather conservative assumptions regarding the average behavioral event duration and probability of short interim times, it seems reasonable to conclude that this treatment is very effective for these three cases.

5.4.2. Log-prevalence and log-prevalence odds ratios

Suppose that both lower and upper bounds on the average event durations can be established, $0 < \mu_L^* \leq \mu^t \leq \mu_U^*$, and that $F_E(l|\lambda^t) < p^*$ for known p^*, $t = 0, 1$. Let

$$z^\phi = \begin{cases} \ln(\mu_L^* + l) - \ln(\mu_L^*) + \ln(\mu_U^*) - \ln(\mu_U^* + (1-p^*)l) & \mu_U^* < \infty \\ \ln(\mu_L^* + l) - \ln(\mu_L^*) & \mu_U^* = \infty. \end{cases}$$

It follows from (5.4) that estimable bounds on the log-prevalence ratio are given by

$$\ln\left(\pi_1^P\right) - \ln\left(\pi_0^P\right) - z^\phi \leq \omega^\phi \leq \ln\left(\pi_1^P\right) - \ln\left(\pi_0^P\right) + z^\phi. \tag{5.14}$$

point estimates. Here, green rectangles represent the estimated bounds for the prevalence ratio, while whisker bars represent approximate 95% confidence intervals for the bounds; the latter are also reported in the right margin of the figure. In conventional forest plots, the marker representing the point estimate is often drawn so that its area is proportional to its precision, which has the effect of drawing attention to more precisely estimated effect sizes. The modified forest plots that I present do not follow this practice.

These bounds can be estimated from partial interval recording data by $L_2^P \pm z^\phi$, with approximate variance V_L^P. A parallel argument leads to bounds on the log-prevalence odds ratio:

$$\begin{aligned}(5.15)\quad \ln\left(\pi_1^P\right) - \ln\left[1 - \pi_1^P + \frac{l}{\mu_L^*}\right] - \ln\left(\pi_0^P\right) + \ln\left[1 - \pi_0^P + \frac{(1-p^*)l}{\mu_U^*}\right] &\leq \psi \\ &\leq \ln\left(\pi_1^P\right) - \ln\left[1 - \pi_1^P + \frac{(1-p^*)l}{\mu_U^*}\right] - \ln\left(\pi_0^P\right) + \ln\left[1 - \pi_0^P + \frac{l}{\mu_L^*}\right].\end{aligned}$$

Moment estimators for the lower and upper bounds in (5.15) are given by

$$\begin{aligned}(5.16)\quad \hat{\psi}_L^P &= \ln\left(\bar{y}_1^P\right) - \ln\left[1 - \bar{y}_1^P + \frac{l}{\mu_L^*}\right] - \ln\left(\bar{y}_0^P\right) + \ln\left[1 - \bar{y}_0^P + \frac{(1-p^*)l}{\mu_U^*}\right] \\ \hat{\psi}_U^P &= \ln\left(\bar{y}_1^P\right) - \ln\left[1 - \bar{y}_1^P + \frac{(1-p^*)l}{\mu_U^*}\right] - \ln\left(\bar{y}_0^P\right) + \ln\left[1 - \bar{y}_0^P + \frac{l}{\mu_L^*}\right]\end{aligned}$$

with corresponding variance estimates

$$\begin{aligned}(5.17)\quad V_{\psi L} &= \frac{s_{P0}^2\left(1 + \frac{(1-p^*)l}{\mu_U^*}\right)^2}{n_0\left(\bar{y}_0^P\right)^2\left(1 - \bar{y}_0^P + \frac{(1-p^*)l}{\mu_U^*}\right)^2} + \frac{s_{P1}^2\left(1 + \frac{l}{\mu_L^*}\right)^2}{n_1\left(\bar{y}_1^P\right)^2\left(1 - \bar{y}_1^P + \frac{l}{\mu_L^*}\right)^2} \\ V_{\psi U} &= \frac{s_{P0}^2\left(1 + \frac{l}{\mu_L^*}\right)^2}{n_0\left(\bar{y}_0^P\right)^2\left(1 - \bar{y}_0^P + \frac{l}{\mu_L^*}\right)^2} + \frac{s_{P1}^2\left(1 + \frac{(1-p^*)l}{\mu_U^*}\right)^2}{n_1\left(\bar{y}_1^P\right)^2\left(1 - \bar{y}_1^P + \frac{(1-p^*)l}{\mu_U^*}\right)^2}.\end{aligned}$$

Example 4. Moes (1998) used a four-phase treatment reversal design to evaluate the effect of providing choice-making opportunities on the disruptive behavior of four children with autism, in the context of homework tutoring sessions. The investigators measured disruptive behavior using partial interval recording with $l = 10$ second active intervals, 5 seconds for recording, and $K = 80$ intervals per observation session. Each case was measured for a total of $n_0 = 10$ sessions in the no-choice condition and $n_1 = 10$

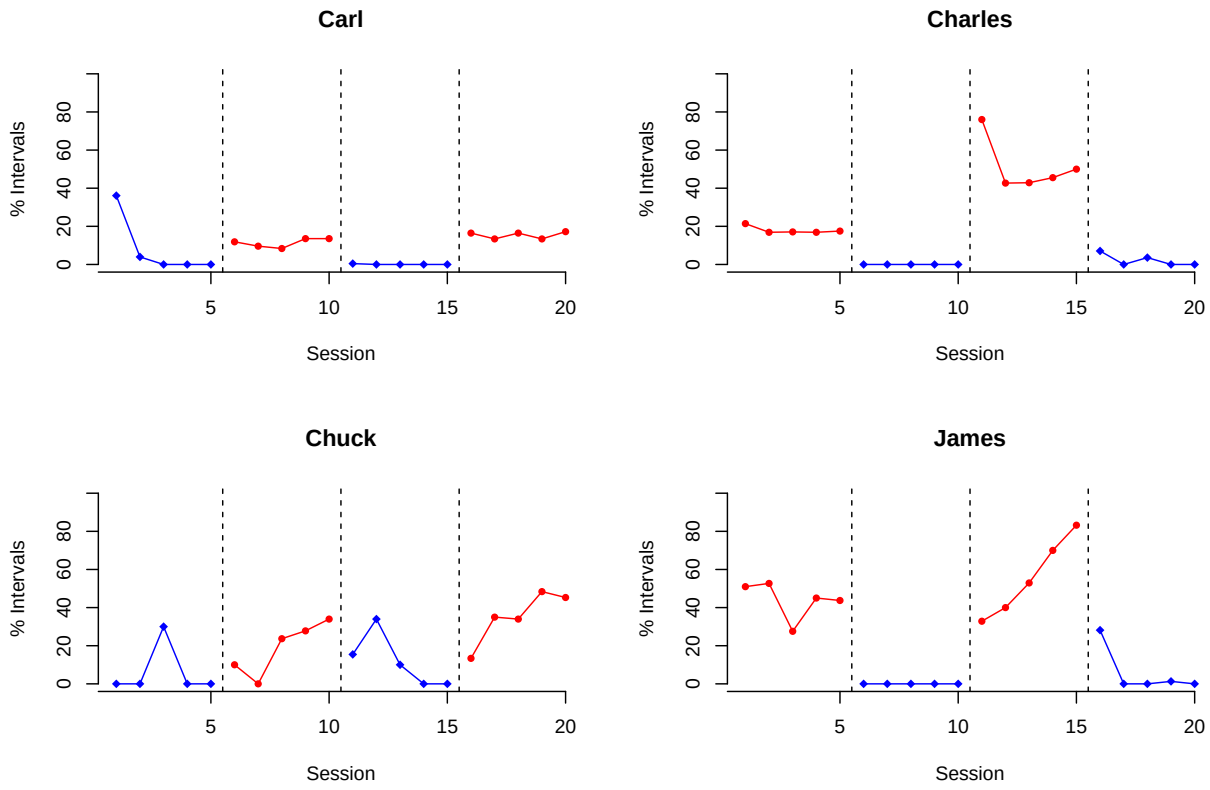


Figure 5.7. Percentage of partial intervals with disruptive behavior for four cases from Moes (1998). Red circles indicate no-choice condition; blue diamonds indicate choice condition.

sessions in the choice condition; each condition was introduced across two phases, using a randomized ABAB/BABA design. Figure 5.7 displays the data from this study.

Suppose that, based on experience with the types of disruptive behaviors exhibited by the study participants, the average length of disruptive behaviors can be established as greater than $\mu_L^* = 10$ seconds and less than $\mu_U^* = 60$ seconds. Also suppose that at most $p^* = 20\%$ of interim times are less than 10 seconds. It then follows from (5.14) that the log-prevalence ratio is within $z^\phi = 0.57$ of the response ratio based on partial interval recording means.[15] Table 5.5 reports estimated bounds on the log-prevalence ratio for

[15]If only a lower bound on mean event duration can be established (i.e., $\mu_U^* = \infty$), then the half-width of the bound increases to $z^\phi = 0.69$.

Table 5.5. Estimated effect size bounds for Moes (1998)

,

	Log-prevalence ratio ω^{ϕ}		Log-prevalence odds ratio ψ		Log-interim ratio ω^{λ}	
Case	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Carl	(-1.39,-0.25)	(-3.20,1.55)	(-1.90,-0.69)	(-3.75,1.18)	(-0.93,-0.87)	(-2.81,0.97)
Charles	(-3.76,-2.63)	(-5.23,-1.16)	(-4.36,-3.04)	(-5.88,-1.54)	(-3.60,-3.39)	(-5.15,-1.89)
Chuck	(-1.59,-0.45)	(-2.60,0.56)	(-1.91,-0.61)	(-3.01,0.49)	(-1.24,-1.13)	(-2.39,-0.05)
James	(-2.93,-1.79)	(-4.85,0.13)	(-3.94,-2.50)	(-5.92,-0.52)	(-3.02,-2.67)	(-5.03,-0.71)
FE meta-analysis	(-2.24,-1.10)	(-2.93,-0.41)	(-2.78,-1.49)	(-3.51,-0.76)	(-2.03,-1.84)	(-2.78,-1.13)

each of the four cases, along with approximate 95% confidence intervals (CI) calculated as $L_2^P \pm \left(z^{\phi} + 2\sqrt{V_L^P}\right)$. The final row of the table reports fixed-effects meta-analyses based on the end-points of the bounds. The average log-prevalence across the four cases is estimated to be between -2.24 and -1.10, equivalent to a 67-89% reduction in disruptive behavior. Based on the confidence interval of (-2.93,-0.41) for the average-log prevalence, which accounts for sampling uncertainty in the bounds estimates, the treatment leads to a reduction in disruptive behavior of 33-95%. While it is apparent from this analysis that the treatment is beneficial, considerable uncertainty remains about the magnitude of the average effect.

The third and fourth columns of Table 5.5 reports estimated bounds $\left(\hat{\psi}_L^P, \hat{\psi}_U^P\right)$ and corresponding confidence intervals for the log-prevalence odds ratio, based on the Moes (1998) data. Approximate 95% confidence intervals are calculated as $\left(\hat{\psi}_L^P - 2\sqrt{V_{\psi L}}, \hat{\psi}_U^P + 2\sqrt{V_{\psi U}}\right)$. Based on a fixed-effects meta-analysis, and accounting for the uncertainty in the bounds estimates, the treatment leads an average log-prevalence odds of between -3.51 and -0.76. The practical implications are much the same as with the other metric: the treatment leads to reductions in disruptive behavior, but there remains uncertainty about the magnitude of the average reduction.

5.4.3. Log-interim ratio

The first two bounding approaches that I have described make assumptions only about the average event duration and the probability of short interim times, but not about the full distribution of event durations or interim times. Alternately, one could entertain a stronger set of distributional assumptions about the behavior stream that will yield narrower identification regions. For instance, S. A. Altmann and Wagner (1970) proposed analyzing partial interval recording data under the assumptions that event durations are negligible and that interim times follow an exponential distribution (i.e., a Poisson point process); they showed that these assumptions point-identify the average interim time, though estimators based on this model are quite sensitive to violations of the assumptions (Fienberg, 1972). I now consider a related but somewhat more general set of assumptions, which yield relatively narrow bounds for the log-interim ratio ω^{λ}.

Suppose that the intervention does not change the average event duration, so that $\mu^0 = \mu^1$. Further suppose that the interim times in each treatment condition follow exponential distributions, so that $F_E\left(t|\lambda^t\right) = 1 - \exp\left(-t/\lambda^t\right)$, $t = 0, 1$. If $\pi_0^P > \pi_1^P$, it then follows that

$$\text{logit}\left(\pi_1^P\right) - \text{logit}\left(\pi_0^P\right) < \omega^{\lambda} < \text{cll}\left(\pi_1^P\right) - \text{cll}\left(\pi_0^P\right), \tag{5.18}$$

where cll() denotes the complementary-log-log function, $\text{cll}(x) = \ln\left[-\ln(1-x)\right]$; similarly, if $\pi_0^P \leq \pi_1^P$, then

$$\text{cll}\left(\pi_1^P\right) - \text{cll}\left(\pi_0^P\right) \leq \omega^{\lambda} \leq \text{logit}\left(\pi_1^P\right) - \text{logit}\left(\pi_0^P\right). \tag{5.19}$$

Proof is given in Appendix B.3. The log-odds ratio, $\text{logit}\left(\pi_1^P\right) - \text{logit}\left(\pi_0^P\right)$, can be estimated using the moment estimator $\hat{\psi}^P$ as given in (5.10) or the bias-corrected estimator $\hat{\psi}_2^P$ as given in (5.12); its approximate variance is given by (5.11) with $p = P$. The complementary-log-log ratio can be estimated using the moment estimator

$$\hat{\kappa}^P = \text{cll}\left(\tilde{y}_1^P\right) - \text{cll}\left(\tilde{y}_0^P\right) \tag{5.20}$$

or the bias-corrected estimator

$$\hat{\kappa}_2^P = \text{cll}\left(\tilde{y}_1^P\right) + \frac{s_{P1}^2\left[\ln\left(1-\tilde{y}_1^P\right)+1\right]}{2n_1\left(1-\tilde{y}_1^P\right)^2\left[\ln\left(1-\tilde{y}_1^P\right)\right]^2} - \text{cll}\left(\tilde{y}_0^P\right) - \frac{s_{P0}^2\left[\ln\left(1-\tilde{y}_0^P\right)+1\right]}{2n_0\left(1-\tilde{y}_0^P\right)^2\left[\ln\left(1-\tilde{y}_0^P\right)\right]^2}. \tag{5.21}$$

The variance of either estimator is approximately

$$V_\kappa^P = \frac{s_{P0}^2}{n_0\left(1-\tilde{y}_0^P\right)^2\left[\ln\left(1-\tilde{y}_0^P\right)\right]^2} + \frac{s_{P1}^2}{n_1\left(1-\tilde{y}_1^P\right)^2\left[\ln\left(1-\tilde{y}_1^P\right)\right]^2}. \tag{5.22}$$

Example 4 (continued). Consider again the study by Moes (1998), which used 10-second partial interval recording. The final two columns of Table 5.5 reports estimated bounds and 95% confidence intervals for the log-interim ratio, under the assumptions that the treatment does not affect the average duration of disruptive behaviors and that the interim times are exponentially distributed. Under the assumption that $\mu^0 = \mu^1$, the log-interim ratio is directly comparable to the log-prevalence odds ratio. Comparing the confidence intervals for the average effect size as estimated using fixed-effects meta-analysis, it can be seen that the estimated bounds for the log-interim ratio are considerably narrower than those for the log-prevalence odds ratio: the former have a width of only

0.18, compared to a width of 1.29 for the latter. However, this gain in precision comes entirely from reliance on additional distributional assumptions.

5.5. Application: Shogren, Faggella-Luby, Bae, & Wehmeyer (2004)

The examples presented in previous sections are all based on studies included in a systematic review by Shogren et al. (2004). This section presents a complete meta-analysis of the studies included in the review, in order to illustrate the extent to which conclusions are sensitive to identifying assumptions. I examine four sets of successively stronger models, with the goal of summarizing the average effect of providing choice-making opportunities to individuals who display problem behavior. In all four models, I focus on prevalence as the most practically relevant dimension of problem behavior.

Recall from Table 5.1 that the studies used a variety of measurement procedures, including event counting, continuous recording, momentary time sampling, interval recording, and other procedures. Effect sizes based on each procedure involve distinct assumptions, which I detail separately. I exclude from the analysis the three studies (including four cases) that used idiosyncratic measurement procedures (Bambara et al., 1995; Cole & Levinson, 2002; Dibley & Lim, 1999), from which measurement-comparable effect sizes cannot be derived. I also excluded one study with a single case (Peterson et al., 2001) that reported a functional assessment but did not use an evaluation design. Of the cases measured using interval recording, only one (Kern et al., 2001, "Kelly") was measured with whole interval recording. With this case, a whole interval recording procedure was used to measure task engagement. For purposes of calculating effect sizes, I re-coded the data as a partial interval recording measure of task dis-engagement.

5.5.1. Effect size assumptions

Model 1 uses the log-prevalence ratio to quantify changes in prevalence. For cases measured using continuous recording and momentary time sampling, effect size estimates were calculated using L_2^C or L_2^M, as given in (5.9). For cases measured using event counting, I assume that the intervention does not alter the mean event duration ($\mu^0 = \mu^1$), so that the log-response ratio, estimated by L_2^E, is equivalent to the log-prevalence ratio. Finally, for cases measured using interval recording, I assume that the average event duration in each phase is greater than $\mu_L^* = 10$ seconds, and estimate bounds for the log-prevalence ratio based on (5.14).

Model 2 incorporates all of the assumptions of Model 1, while imposing an additional assumption for cases measured using interval recording. For these cases, active interval lengths ranged from $l = 10$ seconds to 30 seconds. I assume that the average event duration in each phase is less than $\mu_U^* = 60$ seconds and that the probabilities of interim times less than 10, 15, or 30 seconds are 15%, 25%, and 50%, respectively. These additional assumptions leads to narrower bounds for the log-prevalence ratio.

Model 3 uses the log-prevalence odds ratio to quantify changes in prevalence, while retaining all of the assumptions of Model 2. An additional assumption is needed in order for the effect sizes estimated using event counting to be directly comparable to log-prevalence odds ratio estimates. Namely, I assume that the average event duration in each treatment condition is close to zero, so that proportionate changes in incidence are approximately equal to proportionate changes in interim times. Thus, effect sizes for cases measured by event counting are estimated using L_2^E. For cases measured using continuous recording or momentary time sampling, the log-prevalence odds ratio is estimated using

$\hat{\psi}_2^C$ or $\hat{\psi}_2^M$, as given in (5.12). For cases measured using interval recording, effect size bounds are estimated using (5.16).

Model 4 also uses the log-prevalence odds ratio, but makes different assumptions for cases measured by interval recording. Rather than assuming that the average event durations lie in a particular interval, I assume instead that the average event duration is not affected by intervention, so that $\mu^0 = \mu^1$; this assumption implies that the interim ratio is equivalent to the log-prevalence odds ratio. I also assume that interim times are exponentially distributed. Together, these assumptions lead to bounds for the log-prevalence odds ratio, which I estimate as described in Section 5.4.3. For event counting, continuous recording, and momentary time sampling, log-prevalence odds ratio estimates are formed as in the third model.

5.5.2. Meta-analytic model

Having described the assumptions behind how I calculate effect size estimates, I now describe the modeling assumptions I use to meta-analyze those effect sizes. Typical meta-analytic methods deal with point estimates of effect sizes, rather than bounds estimates. I follow an approach of separately meta-analyzing the lower and upper effect size bounds. For the few cases and studies where direct measures were used, point estimates of effect size are treated as bounds of zero width, and so included in both the lower- and upper-bounds models.[16] Let Z_{ij}^L, Z_{ij}^U denote estimated lower and upper bounds on the effect size

[16]Another approach to synthesizing a collection of studies with some point estimates (from direct measurement procedures) and some bound estimates (from interval recording procedures) would be to separately meta-analyze the point estimates and the bounds estimates. I do not implement this strategy here due to the limited number of studies with cases measured using methods other than interval recording. Only four independent studies used a direct measurement procedure, and in only one of these were multiple cases available; estimation of multiple variance components would be very tenuous with this configuration.

for case j from study i; for point estimates, $Z_{ij}^L = Z_{ij}^U$. Let V_{ij}^b, $b \in \{L, U\}$ be the estimated sampling variance of these estimates. Following Van den Noortgate and Onghena (2008), I assume that that the estimates follow a multi-level model in which

$$Z_{ij}^b = \gamma^b + u_i + \nu_{ij} + \epsilon_{ij}, \tag{5.23}$$

for $b \in \{L, U\}$, where γ^b is the average effect size bound across cases and studies, u_i is a study-level deviation from the average bound, ν_{ij} is a case-level deviation from the bound for study i, and ϵ_{ij} is the sampling error of the estimated effect size bound for case j in study i. I assume that all errors are independently distributed, with $u_i \sim N(0, \tau_b^2)$, $\nu_{ij} \sim N(0, \sigma_b^2)$, and $\epsilon_{ij} \sim N\left(0, V_{ij}^b\right)$.

I estimate the variance components τ_b^2, σ_b^2 and the average effect size γ^b via restricted maximum likelihood. Denoting the estimated bound for the average effect size as $\hat{\gamma}^b$, the corresponding model-based variance estimate as V^b, and the upper 0.025 critical value from a t-distribution with 8 degrees of freedom as t_8, I construct confidence intervals for the average effect size bounds by $\left(\hat{\gamma}^L - t_8\sqrt{V^L}, \hat{\gamma}^U + t_8\sqrt{V^U}\right)$.[17]

This modeling approach leads to estimated bounds for the average effect size $(\hat{\gamma}^L, \hat{\gamma}^U)$. However, it is important to note that the estimates of within- and between-study heterogeneity $\hat{\tau}_b^2, \hat{\sigma}_b^2$ cannot be interpreted as bounds on the true heterogeneity. Instead, estimated variance components serve only as rough indicators of heterogeneity. I know of

[17] In multi-level meta-analysis, it is prudent to examine cluster-robust variance estimates in addition to the model-based variance estimates. The former are asymptotically consistent (as the number of independent studies increases) even if the model for variance components is mis-specified (Hedges, Tipton, & Johnson, 2010). In the present example, the cluster-robust standard errors are nearly identical to the model-based standard errors, though slightly smaller.

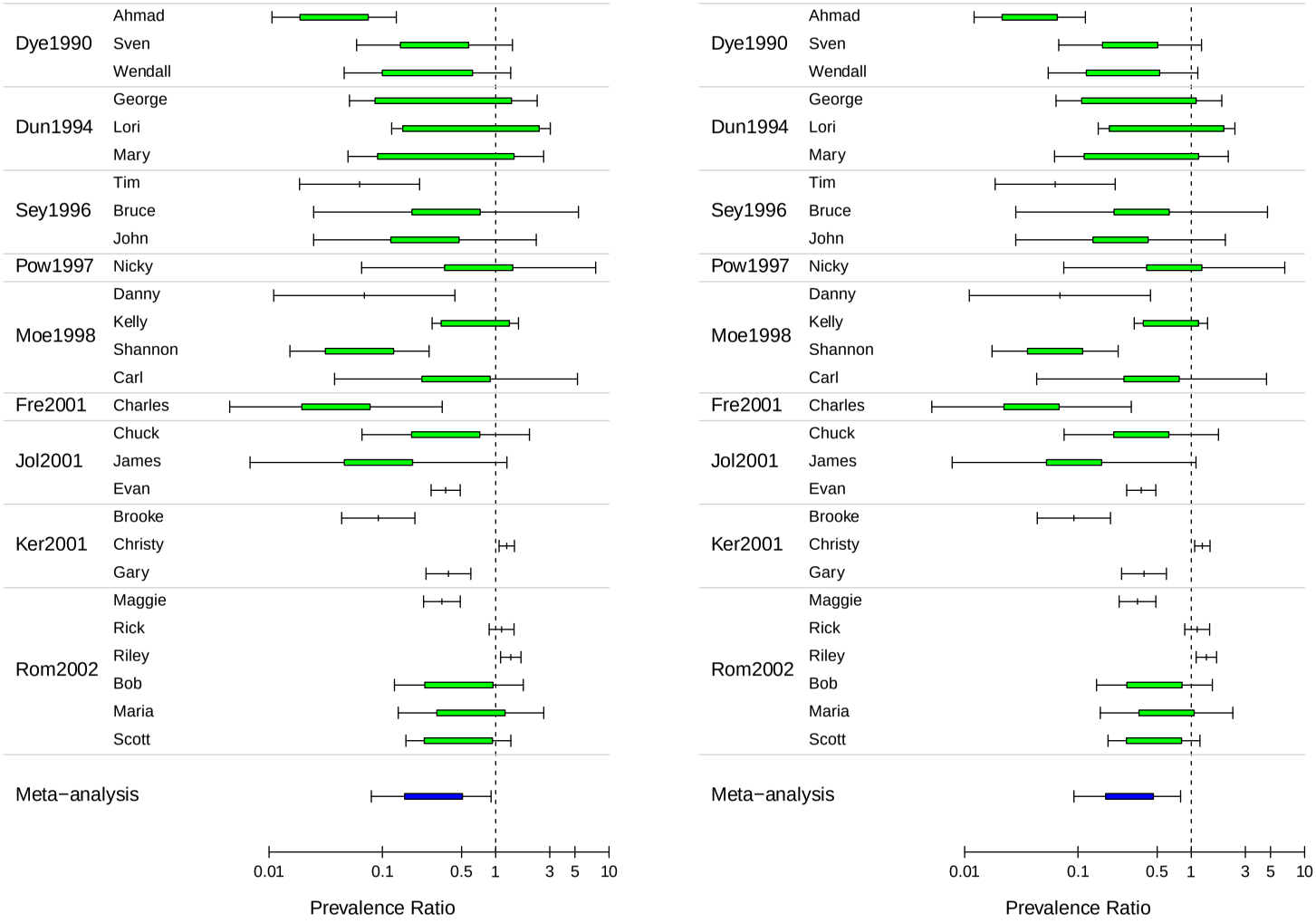
no method for determining bounds on variance components in this context; a more rigorous analysis of variance components based on estimated effect size bounds will require further methodological development.

5.5.3. Results

Figures 5.8a through 5.9b display forest plots of the estimated effect size bounds under Models 1 through 4, respectively.[18] Table 5.6 reports estimated bounds and 95% confidence intervals for the average effect size under each model, along with estimates of between-study heterogeneity $\hat{\tau}^2$ and within-study heterogeneity $\hat{\sigma}^2$.

Models 1 and 2 use the log-prevalence ratio as the effect size metric. Based on the assumptions of Model 1, the average log-prevalence ratio is estimated to be between -1.85 and -0.67, corresponding to a reduction in the prevalence of problem behavior of between 49% and 84%; the 95% confidence interval of -2.53 to -0.09 corresponds to a reduction in prevalence of between 9% and 92%. The range of the estimated bounds is wide due to the large number of cases that were measured using partial interval recording. For many of the cases measured using partial interval recording, the estimated bounds on the individual prevalence ratio are very wide; note in particular that for 6 cases, the estimated bounds includes a prevalence ratio of one, which corresponds to no intervention effect, while for an additional 8 cases, the confidence interval includes unity. Based on the upper bounds of the individual effect size estimates, the total heterogeneity of individual effect sizes (including variation both within and between studies) is estimated to be $\hat{\tau}_U^2 + \hat{\sigma}_U^2 = 1.04$,

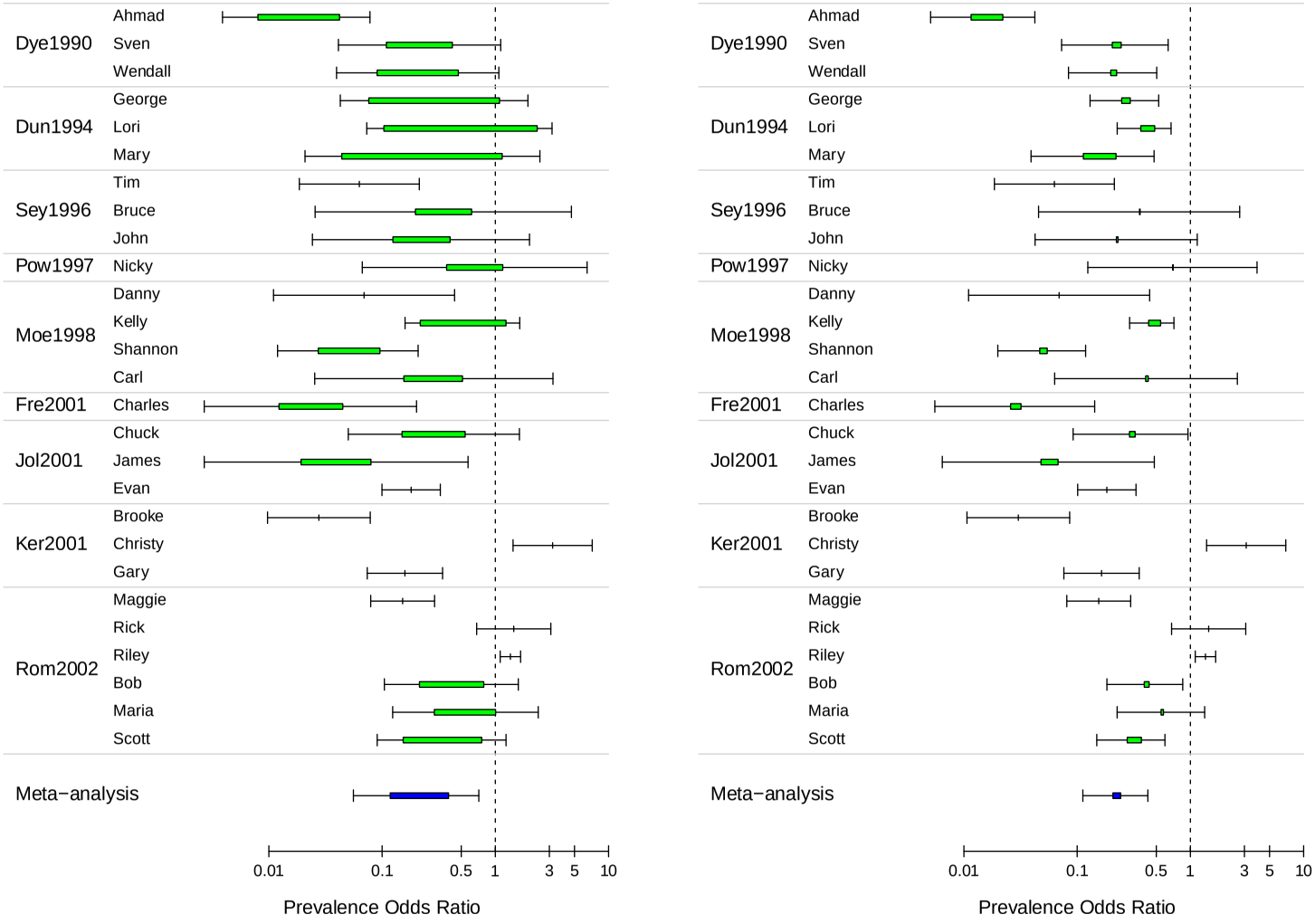
[18] See footnote 14 on the construction of these modified forest plots.



(a) Model 1

(b) Model 2

Figure 5.8. Forest plot of estimated prevalence ratio bounds for studies from Shogren et al. (2004). Green rectangles represent estimated bounds; whisker bars represent approximate 95% confidence intervals for the bounds.



(a) Model 3

(b) Model 4

Figure 5.9. Forest plot of estimated prevalence odds ratio bounds for studies from Shogren et al. (2004). Green rectangles represent estimated bounds; whisker bars represent approximate 95% confidence intervals for the bounds.

Table 5.6. Meta-analysis of studies from Shogren et al. (2004)

Effect size	Model	Estimate (L, U)	95% CI (L, U)	$\hat{\tau}^2$ L	$\hat{\tau}^2$ U	$\hat{\sigma}^2$ L	$\hat{\sigma}^2$ U
Prevalence	1	(-1.85, -0.67)	(-2.53, -0.09)	0.42	0.18	0.70	0.87
ratio ω^ϕ	2	(-1.74, -0.77)	(-2.38, -0.22)	0.35	0.14	0.70	0.85
Prevalence	3	(-2.14, -0.95)	(-2.89, -0.34)	0.37	0.02	1.31	1.59
odds ratio ψ	4	(-1.58, -1.42)	(-2.19, -0.86)	0.06	0.00	1.39	1.29

which indicates a substantial level of variability across cases and studies.[19] Whether based on the lower- or upper-bound meta-analysis, it appears that more of the heterogeneity is within studies than between studies.

Model 2 adds additional assumptions regarding cases measured using partial interval recording; specifically, Model 2 assumes that the average event duration is less than 60 seconds and that at most a certain percentage of interim times are less than the active interval length used for measurement. The additional assumptions reduce the width of the estimated bounds by 0.21 and the width of the confidence interval by 0.27.[20] Under Model 2, the 95% confidence interval for the average log-prevalence ratio corresponds to a reduction in prevalence of between 20% and 91%. The substantive conclusion is much the same as under Model 1: on average, the treatment reduces problem behavior, but the magnitude of the average reduction remains uncertain. Just as in Model 1, there appears to be substantial heterogeneity of treatment effects across cases and studies, though it

[19]Suppose that the average log-prevalence ratio is -0.67 (corresponding to a reduction of about 50%). Total heterogeneity of $\tau^2 + \sigma^2 = 1.04$ implies that a quarter of the population experiences a reduction in prevalence of over 70% while another quarter of the population experiences an increase in prevalence of 1% or more.

[20]The width of the confidence interval is reduced by more than the width of the estimated bounds because the estimated between-study variation $\hat{\tau}^2$ is smaller under Model 2, which leads to a reduced standard errors on the estimated bounds for the average effect size.

should be borne in mind that the estimated variance components do not account for the use of bounds rather than point estimates.

Model 3 relies on the same assumptions regarding partial interval recording as does Model 2, but uses the log-prevalence odds ratio as the effect size metric. Based on the assumptions of Model 3, the average log-prevalence odds ratio is estimated to be between -2.14 and -0.95 (95% confidence interval: [-2.89, -0.34]). Odds ratios can be difficult to interpret; as an aid to interpretation, it is helpful to provide translations into proportionate reductions in prevalence at benchmark levels of baseline prevalence.[21] For a baseline prevalence of $\phi^0 = 0.50$, the confidence interval for the average log-prevalence odds ratio corresponds to a reduction of between 17% and 90%, which is quite close to the confidence interval from Model 2; for a baseline prevalence of $\phi^0 = 0.75$, the corresponding reduction is between 9% and 81%. Differences between Model 2 and Model 3 are due to the change of metric, and so the results are not on the same scale. However, the I^2 statistic (Higgins & Thompson, 2002) can be used to compare heterogeneity across the two models because it is a scale-free measure.[22] The I^2 statistics from Model 3 are both $I^2_L = I^2_U = 0.84$ and slightly larger than those from Model 2 (where $I^2_L = 0.79, I^2_U = 0.78$).

[21]For baseline prevalence ϕ^0 and log-prevalence odds ratio ψ, the proportionate reduction in prevalence is given by

$$\exp\left(\omega^{\phi}\right) - 1 = \frac{\exp(\psi)}{1 - \phi^0\left[1 - \exp(\psi)\right]} - 1.$$

[22]In a basic meta-analysis, I^2 is a function of Cochran's Q statistic and the total number of studies. However, this simple relationship does not hold in multi-level meta-analysis, where multiple components of true variation between effects must be estimated. I therefore calculate I^2 using

$$I^2_b = \frac{\hat{\tau}^2_b + \hat{\sigma}^2_b}{\hat{\tau}^2_b + \hat{\sigma}^2_b + \bar{V}_b},$$

where $\hat{\tau}^2_b$ and $\hat{\sigma}^2_b$ are the restricted maximum likelihood estimates and $\bar{V}_b$ is the simple mean of the effect size variances V^b_{ij}.

This suggests that the log-prevalence ratio metric offer a small advantage for modeling the variability of individual effect sizes, though the difference is slight.

Model 4 is based on different assumptions regarding partial interval recording than those used in Model 3. Under Model 4, the estimated bound for the average log-prevalence odds ratio is -1.58 to -1.42 (95% confidence interval: [-2.19,-0.86]). For a baseline prevalence of $\phi^0 = 0.50$, the confidence interval corresponds to proportionate reductions in prevalence of between 51% and 59%; for a baseline prevalence of $\phi^1 = 0.75$, the corresponding reduction is between 46% and 62%. The stronger assumptions of Model 4 yield an estimated bound for the average log-prevalence odds ratio that is considerably narrower, having a width of 0.16 logits, compared to 1.19 logits under Model 3; the corresponding confidence interval has a width of 1.32 logits, compared to 2.55 logits under Model 3. Also as compared to Model 3, there is less difference in the variance component estimates when based on the lower versus the upper bounds from Model 4. When based on the lower bound estimates, the estimated between-study variance is small ($\hat{\tau}^2 = 0.06$) while the estimated within-study variance remains fairly large ($\hat{\sigma}^2 = 1.39$).[23]

5.5.4. Discussion

I have presented four models for synthesizing case-level estimates of the effect of providing choice-making opportunities on the prevalence of individuals' problem behavior, using data from studies identified by Shogren et al. (2004). A major feature of these data is

[23]Total heterogeneity is also large. For an average log-prevalence odds ratio of -1.58, total heterogeneity of $\tau^2 + \sigma^2 = 1.45$ implies that a quarter of the population has a log-prevalence odds ratio of less than -2.39 (a reduction of more than 70% for baseline prevalence $\phi^0 = 0.50$) while another quarter of the population has a log-prevalence odds ratio of more than -0.76 (a reduction of less than 22% for baseline prevalence $\phi^0 = 0.50$).

the large number of cases measured using partial interval recording, which is not a direct measure of prevalence. Each model entertained a different set of assumptions in order to identify bounds for a meaningful, measurement-comparable effect size that could be estimated from interval recording data. Models 1, 2, and 3 lead to very similar inferences for the average effect of choice-making; across all three models, the treatment is, on average, effective in reducing the prevalence of problem behavior, but the magnitude of the average reduction is imprecisely estimated. The lack of precision is due in part to heterogeneity of effects across cases, but also due to the wide bounds on individual treatment effect estimates that are based on partial interval recording data. Model 4 presents a much more precise picture than the others, but this precision is based on strong modeling assumptions. The results reported for Model 4 should be interpreted in light of the assumptions, and with considerable caution.

The findings based on Models 1 through 3 contrast somewhat with the conclusions reached by Shogren et al. (2004), who inferred based on the same data that "providing choice opportunities resulted in a clinically significant reductions" (p. 229) in problem behavior. Shogren et al. (2004) used as effect size measures the percentage of non-overlapping data and the percentage of zero data, and synthesized those effect sizes by taking simple averages. They also performed several sub-group analyses, using Kolmogorov-Smirnov tests and Kruskal-Wallis analysis of variance procedures for identifying statistically significant differences between groups. I do not yet have access to the covariate information necessary to re-produce these subgroup analyses, but plan to do so in future work. I anticipate that this exercise may not be very informative because it will involve contrasts

between bounds on average treatment effects in each subgroup, leading to even wider bounds for the difference in effects.

Further analyses would be possible if a larger set of studies could be identified that included a greater variety of measurement methods. Among those identified by Shogren et al. (2004), only 4 studies (9 cases) used a measurement method other than interval recording; only three cases were measured using event counting, only one with momentary time sampling, and only five with continuous recording (all from a single study). As a result, there is insufficient data to examine whether there are differences in average effect sizes for cases measured by different methods. However, it may be possible to carry out such an analysis in other applications, and meta-analysts are encouraged to do so. Although the point of using measurement-comparable effect sizes is to reduce irrelevant operational heterogeneity and put different measurement procedures on a comparable basis, it is prudent to test whether the exercise has succeeded. Residual differences between measurement methods may indicate violation of one's modeling assumptions, which can in turn lead the analyst towards more refined assumptions and signal caution in the interpretation of effect sizes averaged across measurement procedures.

5.6. General discussion

In this chapter, I have presented an alternating renewal process model for free-operant behavior that can be used to describe data collected via several common measurement procedures. I used the model to define measurement-comparable effect size metrics and proposed some estimators that are applicable under a simple between-phase model. When data are collected using a direct measurement procedure, several of the estimators are

special cases of the log-response ratio, a well-known effect size used in other areas of meta-analysis (Hedges et al., 1999). However, data based on interval recording procedures present special difficulties because they cannot be interpreted as direct measurements of either prevalence or incidence. I have proposed several different approaches to estimating measurement-comparable effect sizes based on interval recording data, all of which involve further modeling assumptions that may be difficult to verify in practice, and all of which yield bounds rather than point estimates.

A working meta-analyst, interested in synthesizing evidence from single-case studies of free-operant behavior, might sensibly question the need for an elaborate, notation-laden model to define effect sizes. I see three advantages to this model. First, using a model that captures the essential features of the outcome measurement procedures improves the interpretability of effect sizes defined with respect to it. The effect sizes that I have proposed are defined in terms of prevalence and incidence, both readily understood aspects of a behavior. In contrast, other effect size proposals such as the within-phase standardized mean difference or the percentage of non-overlapping data do not correspond closely with interpretable constructs. Instead, both of these measures are sensitive to the precision of the measurements on which they are based, and thus to operational aspects of the study such as the chosen recording procedure, the length of observation sessions, and (for interval recording procedures) the active interval length.

Second, a measurement-comparability model is all the more necessary when dealing with measurements that are difficult to interpret. Given that interval recording procedures are widely used for measurement of free-operant behavior, meta-analytic methods for single-case research cannot simply ignore them. Consider that, in the application to

studies of choice-making interventions, interval recording was used with two thirds of the cases to be synthesized. Taking a naive approach by treating interval recording data just as other data would compromise the construct validity of the synthesis. On the other hand, restricting the meta-analysis to cases measured using procedures other than interval recording would drastically reduce the sample size and possibly compromise external validity. Methods are therefore needed that retain cases measured using interval recording while also making use of interpretable, measurement-comparable effect sizes.

Third, use of a measurement-comparability model has implications for current and future research practices. Effect sizes defined under such a model allow meta-analysts to formulate research questions using more exact terms, such as whether an intervention affects the prevalence of a behavior, the incidence of a behavior, or both. As I noted in Section 5.2.1, it will rarely be possible to estimate theoretically interesting effect sizes such as the log-duration ratio and log-interim ratio based only on data collected from published graphs. However, such limitations do not pertain to primary researchers planning future studies; data collection procedures and reporting practices could certainly be adjusted so that effects on event duration and interim time could be separately estimated. Similarly, meta-analytic methods that account for the relative precision of different measurement procedures (and, in the case of interval recording, their validity) provide a rational guide for the design of new research. Researchers planning a new study can assess the state of existing research to identify outcomes or sub-groups where average effect sizes are imprecise and choose measurement methods, designs, and sample sizes accordingly.

5.6.1. Choosing an effect size

Several factors are relevant when choosing an effect size metric to use for summarizing the results of a single-case study or meta-analyzing the results of several studies. I consider both purposes in turn. For summarizing the results of a study on a single case, the choice of effect size should follow directly from the goals of the study. Primary investigators should select both a measurement procedure and an effect size based on which dimension of a case's behavior is of theoretical or practical concern. If prevalence is the primary dimension of interest, continuous recording or momentary time sampling is recommended, and the choice between the log-prevalence ratio or log-prevalence odds ratio can likely be made based on which metric is more easily interpreted. Similarly, if incidence is the primary concern, event counting is recommended, together with the log-incidence ratio as a summary effect size.[24] If an investigator chooses to use an interval recording method, they should nonetheless choose (and defend the choice of) a dimension of behavior that is of primary interest, then summarize their results using an effect size appropriate for that dimension.

In the context of meta-analyzing several studies, the choice of a summary effect size will be limited to those that can be estimated from the available data. If the set of included studies used varying measurement procedures, then it may be necessary to use several different effect sizes and to evaluate whether the those effect sizes are directly comparable, on the basis of the assumptions described in Section 5.2.5. For example, if some studies measure incidence with an event counting procedure while the remainder measure prevalence with continuous recording, each subset of studies might first be

[24]If both prevalence and incidence are of interest, the most robust approach is to use an electronic system for continuous recording, such as the MOOSES software (Tapp, Wehby, & Ellis, 1995).

summarized and meta-analyzed separately; the meta-analyst might then assess whether it is reasonable to assume that average event durations can be treated as constant for the studies that use event counting, in which case a combined meta-analysis of both subsets would be warranted.

If the preponderance of included studies focus on prevalence, then the meta-analyst must further determine whether to use the log-prevalence ratio or the log-prevalence odds ratio. In this circumstance, it may be possible to choose an effect size metric by comparing the empirical fit of the meta-analytic models applied to each effect size, with preference given to the effect size metric that exhibits less heterogeneity. This approach was used by Engels, Schmid, Terrin, Olkin, and Lau (2000) for meta-analysis of medical studies with binary outcomes (see also Deeks, 2002). These authors found that it was often difficult to distinguish between models based on risk ratios versus those based on odds ratios, particularly when effects were small and the number of studies was limited. In applying this approach to the studies of the effects of choice-making problem behavior, as identified by Shogren et al. (2004), the ability to distinguish between models was hampered not by small effects, but by the limitations of partial interval recording data, which necessitated the use of effect size bounds. However, other single-case meta-analyses, particularly those in which partial interval recording is less common, might have comparatively greater power for determining the more homogeneous effect size metric, due both to the availability of case-level data and the potential for very large effects on prevalence.

5.6.2. Limitations and Extensions

The models and methods presented in this chapter have several limitations that should be noted, related mostly to the modeling assumptions upon which I have relied. Here I address, in turn, assumptions related to the within-session model, between-session model, effect sizes, and meta-analytic methods. I also note possible extensions and areas for future research.

The equilibrium alternating renewal process model described in Section 5.1.2 is general in the sense that only first-moment assumptions are specified regarding event durations and interim times. However, the assumption that the process is in equilibrium may strike some as unrealistic, particularly when observation sessions coincide with another event, such as the beginning of math class, or entrances into a novel setting, such as a therapist's office. Others may object that the model does not account for relationships between behavioral events and environmental contingencies, which are of great interest in behavior-analytic theories. Regarding both objections, I can offer no response but that available data (in the form of published graphs) do not provide sufficient information to model the aspects of interest. Without the equilibrium assumption, all of the measurement procedures that I have described become somewhat sensitive to initial conditions and to the length of observation sessions, but the meta-analyst will not have information about these dimensions. The equilibrium assumption implies that the process is uniform over the course of a session, which is consistent with how the recorded data generated by various measurement procedures are typically summarized into a single reported datum.

Next, I have used the simplest possible between-session model to describe change in behavior over time. The stable-phase model allows for neither trends in the process

over time nor for serial dependence of repeated measurements, both of which are prime concerns in quantitative single-case research methodology (Horner et al., 2012; Maggin, Swaminathan, et al., 2011; Wolery et al., 2010). Extensions to the between-session model are possible and will be described in Chapter 6. Even so, the simple stable-phase model may be plausible in many applications, particularly for designs with very short phase lengths.

There are two shortcomings to the case-level effect size metrics that I have described. The goal of some behavioral interventions is complete elimination of an undesirable behavior, and a researcher or meta-analyst may be interested in the probability that elimination will be achieved. If the complete absence of a behavior is achieved in one treatment condition, prevalence or incidence will be equal to zero and effect size metrics such as the log-prevalence ratio, log-prevalence odds ratio, and log-incidence ratio will have a value of negative infinity, making them impossible to analyze using conventional meta-analytic models. Specialized meta-analytic models and methods will be needed in such contexts.

The other short-coming of the individual-level effect size metrics as proposed is that they are not design-comparable in the sense of Section 1.1.1, and thus not useful for syntheses containing both single-case and group research. To address this, the general framework outlined in Chapter 2 could be applied to define design-comparable analogs of the effect sizes described in this chapter. An alternative approach would be to develop methods for bounding averages of individual-level effect sizes based on data from group designs. The latter approach is closely related to methods for detecting unit-treatment

interactions in clinical trials (Gadbury & Iyer, 2000; Gadbury, Iyer, & Albert, 2004; Poulson, Gadbury, & Allison, 2012). Future work on either approach will need to be developed in the context of specific applications.

A final, important limitation of this work is that the key assumptions have yet to be vetted by applied researchers who have experience with measurement procedures for free-operant behavior. Particularly for analysis of interval recording data, I have suggested several different approaches that strike me as reasonable and useful, but those with more direct experience with application will be in a better position to judge their plausibility and utility. For example, in the application described in Section 5.5, I have employed assumptions uniformly across cases measured using partial interval recording. A researcher with knowledge of the individuals or classes of behavior might be able to develop assumptions that are more carefully tailored to each case. Similarly, a better understanding of how researchers choose between alternative measurement procedures would be helpful in assessing which assumptions regarding those procedures are most reasonable. This final limitation speaks to the need for–and importance of–greater collaboration between applied single-case researchers and statistical methodologists, calls for which have also been made by J. M. Campbell and Herzinger (2010) and others.

CHAPTER 6

Generalized linear models for free-operant behavior

The effect size estimators proposed in Chapter 5 are motivated by an equilibrium alternating renewal process (ARP) for the stream of free-operant behavior observed during the course of an observation session, based on which the properties of several different types of recorded data can be derived. They are further motivated by a between-session model in which the behavior stream process is stable, leading to reported data that are independent and identically distributed within each treatment condition. In this chapter, I consider models that relax the stability assumption in two ways: by allowing for deterministic time trends and by allowing for stochastic, possibly serially correlated variation in the parameters of the behavior stream process. Both of these extensions involve the framework of quasi-likelihood for generalized linear models. I focus on the types of reported data that are direct measures of behavioral parameters, namely event counting (a direct measure of incidence) and continuous recording and momentary time sampling (both direct measures of prevalence); applications to interval recording methods remain a topic for future work. I also limit consideration to one effect size metric for each of these measurement procedures: for eventing counting, I consider the log-incidence ratio; for continuous recording and momentary time sampling, I consider the log-prevalence odds ratio. I do not examine the log-prevalence ratio for the latter types of data because it is difficult to specify sensible data-generating models for the log-prevalence.

The central challenges in extending the model from the previous chapter stem from a lack of full probability distributions for the types of reported data under consideration that are both plausible and tractable. Under the posited ARP, the first moments of data generated by direct measurement procedures depend only on the first moments of the event durations and interim times that constitute the latent behavior stream. However, full probability distributions for the recorded data will depend on the event duration distribution F_D and the interim time distribution F_E, about which little is known. Moreover, even if these could be specified, the resulting probability distributions for the recorded data would typically be cumbersome if not entirely intractable.

The quasi-likelihood framework, described by Wedderburn (1974; see also McCullagh, 1983), is an estimation criteria based on models for the mean and variance of the outcome, rather than the full distributional form. For example, consider the stable phase model from the previous chapter. For outcome data based on direct measures, $r \in \{E, C, M\}$, denote $\mathrm{E}\left(Y_j^r\right) = \pi_j^r$. The model can then be written as

$$
\begin{aligned}
g_r\left(\pi_j^r\right) &= \beta_0 + \beta_1 T_j, \\
\mathrm{Var}\left(Y_j^r\right) &= \left[\gamma_0(1 - T_j) + \gamma_1 T_j\right] V_r\left(\pi_j^r\right),
\end{aligned}
\tag{6.1}
$$

$j = 1, ..., n$, where g_r is a known link function and V_r is a known function expressing the relationship between the mean and the variance, both for an outcome of type r. A logit link $g_r(x) = \ln(x) - \ln(1-x)$ with continuous recording data or momentary time sampling data leads to a model in which β_1 is the log-prevalence odds ratio; taking $g_r(x) = \log(x)$ with event counting data leads to a model in which β_1 is the log-incidence ratio.

As written in (6.1), the parameter γ_0 measures the dispersion of the outcome in the baseline phase above or below what would be expected for a given variance function V_r; similarly, γ_1 measures dispersion in the treatment phase. Of course, for this trivial model, it is not strictly necessary to use a variance function V_r because, with only two possible values of the mean, V_r is not identified; in fact, the variance model can be written more simply as $\text{Var}\left(Y_j^r\right) = \sigma_0^2(1 - T_j) + \sigma_1^2 T_j$. This is why mean-variance relationships were not a concern in the previous chapter. However, in the more general models considered in this chapter, it will be necessary to consider how V_r should be specified.

A natural way to extend (6.1) is to add further covariates to the mean specification, such as for describing time trends. Here, the primary challenges in applying the quasi-likelihood framework are in determining how to choose a variance function and how to estimate the sampling variance of the effect size estimate. I illustrate this modeling extension and address these challenges in Section 6.1.

One way to introduce serial dependence into a model such as (6.1) is by adding another source of error to the mean specification–that is, by allowing stochastic variation in the parameters of the behavior stream. If these errors are serially dependent, so to will be the reported outcome data, resulting in what Cox (1981) termed a parameter-driven time series model. For outcome data in the form of counts, parameter-driven models have been widely studied, including notably by Zeger (1988) and Davis, Dunsmuir, and Wang (2000). For outcome data in the form of proportions (as produced by continuous recording and momentary time sampling), parameter-driven models have been studied by Song and Tan (2000), Molenberghs and Verbeke (2005, Chp. 22), and Czado and Song (2007), among others. Models with this feature create two additional challenges for effect

size estimation in single-case studies. The first is definitional: how should effect sizes be defined under such models? Beyond this, how should effect sizes be estimated? I consider parameter-driven models involving serial dependence in Sections 6.2 and 6.3. Section 6.4 demonstrates both extensions in applications and Section 6.5 concludes.

In both of these extensions, the main advantage of adopting a quasi-likelihood approach is that its modeling assumptions are reasonable and concordant with the current state of knowledge about the processes under study. Absent more fine-grained, within-session data, it seems judicious to use methods that do not involve committing to specific distributional assumptions about the behavior stream. However, quasi-likelihood also carries a serious caveat: the performance and properties of estimators in this framework are justified on the basis of asymptotic consistency arguments, rather than exact, small-sample results. Asymptotic consistency is rather cold comfort for the applied single-case researcher or the meta-analyst who must deal with limited available data. I thus rely on small-sample simulation results to make initial assessments regarding the performance of the methods described in the following sections.

6.1. Defining and estimating models with time trends

The mean specification from (6.1) can be extended to capture time trends in the behavior stream process. In Chapter 3, I considered a case-level model for the multiple baseline design with continuous, interval-scale outcome measures; model (3.2) included a baseline time trend, a treatment effect, and a treatment-by-trend interaction term. Following a specification similar to (3.2), the mean of the outcome process can be modeled

as

$$g_r\left(\pi_j^r\right) = \beta_0 + \beta_1 T_j + \beta_2 \times j + \beta_3 \sum_{k=1}^{j} T_k. \tag{6.2}$$

The coefficients in this model have interpretations very similar to those in (3.2), but their units depend on the choice of link function. When the outcome is a measure of prevalence and g_r is a logit link, the baseline level β_0 is in log-prevalence-odds units; the initial treatment effect β_1 is the difference in log-prevalence-odds (the log-prevalence odds ratio) immediately upon introduction of the treatment; the baseline trend β_2 is the linear change in log-prevalence odds per time unit (i.e., session); and β_3 is the difference in linear trends between baseline and treatment phases, also in log-prevalence odds per unit time. When the outcome is a measure of incidence and g_r is a log link, the coefficients have units of log-incidence, difference in log-incidence, or change in log-incidence per time unit.

There are two possible approaches to defining an effect size under model (6.2). One approach would be to use the initial treatment effect β_1 and the effect on trend β_3 as a two-dimensional description of the effect size. The other approach would be to choose a clinically meaningful, fixed duration of treatment for purposes of summary; for example, one could take as the target parameter the effect of treatment four sessions after introduction, $\beta_1 + 4\beta_3$. When the outcome is a measure of prevalence and g_r is a logit link, the latter effect size is a log-prevalence odds ratio; when the outcome is a measure of incidence and g_r is a log link, it is a log-incidence ratio.

The latter approach is similar to the approach to defining design-comparable effect sizes described in Section 2.4.1. For purposes of summarizing a single study, either approach may be reasonable. For purposes of research synthesis, the use of a bi-variate

summary effect size has the advantage of retaining greater detail, but requires the use of multi-variate meta-analytic models. On the other hand, using a single endpoint requires the analyst to choose and justify a particular follow-up time, but also makes possible simpler, more conventional approaches to meta-analysis. In what follows, I take the latter approach because it leads to a somewhat simpler presentation. The remainder of this section describes how to obtain an estimate of the target effect size parameter and a corresponding variance estimate.

6.1.1. Effect size estimation with quasi-likelihood

For ease of notation, I describe methods of estimating a model given by

$$\begin{aligned} g_r\left(\pi_j^r\right) &= \mathbf{x}_j'\boldsymbol{\beta} \\ \operatorname{Var}\left(Y_j^r\right) &= \left(\mathbf{t}_j'\boldsymbol{\gamma}\right) V_r\left(\pi_j^r\right) \end{aligned} \tag{6.3}$$

for $j = 1, ..., n$, where $\mathbf{x}_j$ is a p-dimensional covariate vector that includes the treatment indicator T_j, $\boldsymbol{\beta}$ is a p-dimensional mean parameter, $\mathbf{t}_j = (1 - T_j, T_j)'$, and $\boldsymbol{\gamma}$ is a 2-dimensional vector of dispersion parameters. Working with this general formulation also has the advantage that the estimation methods I describe can be applied to other models in addition to (6.2). Model (6.2) is of course a special case with $\mathbf{x}_j = (x_{0j}, x_{1j}, x_{2j}, x_{3j})'$, where $x_{0j} = 1$, $x_{1j} = j$, $x_{2j} = T_j$, and $x_{3j} = \sum_{k=1}^{j} T_k$. Throughout, I assume that $g_r(x) = \text{logit}(x)$ for $r = C, M$ and $g_E(x) = \log(x)$. Let $h_r = g_r^{-1}$, so that $h_r\left(\mathbf{x}_j'\boldsymbol{\beta}\right) = \pi_j^r$; also denote $\eta_j = \mathbf{x}_j'\boldsymbol{\beta}$.

The effect size parameter of interest is a linear combination of the mean parameter components, $\mathbf{c}'\boldsymbol{\beta}$ for fixed, p-dimensional $\mathbf{c}$. In the quasi-likelihood framework, an estimator of $\boldsymbol{\beta}$ is defined as the solution to the p-dimensional quasi-score equation for the mean parameters

$$\mathbf{U}_1(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \sum_{j=1}^{n} \mathbf{x}_j \left(\frac{dh\left(\mathbf{x}_j'\boldsymbol{\beta}\right)}{d\eta_j} \right) \frac{Y_j^r - h_r\left(\mathbf{x}_j'\boldsymbol{\beta}\right)}{\left(\mathbf{t}_j'\boldsymbol{\gamma}\right) V_r\left(h_r\left(\mathbf{x}_j'\boldsymbol{\beta}\right)\right)} = \mathbf{0} \tag{6.4}$$

and a 2-dimensional quasi-score equation for the dispersion parameters

$$\mathbf{U}_2(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \sum_{j=1}^{n} \mathbf{t}_j \left[\frac{\left[Y_j^r - h_r\left(\mathbf{x}_j'\boldsymbol{\beta}\right)\right]^2 / V_r\left(h_r\left(\mathbf{x}_j'\boldsymbol{\beta}\right)\right) - \mathbf{t}_j'\boldsymbol{\gamma}}{\left(\mathbf{t}_j'\boldsymbol{\gamma}\right)^2} \right] = \mathbf{0} \tag{6.5}$$

for a variance function V_r that is yet to be specified. Equivalently, an estimator of $\boldsymbol{\beta}$ can be defined as the maximizer of a quasi-likelihood criterion function and an estimator of $\boldsymbol{\gamma}$ can be defined as the maximizer of an extended quasi-likelihood criterion, as described in McCullagh and Nelder (1989, Chps. 9-10). Equations (6.4) and (6.5) can be solved for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ using an interlinked, iteratively re-weighted least squares (IRLS) fitting algorithm (McCullagh & Nelder, 1989). An R implementation of this algorithm can be found in the package `dglm` (Dunn & Smyth, 2012). Let $\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}}$ denote solutions to the quasi-score equations; the effect size estimator is then $\mathbf{c}'\hat{\boldsymbol{\beta}}$. Also let $\hat{\eta}_j = \mathbf{x}_j'\hat{\boldsymbol{\beta}}$ and $\hat{\pi}_j^r = h\left(\hat{\eta}_j\right)$.

When modeling multiple baseline designs, piece-wise linear regressions for each phase, as in (6.2), will be a common specification for the mean. In this case, $\boldsymbol{\beta}$ can be re-parameterized into $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$, each of which is estimated based only on the data from a single phase; the quasi-score equation for the mean parameters (6.4) will then be independent of $\boldsymbol{\gamma}$. Given an estimate $\hat{\boldsymbol{\beta}}$, estimates of $\boldsymbol{\gamma}$ for the piece-wise model can be computed

as

$$\gamma_0 = \frac{1}{\sum_{j=1}^{n}(1-T_j)} \sum_{j=1}^{n}(1-T_j)\frac{\left(Y_j^r - \hat{\pi}_j^r\right)^2}{V_r\left(\hat{\pi}_j^r\right)}, \qquad \gamma_1 = \frac{1}{\sum_{j=1}^{n} T_j} \sum_{j=1}^{n} T_j \frac{\left(Y_j^r - \hat{\pi}_j^r\right)^2}{V_r\left(\hat{\pi}_j^r\right)}. \tag{6.6}$$

The divisors in (6.6) are the sample sizes in each phase; alternately, one could divide by the sample size minus $p/2$ as an approximate degrees-of-freedom correction. With this correction, the estimators $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\gamma}}$ are exactly equivalent to the results of fitting separate models to each phase. It is this approach that I follow in the simulations and examples described later. For fitting the models, I use a modified IRLS algorithm with a step-halving provision that ensures a monotonically increasing quasi-likelihood. The modification provides improved the convergence properties (I. Marschner, 2011) and is implemented in the `glm2` package (I. C. Marschner, 2012) in R.

The quasi-score function $\mathbf{U}_1(\boldsymbol{\beta}, \boldsymbol{\gamma})$ is an unbiased estimating equation for $\boldsymbol{\beta}$, given an arbitrary value for $\boldsymbol{\gamma}$. As a consequence, the estimator $\hat{\boldsymbol{\beta}}$ is an asymptotically consistent estimator of $\boldsymbol{\beta}$ provided only that some rather general regularity conditions are satisfied (Liang & Zeger, 1986). Crucially, this is true regardless of whether the variance function V_r is correctly specified. The most salient of the regularity conditions is that the matrices

$$\frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j' \frac{\mathrm{Var}\left(Y_j^r\right)}{\left(\mathbf{t}_j'\boldsymbol{\gamma}\right)^2 V_r^2\left(\pi_j^r\right)} \left(\frac{dh\left(\eta_j\right)}{d\eta_j}\right)^2 \quad \text{and} \quad \frac{1}{n}\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j' \frac{1}{\left(\mathbf{t}_j'\boldsymbol{\gamma}\right) V_r\left(\pi_j^r\right)} \left(\frac{dh\left(\eta_j\right)}{d\eta_j}\right)^2$$

must have positive-definite limits as $n \to \infty$. This condition will not be satisfied if, for instance, the series length increases but the number of observations in the baseline phase $\sum_{j=1}^{n}(1-T_j)$ remains fixed. The practical implication is that collecting a long data

series is not enough; instead, the sample size within each treatment condition must be sufficiently large to ensure that the estimator $\hat{\boldsymbol{\beta}}$ will perform well.

6.1.2. Selecting a variance function

I have yet to describe the exact form for the variance function V_r in (6.4) and (6.5). This may seem to be a difficult problem because the variance of the recorded data will depend on parameters of the event duration and interim time distributions over an above just their expectations. In fact, even assuming specific parametric forms for F_D and F_E, it is difficult to obtain analytic expressions for the variance of event counting, continuous recording, or momentary time sampling data except in a few special cases.

The most tractable case is for an alternating Poisson process, a special case of an ARP in which event durations and interim times are assumed to be exponentially distributed, with $D_1 \sim Exp(1/\mu)$ and $E_1 \sim Exp(1/\lambda)$. This process has the property that non-overlapping time increments are both stationary and independent, which considerably simplifies the derivation of moments. Denoting the length of the observation session as L, prevalence as $\phi = \mu/(\mu+\lambda)$, and incidence as $\zeta = 1/(\mu+\lambda)$, Table 6.1 reports the variance of a reported datum generated by each of the three direct measurement procedures (see Appendix B.4 for derivations of these expressions). Even in this basic case, the mean-variance relationships are unwieldy expressions that involve nuisance parameters: for the procedures that measure prevalence, the variances depend on incidence, and for event counting, which measures incidence, the variance depends on prevalence.

Several possible modeling strategies could be followed here. One would be to assume that the behavior stream follows an alternating Poisson process and that the nuisance

Table 6.1. Moments of reported datum under an alternating poisson process

Recording procedure	E(Y)	Var(Y)
Event counting	ζL	$\zeta L\left[\phi^2+(1-\phi)^2\right]+2\phi^2(1-\phi)^2\left[1-\exp\left(\frac{-\zeta L}{\phi(1-\phi)}\right)\right]$
Continuous recording	ϕ	$\frac{2\phi^2(1-\phi)^2}{\zeta L}\left(1-\frac{\phi(1-\phi)\left[1-\exp\left(\frac{-\zeta L}{\phi(1-\phi)}\right)\right]}{\zeta L}\right)$
Momentary time sampling	ϕ	$\frac{\phi(1-\phi)}{K}\left[1+\frac{2}{K}\sum_{k=1}^{K-1}(K-k)\exp\left(\frac{-\zeta k L}{\phi(1-\phi)K}\right)\right]$

parameters are constant within each phase; on the basis of these assumptions, one could develop estimators for the nuisance parameters implicated in each variance expression. An alternative strategy would be to forgo estimation and just assume fixed values for the nuisance parameters, based on prior knowledge of the behavior under study. Rather than follow either of these strategies, I propose to approximate the mean-variance relationships of an alternating Poisson process using functions that do not depend on the nuisance parameters. Even when using these approximate variance functions, the estimator $\hat{\boldsymbol{\beta}}$ remains asymptotically consistent, though the form of approximation may affect its finite-sample bias and precision.

For event counting data, the variance is approximately proportional to the mean except when the mean is very small; this suggests taking $V_E(x) = x$, which is the mean-variance relationship of a Poisson distribution.

For continuous recording data, the second term in the variance is close to one except when ζL is small, that is, when there are few events per observation session. This suggests using $V_C(x) = x^2(1-x)^2$, which is sometimes known as the "Wedderburn" variance

function (McCullagh & Nelder, 1989, p. 330). This variance function does not correspond to any exponential family distribution, and is not always available in software for fitting generalized linear models. Thus I also consider using $V_{C2}(x) = x(1-x)$, though this approximation has no particular justification other than that it is a convenient default option in widely-available software.

Finally, the variance of momentary time sampling data could be approximated by its first term, using $V_M(x) = x(1-x)$. This is the mean-variance relationship of a binomial distribution. The accuracy of this approximation depends on both the expected number of events per session ζL and the number of intervals sampled K. The approximation can be justified in a rather different sense than the others, by considering a model not for the reported data, but for the within-session recorded data as defined in Section 5.1.1. If these within-session data are modeled using covariates that are constant within session, then the posited model for the reported data (6.3) with a binomial variance function is equivalent to a generalized estimating equation with independence working assumptions (Liang & Zeger, 1986).

6.1.3. Variance estimation

I consider two approaches to estimating the variance of the effect size $\mathbf{c}'\hat{\boldsymbol{\beta}}$. In the modeling strategy that I have outlined, the variance model given in (6.3) could be only approximately correct or even entirely incorrect. In this situation, a commonly used approach is to turn to variance estimators that are asymptotically consistent in the presence of heteroskedasticity; such estimators are sometimes called "sandwich" or "robust" variance estimators. H. White (1980) and MacKinnon and White (1985) proposed several different

heteroskedasticity-consistent variance estimators for linear regression models; Liang and Zeger (1986) proposed their use in the context of generalized linear models for longitudinal data. Several different, asymptotically equivalent versions of these estimators exist; the one described below uses an approximate first-order bias correction based on the hat matrix (Kauermann & Carroll, 2001, p. 1931). I define the first, robust variance estimator as

$$V_R = \mathbf{c}'\mathbf{B}^{-1}\mathbf{M}\mathbf{B}^{-1}\mathbf{c}, \tag{6.7}$$

where

$$\mathbf{B} = \sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j' \frac{1}{\left(\mathbf{t}_j'\hat{\boldsymbol{\gamma}}\right) V_r\left(\hat{\pi}_j^r\right)} \left(\frac{dh\left(\hat{\eta}_j\right)}{d\eta_j}\right)^2,$$

$$k_j = \frac{1}{\left(\mathbf{t}_j'\hat{\boldsymbol{\gamma}}\right) V_r\left(\hat{\pi}_j^r\right)} \left(\frac{dh\left(\hat{\eta}_j\right)}{d\eta_j}\right)^2 \mathbf{x}_j'\mathbf{B}^{-1}\mathbf{x}_j,$$

and

$$\mathbf{M} = \sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j' \left(\frac{Y_j^r - \hat{\pi}_j^r}{\sqrt{1-k_j}\left(\mathbf{t}_j'\hat{\boldsymbol{\gamma}}\right) V_r\left(\hat{\pi}_j^r\right)}\right)^2 \left(\frac{dh\left(\hat{\eta}_j\right)}{d\eta_j}\right)^2.$$

I use the `sandwich` package (Zeileis, 2004, 2006) in R to compute V_R.

The robust variance estimator V_R is asymptotically consistent provided only that the model for the mean outcome is correct. An alternative approach is to estimate the variance of $\mathbf{c}'\hat{\boldsymbol{\beta}}$ using

$$V_M = \mathbf{c}'\mathbf{B}^{-1}\mathbf{c}. \tag{6.8}$$

The asymptotic consistency of this model-based variance estimator is predicated on having a correct model for the variance of the outcome, about which I have expressed skepticism. Still it seems reasonable to examine its performance, given that it is generated by default in most software for fitting generalized linear models. Furthermore, it is known that robust variance estimators can be inferior to model-based variance estimators, particularly when based on only a small sample of data and when the model is close to correct (i.e., when the degree of heteroskedasticity is mild). Kauermann and Carroll (2001) provided several examples where, under a correct variance model, the asymptotic efficiency of the robust variance estimator is very low relative to the model-based variance estimator. Of course, neither the model-based nor the robust variance estimator carries anything but an asymptotic guarantee. I turn therefore turn to simulation evidence to examine their performance with sample sizes typical of those found in single-case designs.

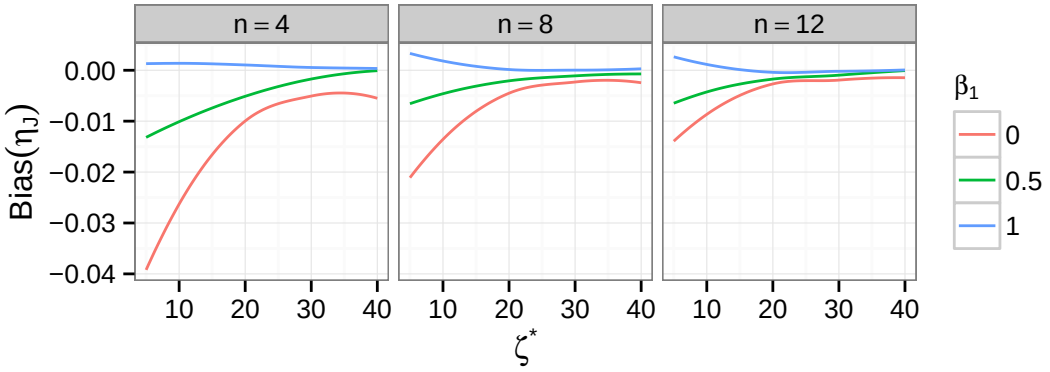
6.1.4. Small-sample performance

I used simulation to evaluate the magnitude of the bias of proposed estimators when based on continuous recording, momentary time sampling, or event counting data; to compare the performance of model-based variance estimators versus robust variance estimators; and to compare the performance of the binomial variance function versus the Wedderburn variance function for modeling continuous recording data. In order to moderate the dimension of the simulations, I examined models involving only a linear trend, as would be used to describe one phase of a design, rather than simulating a full single-case design with multiple phases. For recording procedure $r \in \{E, C, M\}$ and a series of length n, the data-generating model for the mean of Y_j^r is $g_r(\pi_j^r) = \beta_0 + \beta_1 t_j$, $j = 1, ..., n$, where

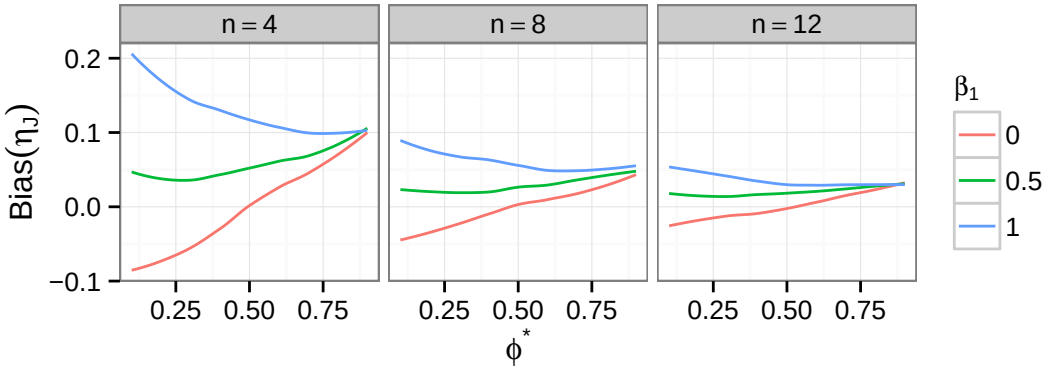
$t_j = (2j - n - 1)/(n - 1)$. For brevity, the presentation in this section is focused on a single target parameter, the predicted value of the transformed mean (i.e., the log-prevalence odds or the log-incidence) at a fixed point in time J, extrapolating beyond the final point in the phase by one quarter of the length of the series; denote this parameter as $\eta_J = g_r(\pi_J^r) = \beta_0 + 1.5 \times \beta_1$.

The simulations varied β_0, β_1, and the data-generating model for the variance of Y_j^r, and included models for the variance in which the assumed analytic model and the actual data-generating model were inconsistent. I varied the phase length from $n = 4$ to $n = 12$, typical of phase lengths found in empirical designs (Shadish & Sullivan, 2011). A complete description of the data-generating model and simulation design can be found in Appendix sections C.2 and C.3, along with a more detailed presentation of simulation results.

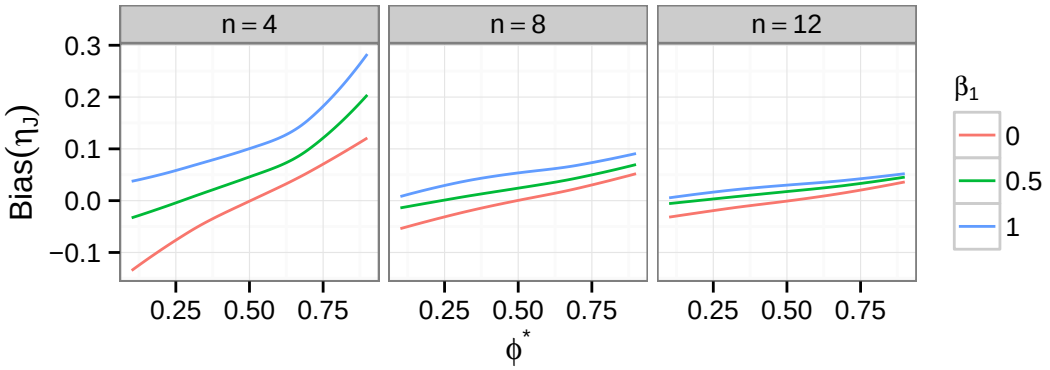
Across the levels of the simulation parameters, the proposed quasi-likelihood estimators generally have small biases, particularly for sample sizes of $n \geq 8$. As would be expected, the biases of the estimators are larger when the underlying ARP is more variable. Figure 6.1 displays the biases of the estimate $\hat{\eta}_J$ for each type of data, using the levels of the simulation parameters that produce the most extreme biases. For event counting data, where $\eta_J = \ln(\zeta_J)$, the maximal bias of $\hat{\eta}_J$ is negligible even for the smallest sample size considered; if the average incidence is at least 10 events per session, then the bias is less than 0.03. For continuous recording data and momentary time sampling data, where $\eta_J = \text{logit}(\phi_J)$, the bias is increasing in the average prevalence and in the magnitude of the trend. When the estimator is based on continuous recording data, η_J tends to be be over-estimated, particularly when there is a steep trend in the series, but the magnitude of the maximum bias is small for $n \geq 8$. When based on momentary time sampling, $\hat{\eta}_J$



(a) Event counting, $\phi^* = 0.1$, $G =$ Exp-Exp.

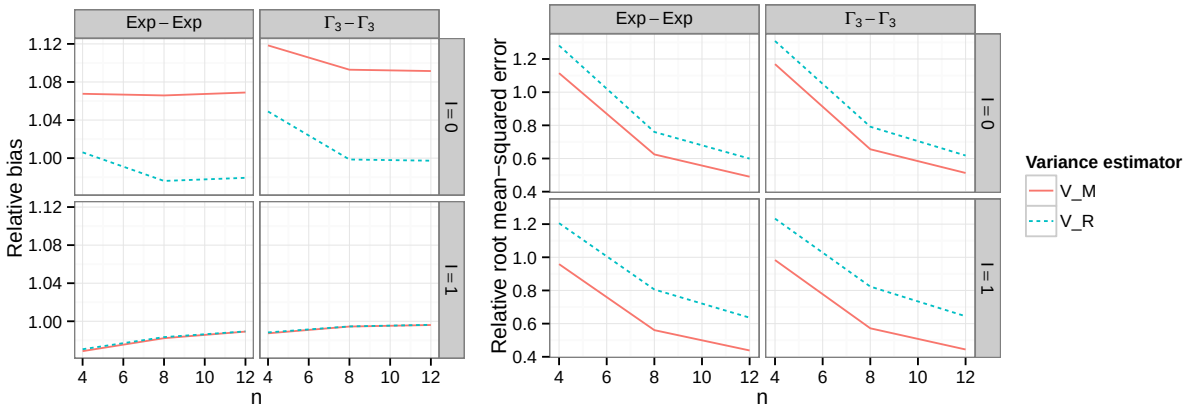


(b) Continuous recording with Wedderburn variance function, $\zeta^* = 5$, $I = 0$, $G =$ Exp-Exp.



(c) Momentary time sampling, $\zeta^* = 5$, $I = 0$, $G =$ Exp-Exp.

Figure 6.1. Maximal bias of quasi-likelihood estimator for η_J based on (a) event-counting data, (b) continuous recording data, and (c) momentary time sampling data.

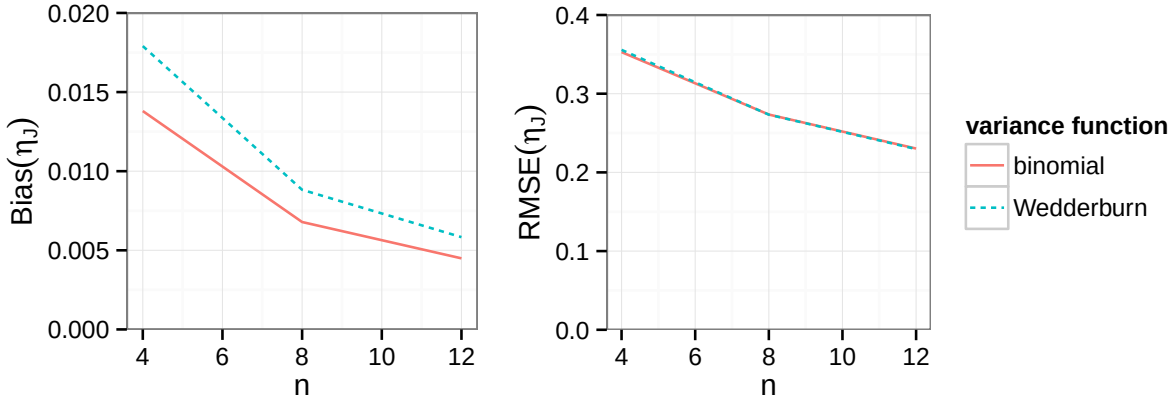


(a) Relative Bias (b) Relative root mean-squared error

Figure 6.2. Average relative bias and relative rmse of model-based and robust variance estimators, based on continuous recording data with Wedderburn variance function.

may have large bias for $n = 4$, but the maximum bias is much reduced for $n \geq 8$. In general, the proposed estimators have tolerable biases even for the fairly small sample sizes typical of single-case studies, at least for the data-generating models considered in these simulations.

Regarding the performance of the alternative variance estimators, a clear pattern holds: though the robust variance estimator is usually less biased than the model-based variance estimator, it also has much larger sampling variability. A typical result is depicted in Figure 6.2, which plots the average relative bias and relative root mean-squared error of the model-based variance estimator $V_M(\hat{\eta}_J)$ and the robust variance estimator $V_R(\hat{\eta}_J)$ based on continuous recording data and using the Wedderburn variance function. The horizontal axis corresponds to sample size n. The rows of the lattice correspond to more mis-specified ($I = 0$) versus approximately correctly specified ($I = 1$) analytic models for the variance; the columns correspond to different event duration and interim



(a) Bias (b) Root mean-squared error

Figure 6.3. Average bias and rmse of $\hat{\eta}_J$ based on continuous recording data, comparing the binomial and Wedderburn variance functions.

time distributions. In Figure 6.2a, it can be seen that both variance estimators are approximately unbiased when the analytic model for the variance is close to correct (bottom row, $I = 1$), but the model-based estimator has positive bias when the analytic model is mis-specified (top row, $I = 0$). However, from Figure 6.2b, V_M has lower relative root mean-squared error, on average, even at the largest sample size considered. The same general pattern of results holds for event counting data and for momentary time sampling data, though the robust estimator is severely biased for momentary time sampling data when $n = 4$. In general, the greater precision of the model-based variance estimator leads me to recommend its use, at least for sample sizes similar to those considered here and when the variance model does not appear to be drastically mis-specified.

For modeling continuous recording data, I have proposed to use the Wedderburn variance function, though I also noted that the binomial variance function might be considered because it is a convenient default. Figure 6.3 displays the average bias and rmse of the estimator $\hat{\eta}_J$ based on each function versus n. Use of the Wedderburn variance function

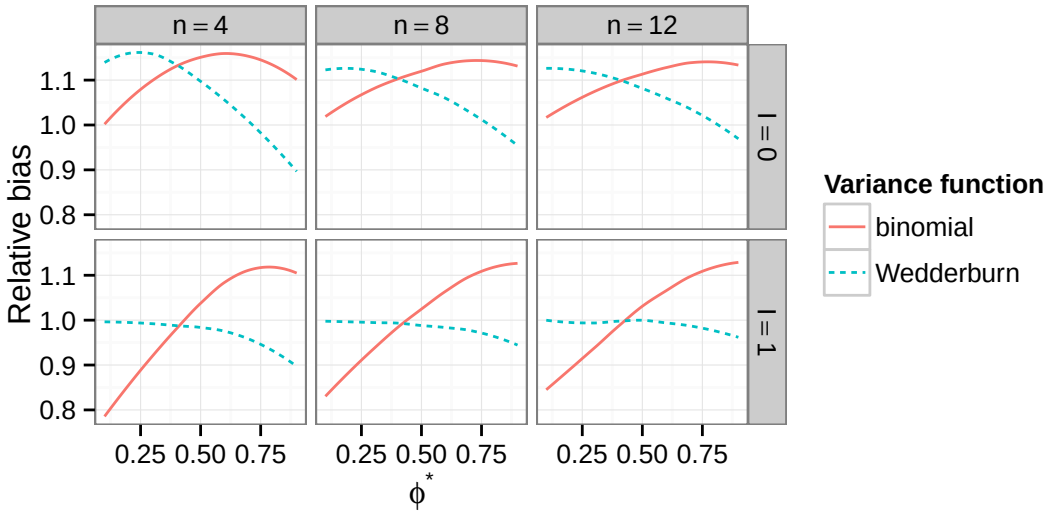


Figure 6.4. Average relative bias of $V_M(\hat{\eta}_J)$ based on continuous recording data with G = Exp-Exp, comparing the binomial and Wedderburn variance functions.

leads to slightly larger bias, but the difference in bias is practically negligible; the rmse of both estimators is nearly identical. The choice of variance function also has implications for the performance of the model-based variance estimator. Figure 6.4 plots the relative bias of $V_M(\hat{\eta}_J)$ when based on each variance function. The generating distributions are taken to be exponential in order to align with the approximation that motivates the Wedderburn variance function. In the bottom row of the lattice ($I = 1$), the analytic model for the variance is approximately correct and the Wedderburn variance function leads to less bias than the binomial variance function; the bias produced by the latter depends strongly on the average level of prevalence. In the top row of the lattice ($I = 0$), the analytic model for the variance is more severely mis-specified, and neither variance function is clearly superior. It would seem that use of the Wedderburn variance function is to be preferred only when the variance model is approximately correct. The practical

implication is that the analyst should assess the fit of the variance model (e.g., through residual analysis) to try to diagnose possible mis-specifications.

6.2. Models with serial dependence

In this section, I describe one approach to extending the model given in (6.3) to incorporate serial dependence between measurements from successive observation sessions. I assume that the serial dependence arises from change in the latent parameters of the behavior stream over time (beyond what can be captured in deterministic trends), rather than from dependence in the measurement process itself. The introduction of variability in the parameters of the behavior stream process leads to a technical distinction between the conditional mean of the process and the marginal mean of the outcome, for which some additional notation will be required. As previously, I assume that the datum from session j is generated by applying a measurement procedure r to a realization of an ARP:

$$\left(Y_j^r | \mu_j, \lambda_j\right) \overset{\text{iid}}{\sim} M_r\left[ARP(\mu_j, \lambda_j)\right] \tag{6.9}$$

(c.f. Equation 5.6). Now though, the parameters of the process μ_j, λ_j (or ϕ_j, ζ_j) will themselves be treated as random. Let $\pi_j^{*r} = \mathrm{E}\left(Y_j^r | \mu_j, \lambda_j\right)$ be the conditional expectation of the reported datum from session j, based on procedure r. With event counting data $\pi_j^{*E} = \zeta_j L$; with continuous recording or momentary time sampling data, $\pi_j^{*r} = \phi_j$, $r \in \{C, M\}$. I will continue to denote the unconditional expectation of the reported datum as $\pi_j^r = \mathrm{E}\left(Y_j^r\right)$; thus, $\pi_j^r = \mathrm{E}\left(\pi_j^{*r}\right)$ where the expectation is with respect to the distribution of the ARP parameters. Note that π_j^r and π_j^{*r} need not be equal.

A between-session model for change in the behavior stream process–and thus an effect size–could conceivably be specified in terms of either the conditional or the marginal mean of the process.[1] The main advantage of conditionally-specified models is that their parameters, including treatment effects, correspond very directly to changes in the behavior stream process. The main drawback of conditionally specified models is that their parameters are sensitive to distributional assumptions regarding the latent process, which can create difficulties for estimation and model-checking (Heagerty & Kurland, 2001; Heagerty & Zeger, 2000). In the context of models for free-operant behavior, where one will have only limited information regarding the conditional distribution of the measurements (i.e., errors arising purely from measurement), it therefore seems more prudent to model the marginal mean of the behavior stream process.

A marginal model for measurements $Y_1^r, ..., Y_n^r$ based on procedure $r \in \{E, C, M\}$ can be written using an implicit specification, as demonstrated by Heagerty and Zeger (2000). In its general form, the model has several parts. First, I posit a generalized linear model for the marginal mean of the process, just as in the previous section:

$$g_r\left(\pi_j^r\right) = \mathbf{x}_j'\boldsymbol{\beta}, \tag{6.10}$$

[1] In a formal sense, the distinction between the two approaches corresponds to a distinction drawn between population-averaged and subject-specified models for analysis of clinical trials (Neuhaus, Kalbfleisch, & Hauck, 1991). However, in the present setting, the marginal mean is averaged over realizations of a posited stochastic process, rather than over individuals. In my view, the advantages of conditionally-specified models (c.f. Lindsey & Lambert, 1998) do not carry the same weight in this context.

where g_r has inverse h_r. Next, I posit a model for the conditional mean of the process, which could be non-linear in $\boldsymbol{\beta}$:

$$g_r\left(\pi_j^{*r}\right) = f_r\left(\pi_j^r, \sigma^2\right) + \nu_j, \tag{6.11}$$

where $\nu_j \sim N(0, \sigma^2)$ and $(\nu_1, ..., \nu_n)$ follow an AR(1) process with serial correlation ρ.[2] The function f_r in (6.11) is defined implicitly by the relationship between the conditional and marginal mean,

$$h_r\left(\mathbf{x}_j'\boldsymbol{\beta}\right) = \pi_j^r = \mathrm{E}\left(\pi_j^{*r}\right) = \mathrm{E}\left\{h_r\left[f_r\left(\pi_j^r, \sigma^2\right) + \nu_j\right]\right\}. \tag{6.12}$$

In order to completely describe the data-generating process, one would need a model for the dimension of the ARP not measured by the reported data (i.e., a model for ϕ_j if $r = E$ or a model for ζ_j if $r \in \{C, M\}$). The complete data-generating process would then be described by (6.9). However, in the previous section, I avoided committing to a model for this nuisance parameter by introducing an approximation for the variance of the reported datum. Following the same approach, I propose to approximate the variance of Y_j^r conditional on the behavior stream as:

$$\mathrm{Var}\left(Y_j^r | \nu_j\right) = \left(\mathbf{t}_j'\boldsymbol{\gamma}\right) V_r\left(\pi_j^{*r}\right), \tag{6.13}$$

where V_r is a variance function chosen according to the recording procedure, $\mathbf{t}_j = (1 - T_j, T_j)'$, and $\boldsymbol{\gamma}$ is a 2-dimensional vector of dispersion parameters. In the remainder of

[2] I focus on an AR(1) process because it is conventional for analysis of single-case designs. With sufficient data, other dependence structures could certainly also be considered.

this section, I give several specific examples of the general model and discuss effect size definition.

6.2.1. Stable-phase models

As an initial example of the proposed model, consider the stable-phase model $g_r\left(\pi_j^r\right) = \beta_0 + \beta_1 T_j$. Because the covariate takes on only two possible values, the form of f_r is irrelevant and the conditional mean follows a generalized linear model where $g_r\left(\pi_j^{*r}\right) = \beta_0^* + \beta_1^* T_j + \nu_j$, with conditional regression coefficients satisfying

$$h_r\left(\beta_0 + \beta_1 T\right) = \mathrm{E}\left[h_r\left(\beta_0^* + \beta_1^* T + \nu_j\right)\right]$$

for $T = 0, 1$.

6.2.2. Log-linear models for incidence

In the model from Section 6.1, I showed that use of the log link function $g_E(x) = \ln(x)$ with event-counting data leads to regression coefficients that represent proportionate changes in incidence, which is useful for estimating the log-incidence ratio. The log link remains useful in models with serial dependence, only now the coefficients represent proportionate changes in *average* incidence, over realizations of the process. The log link also creates a simple and convenient relationship between the conditional mean specification and the marginal mean specification. With $h_E(x) = \exp(x)$, (6.12) becomes

$$\exp\left(\mathbf{x}_j'\boldsymbol{\beta}\right) = \pi_j^E = \mathrm{E}\left(\zeta_j L\right) = \exp\left[f_E\left(\pi_j^r, \sigma^2\right)\right] \mathrm{E}\left[\exp\left(\nu_j\right)\right],$$

by which it follows that $f_E\left(\pi_j^r, \sigma^2\right) = \mathbf{x}_j'\boldsymbol{\beta} - \sigma^2/2$. Thus, if the marginal mean follows a log-linear model, then so too does the conditional mean. What is more, all of the conditional regression coefficients except for the intercept term are identical to the marginal coefficients.

I have assumed a simple auto-correlation structure for the latent errors $(\nu_1, ..., \nu_n)$ in (6.11), which induces dependence among the recorded data $(Y_1, ..., Y_n)$. However, the pattern of dependence in the latter is more complex. Taking $V_E(x) = x$ and treating the variance model (6.13) as correct, it follows that

$$\text{Cov}\left(Y_j^E, Y_k^E\right) = \pi_j^E \pi_k^E \left[\exp\left(\rho^{|j-k|}\sigma^2\right) - 1\right] + \left(\mathbf{t}_j'\boldsymbol{\gamma}\right)\pi_j^E I(j=k) \tag{6.14}$$

where $I()$ is the indicator function. The correlation in the reported data is then

$$\text{corr}\left(Y_j^E, Y_k^E\right) = \frac{\exp\left(\rho^{|j-k|}\sigma^2\right) - 1}{\sqrt{\left(\frac{\mathbf{t}_j'\boldsymbol{\gamma}}{\pi_j^E} + \exp\left(\sigma^2\right) - 1\right)\left(\frac{\mathbf{t}_k'\boldsymbol{\gamma}}{\pi_k^E} + \exp\left(\sigma^2\right) - 1\right)}},$$

which depends on π_j^E, π_k^E, and $\boldsymbol{\gamma}$ in addition to the variability of the latent errors σ^2 and the latent autocorrelation ρ. Zeger (1988) gives a similar formula under a different parameterization of the model, pointing out that the auto-correlation in the observed data will always be less than that in the latent errors.

6.2.3. Logit-linear models for prevalence

In the models for continuous recording and momentary time sampling described in Section 6.1 (which did not allow for serial dependence), treatment effects are measured using log-prevalence odds ratios, leading to use of the logit link function $g_C(x) = g_M(x) =$

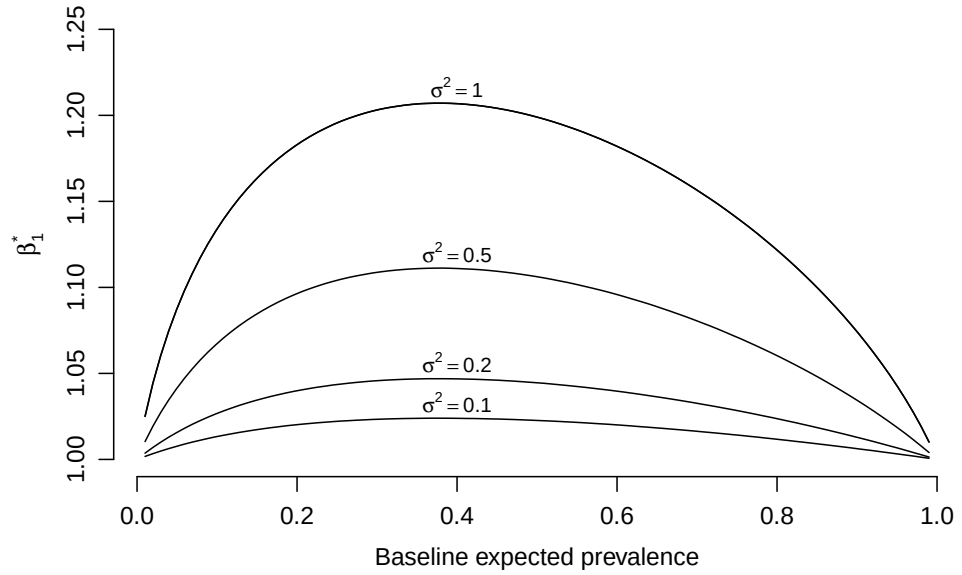


Figure 6.5. Conditional treatment effect β_1^* as a function of baseline expected prevalence β_0 and variability σ^2 in the logistic-linear stable-phase model, fixing marginal treatment effect $\beta_1 = 1$.

$\ln(x) - \ln(1-x)$. Applying the same link to the model with serial dependence leads to treatment effects that are log-odds ratios of the *expected* prevalence over realizations of the data-generating process. Unfortunately, the logit link does not lead to any analytically convenient relationship between the conditional and marginal specifications.

As an illustration of the difference between the conditional and marginal parameters, consider again the stable phase model described in Section 6.2.1, assuming that $g_r(x) = \text{logit}(x)$. The magnitude of the conditional treatment effect depends on both the expected prevalence during baseline (a function of β_0) and on the degree of variability in prevalence (as measured by σ^2), in addition to the marginal treatment effect β_1. Figure 6.5 plots the conditional treatment effect β_1^* as a function of β_0 and σ^2, holding the marginal treatment effect fixed at $\beta_1 = 1$.

A linear model for the log-odds of the marginal mean will generally lead to a model that is non-linear in the log-prevalence odds of the conditional mean (except in the case of the stable phase model, as previously noted). To illustrate, consider a model for a single phase that includes a linear time trend in the marginal mean: $\text{logit}(\pi_j^r) = \beta_0 + \beta_1 j$. Figure 6.6 displays the difference between the conditional log-odds $f_r(\beta_0, \beta_1, j, \sigma^2)$ and the marginal log-odds $\beta_0 + \beta_1 j$ as a function of session j, for several values of β_1 and σ^2. Though the conditional and marginal log-odds clearly differ, the trend in the conditional log-odds remains close to linear except when both β_1 and σ^2 are large. Zeger, Liang, and Albert (1988) suggested an approach that is useful for approximating $f_r(\pi_j^r, \sigma^2)$ in the present context. They noted that approximating $h(x) = 1/(1 + e^{-x})$ by a Gaussian cumulative distribution function leads to a linear approximation for the conditional mean specification:

$$f_r(\pi_j^r, \sigma^2) \approx \mathbf{x}_j' \boldsymbol{\beta} \sqrt{1 + c^2 \sigma^2} = \mathbf{x}_j' \boldsymbol{\beta}^*, \tag{6.15}$$

where $c = 16\sqrt{3}/(15\pi)$.

Unlike in the log-linear model for event-counting, exact expressions for the marginal covariance of the reported data are not available with the logit link. The marginal covariance can be decomposed as

$$\begin{aligned}
\text{Cov}(Y_j^r, Y_k^r) &= \text{Cov}\left[\text{E}(Y_j^r | \nu_j), \text{E}(Y_k^r | \nu_k)\right] + \text{E}\left[\text{Var}(Y_j^r | \nu_j)\right] I(j = k) \\
&= \text{Cov}\left[h_r\left(f_r(\pi_j^r, \sigma^2) + \nu_j\right), h_r\left(f_r(\pi_k^r, \sigma^2) + \nu_k\right)\right] \\
&\qquad + \left(\mathbf{t}_j' \boldsymbol{\gamma}\right) \text{E}\left[V_r\left(\pi_j^{*r}\right)\right] I(j = k).
\end{aligned}$$

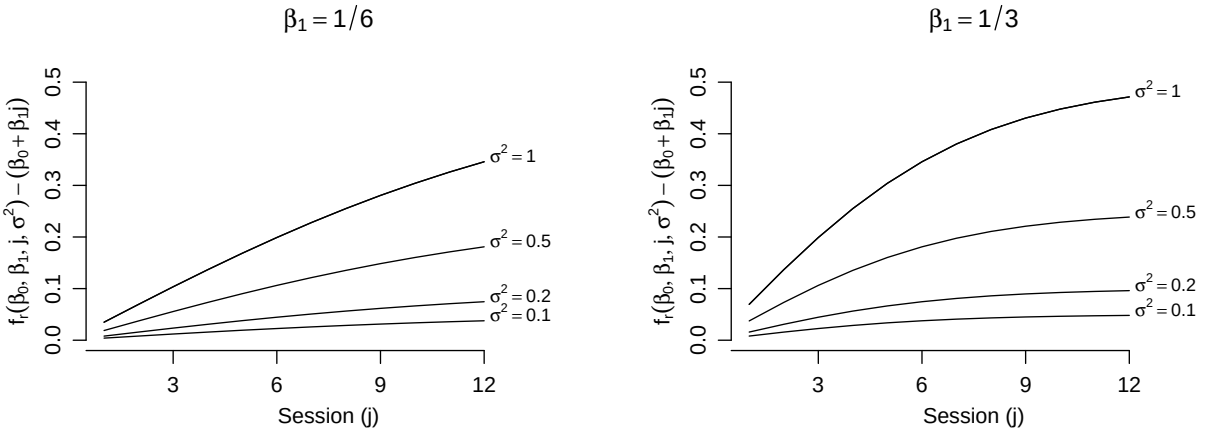


Figure 6.6. Degree of nonlinearity in the logistic-linear trend model, for various values of σ^2 and β_1. Vertical axis corresponds to $f_r(\beta_0, \beta_1, j, \sigma^2) - (\beta_0 + \beta_1 j)$. The initial baseline prevalence is fixed at $\beta_0 = \text{logit}(0.5)$, which leads to the highest possible degree of non-linearity.

Zeger et al. (1988) proposed to approximate the covariance of the conditional expectations using a first-order Taylor series approximation to $h_r(x)$ about the mean of the random effects ν_j, ν_k:

$$\begin{aligned}\text{Cov}\left[\text{E}(Y_j^r|\nu_j), \text{E}(Y_k^r|\nu_k)\right] &\approx \text{Cov}\left[h_r\left(\mathbf{x}_j'\boldsymbol{\beta}^*\right) + h_r^{(1)}(\mathbf{x}_j'\boldsymbol{\beta}^*)\nu_j, , [h_r\left(\mathbf{x}_k'\boldsymbol{\beta}^*\right) + h_r^{(1)}(\mathbf{x}_k'\boldsymbol{\beta}^*)\nu_k\right] \\ &= h_r^{(1)}(\mathbf{x}_j'\boldsymbol{\beta}^*)h_r^{(1)}(\mathbf{x}_k'\boldsymbol{\beta}^*)\rho^{|j-k|}\sigma^2,\end{aligned}$$

where $h^{(p)}$ is the p^{th} derivative of $h(x)$, so $h^{(1)}(x) = h(x)[1 - h(x)]$. Further replacing the conditional mean with the marginal mean in the variance function, the marginal covariance is approximately

$$\text{Cov}(Y_j^r, Y_k^r) \approx h_r^{(1)}(\mathbf{x}_j'\boldsymbol{\beta}^*)h_r^{(1)}(\mathbf{x}_k'\boldsymbol{\beta}^*)\rho^{|j-k|}\sigma^2 + \left(\mathbf{t}_j'\boldsymbol{\gamma}\right) V_r\left(\pi_j^r\right) I(j = k). \tag{6.16}$$

For modeling binary outcome data, Goldstein and Rasbash (1996) suggested using a second-order Taylor series approximation to h_r; this leads to another term in the covariance:

$$\begin{aligned}(6.17)\quad \mathrm{Cov}(Y_j^r, Y_k^r) \approx\ & h_r^{(1)}(\mathbf{x}_j'\boldsymbol{\beta}^*)h_r^{(1)}(\mathbf{x}_k'\boldsymbol{\beta}^*)\rho^{|j-k|}\sigma^2 \\ & + \frac{1}{4}h_r^{(2)}(\mathbf{x}_j'\boldsymbol{\beta}^*)h_r^{(2)}(\mathbf{x}_k'\boldsymbol{\beta}^*)\left(1+2\rho^{2|j-k|}\right)\sigma^4 \\ & + \left(\mathbf{t}_j'\boldsymbol{\gamma}\right)V_r\left(\pi_j^r\right)I(j=k),\end{aligned}$$

where $h^{(2)}(x) = h(x)[1-h(x)][1-2h(x)]$. Though these approximations are ad-hoc, they may nonetheless be reasonable for relatively simple between-phase models. Figure 6.7 plots the exact marginal auto-correlation of continuous recording data based on simulating from a stable-phase model in which $\mathrm{E}\left(Y_1^C\right) = \cdots = \mathrm{E}\left(Y_n^C\right) = \pi^C$, for various values of π^C and latent auto-correlation ρ. The exact auto-correlations fit the first-order approximation remarkably well.

6.2.4. Effect size definition

In models where the parameters of the behavior stream process are allowed to vary stochastically over time, marginal treatment effect sizes no longer have exact, design-comparable interpretations. The effect sizes defined in Section 5.2 must therefore be revisited. As before, interest is in comparing a parameter of the behavior stream under two different conditions, indicated by $t = 0, 1$. Under model (6.11), event counting data under condition t measures ζ_t, which is itself a random quantity, measuring its expectation $\pi_t^E = \mathrm{E}(\zeta_t)$.

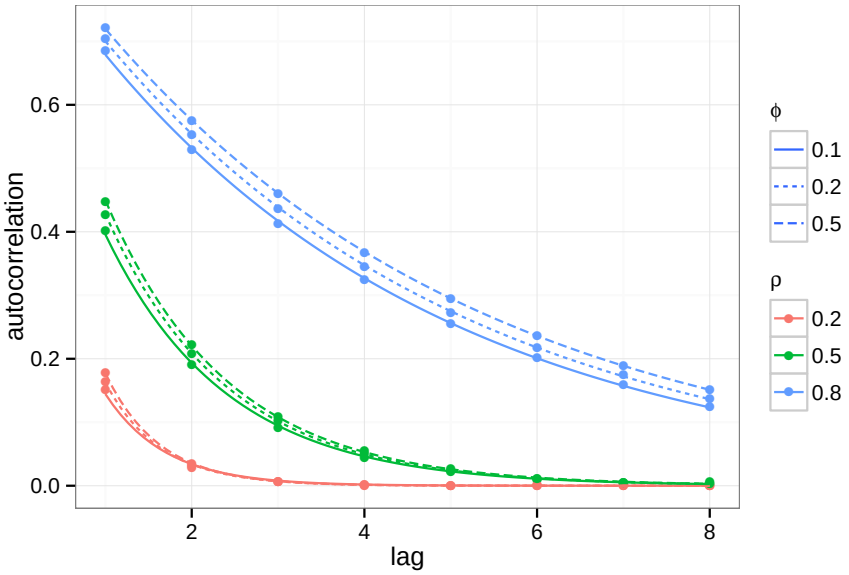


Figure 6.7. Marginal auto-correlation of continuous recording data in the stable-phase model, for varying values of latent auto-correlation ρ and expected prevalence ϕ. Dots represent exact auto-correlation based on simulated data with $\zeta = 20$ and $\sigma^2 = 1$. Lines are fitted curves corresponding to (6.16).

The marginal log-incidence ratio is then defined as $\omega^\zeta = \ln\left(\pi_1^E/\pi_0^E\right)$. Continuous recording and momentary time sampling data measure ϕ_t, which is itself a random quantity, measuring its expectation $\pi_t^C = \pi_t^M = \mathrm{E}(\phi_t)$. The marginal log-prevalence odds ratio is then defined as

$$\psi = \ln\left[\frac{\pi_1^C(1-\pi_0^C)}{(1-\pi_1^C)\pi_0^C}\right] = \ln\left[\frac{\pi_1^M(1-\pi_0^M)}{(1-\pi_1^M)\pi_0^M}\right],$$

Other marginal effect sizes such as log-prevalence ratios are defined similarly. Note that if there is no variability in the parameters of the behavior stream process, these definitions reduce to those given in Section 5.2. In Section 5.2.5, I noted various conditions under which different effect size measures are comparable. These conditions no longer lead to exact measurement-comparability if there is extra variability in the parameters of the

behavior stream. Still, they might serve as approximate guides, especially when the variability in the behavior stream parameters is not too large.

6.3. Estimation for models with serial dependence

This section examines two methods of estimating the marginally-specified model for serially dependent measurements of free-operant behavior as given in (6.10)-(6.13). Having argued that the marginal parameters in (6.10) are of primary interest as measures of effect size, I focus on an estimating-equation approach that maintains a certain robustness to misspecification of the other model components (i.e., those involving second moments). Denote $\mathbf{Y}^r = (Y_1^r, ..., Y_n^r)'$, $\boldsymbol{\pi}_r = (\pi_1^r, ..., \pi_n^r)'$, $\boldsymbol{\Sigma}_r = \mathrm{Cov}\left(\mathbf{Y}^r\right)$, $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)'$, and $\mathbf{D}_r(\boldsymbol{\eta}) = \mathrm{diag}\left(dh_r\left(\eta_1\right)/d\eta, ..., dh_r\left(\eta_n\right)/d\eta\right)$ for a given $n \times 1$ vector $\boldsymbol{\eta}$ with components $\eta_1, ..., \eta_n$. Collect the parameters describing the covariance of $\mathbf{Y}^r$ in $\boldsymbol{\theta} = (\boldsymbol{\gamma}', \sigma^2, \rho)'$; these will be treated as nuisance parameters. An estimator of $\boldsymbol{\beta}$ can be defined as the solution to the linear, unbiased estimating equation

$$\mathbf{U}_3(\boldsymbol{\beta}, \boldsymbol{\theta}) = \mathbf{X}'\mathbf{D}_r\left(\mathbf{X}\boldsymbol{\beta}\right)\mathbf{W}(\boldsymbol{\beta}, \boldsymbol{\theta})\left[\mathbf{Y}^r - h_r\left(\mathbf{X}\boldsymbol{\beta}\right)\right] = \mathbf{0} \tag{6.18}$$

for some $n \times n$ matrix $\mathbf{W}(\boldsymbol{\beta}, \boldsymbol{\theta})$ that may depend on the nuisance parameters; let $\hat{\boldsymbol{\beta}}$ be the solution to (6.18). If $\hat{\boldsymbol{\beta}}$ is asymptotically consistent, then its variance is approximately

$$\mathrm{Cov}(\hat{\boldsymbol{\beta}}) \approx \left(\mathbf{X}'\mathbf{D}_r\mathbf{W}\mathbf{D}_r\mathbf{X}\right)^{-1}\left(\mathbf{X}'\mathbf{D}_r\mathbf{W}\boldsymbol{\Sigma}_r\mathbf{W}\mathbf{D}_r\mathbf{X}\right)\left(\mathbf{X}'\mathbf{D}_r\mathbf{W}\mathbf{D}_r\mathbf{X}\right)^{-1}. \tag{6.19}$$

The most asymptotically efficient choice for $\mathbf{W}$ is to use the inverse of the covariance of the observations $\boldsymbol{\Sigma}^{-1}$, in which case the estimator's variance reduces to $\left(\mathbf{X}'\mathbf{D}_r\boldsymbol{\Sigma}^{-1}\mathbf{D}_r\mathbf{X}\right)^{-1}$

(see McCullagh & Nelder, 1989, Section 9.5).[3] Of course, even if an exact form for $\mathbf{\Sigma}$ can be derived, it will involve the unknown parameters $\boldsymbol{\theta}$. The problems then are how to choose $\mathbf{W}$ without sacrificing too much efficiency and how to estimate the variance of $\hat{\boldsymbol{\beta}}$.

6.3.1. Ignoring serial dependence

One possible approach to estimating $\boldsymbol{\beta}$ is simply to ignore the serial dependence structure. Observe that if $\mathbf{W}$ is taken to be diagonal with non-zero entries $W_{jj} = \left[(\mathbf{t}_j'\boldsymbol{\gamma})V_r(h_r(\mathbf{x}_j'\boldsymbol{\beta}))\right]^{-1}$, then (6.18) reduces to the quasi-likelihood estimating equation for independent data, as given in (6.4); denote this estimator $\hat{\boldsymbol{\beta}}_I$. Davis et al. (2000) studied this approach for a log-linear model like the one I have described for event counting data. They showed that, provided the covariate matrix $\mathbf{X}$ satisfies certain conditions, the quasi-likelihood estimator based on assuming independence of repeated measurements is asymptotically consistent and normally distributed even if the data actually exhibit dependence due to a latent, stationary process. Whether a similar result holds for logit-linear models such as those proposed for continuous recording and momentary time sampling data is an important question for future research; I defer it for now because asymptotic results are not my primary concern. Instead, I investigate the performance of $\hat{\boldsymbol{\beta}}_I$ using simulation, under a simple model for continuous recording data.

[3] In the context of models for repeated measures on many individual cases, (6.18) is referred to as a generalized estimating equation (GEE). Liang and Zeger (1986) proposed the use of GEE for longitudinal data analysis and showed that, if $\mathbf{Y}^r$ can be partitioned into independent sub-vectors, (6.18) leads to consistent, asymptotically normal estimators for the marginal regression parameters even if $\mathbf{W}$ is not chosen optimally, or is based on rough approximations to $\mathbf{\Sigma}$. They showed further that the variance of $\hat{\boldsymbol{\beta}}$ can be estimated consistently using sandwich-type estimators based on (6.19). Unfortunately, these results are predicated on having multiple, independent cases, but in the present context I am concerned with methods for modeling a single case.

Consider a model for $\mathbf{Y}^C$ in which the marginal expected prevalence is constant, with logit $\left(\pi_j^C\right) = \beta$ for $j = 1, ..., n$. Assume that the conditional mean includes a latent error $\pi_j^{*C} = \beta^* + \nu_j$, where $\nu_1, ..., \nu_n$ are normally distributed and follow an AR(1) dependence model with variance σ^2 and auto-correlation ρ. To complete the data-generating model, assume that the reported datum from session j is based on a realization from a behavior stream that follows an alternating Poisson process with prevalence $\phi_j = \pi_j^{*C}$ and constant incidence ζ:

$$\left(Y_j^C | \phi_j, \zeta\right) \overset{\text{iid}}{\sim} M_C \left[APP(\phi_j, \zeta)\right].$$

Under this model, the estimator β_I is equivalent to the logit of the sample mean, as used in (5.10). The optimal (minimum-variance) estimator would use a weighted mean instead of the simple arithmetic mean; denote this statistic $\hat{\beta}_{opt} = \text{logit}\left[\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{Y} / \left(\mathbf{1}'\mathbf{\Sigma}^{-1}\mathbf{1}\right)\right]$.

I use simulation to compare the efficiency of β_I to the theoretical ideal β_{opt}, for varying levels of π^C, ζ, σ, ρ, and n. The simulation methods and results are described fully in Appendix Section C.4. I define relative efficiency as $\sqrt{\text{E}[(\hat{\beta}_{opt} - \beta)^2] / \text{E}[(\hat{\beta}_I - \beta)^2]}$. Figure 6.8 presents a typical result with $\pi^C = 0.5$ and $\zeta = 20$; the relative efficiency of the β_I versus β_{opt} is plotted versus sample size n, for varying levels of latent variability σ^2 and auto-correlation ρ. The independence estimator β_I is over 95% efficient except when the latent errors are highly dependent. Even at the highest level of auto-correlation considered ($\rho = 0.8$), the independence estimator is over 90% efficient, though the relative efficiency decreases as sample size increases and as the latent variability σ^2 increases. Based on analogous simulations, similar patterns hold for momentary time sampling and event counting data (See Appendix Sections C.4 and C.5, respectively). Thus, if one's only goal

is to obtain an estimate of the marginal regression coefficients, ignoring possible serial dependence may be less foolhardy that it initially seems.

The primary drawback of ignoring serial dependence is that both the robust estimator (6.7) and the model-based estimator (6.8) will be highly inaccurate, underestimating the variance of $\boldsymbol{\beta}_I$ to an extent that depends on the latent variability and autocorrelation. Still, the point estimate need not be discarded; for purposes of meta-analysis, it could be used validly in combination with empirical variance estimation. This may be particularly useful if the meta-analyst must collect effect size estimates based on reported analyses, rather than raw data.[4] For primary analysis though, a principled approach to variance estimation is required.

6.3.2. Estimating serial dependence

Another approach to estimation involves estimating the nuisance parameters $\boldsymbol{\theta}$ that describe the serial dependence structure of $\mathbf{Y}^r$. An estimate of $\boldsymbol{\Sigma}(\boldsymbol{\theta})$ can then be used either to better estimate the variance of $\hat{\boldsymbol{\beta}}_I$ or to form the weighting matrix $\mathbf{W}$ in (6.18) and re-estimate the effect size and its variance. Recall that (6.14) gives an exact form for $\boldsymbol{\Sigma}$ under the log-linear model for event counting data described in Section 6.2.2. For the logit-linear model for measures of prevalence, approximations for $\boldsymbol{\Sigma}$ are given in (6.16) and (6.17).

[4]For purposes of effect size weighting, one could use a rough adjustment to the reported variance based on a prior value of the autocorrelation. For example, the model-based variance estimator V_M from a stable-phase model underestimates the true variance proportionally to

$$F = 1 - \frac{2}{n}\sum_{j=1}^{n-1}(n-j)\rho^j.$$

A rough adjustment to the variance is therefore given by $V_m^* = V_m/F$.

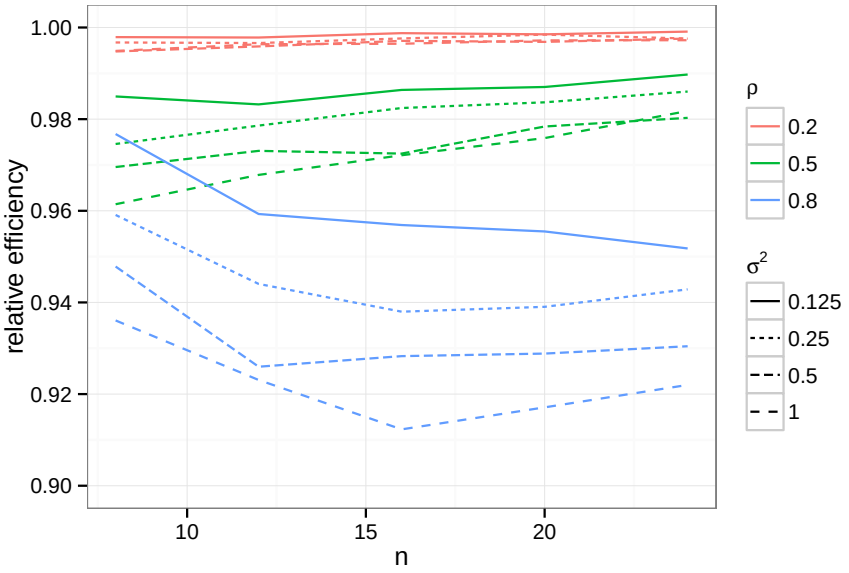


Figure 6.8. Relative efficiency of $\hat{\beta}_I$ versus $\hat{\beta}_{opt}$ based on continuous recording data in the stable-phase model, for varying values of latent variability σ^2 and latent auto-correlation ρ. Expected prevalence is fixed at $\pi^C = 0.5$, incidence is fixed at $\zeta = 20$.

Zeger (1988) proposed moment estimators of the nuisance parameters in log linear models of count data; Davis et al. (2000) proposed bias-corrected moment estimators motivated by the assumption that the conditional distribution of $\mathbf{Y}^r$ is Poisson. In the context of GEE models for multiple individual cases, a variety of other approaches to nuisance parameter estimation are reviewed by Lipsitz and Fitzmaurice (2008). Recently, Gaussian estimating equations have been proposed for estimation and inference on nuisance parameters in longitudinal models (Hall & Severini, 1998; Lipsitz, Molenberghs, Fitzmaurice, & Ibrahim, 2000; Wang & Carey, 2004). Heuristically, this method is equivalent to assuming that the outcomes $\mathbf{Y}^r$ follow a multivariate normal distribution for purposes of estimating the covariance matrix; the estimators are defined as the maximizers of the Gaussian (pseudo)-likelihood. Wang and Carey (2004) demonstrated that

the Gaussian estimating equations are unbiased; the parameter estimators are therefore asymptotically consistent when based on data from multiple, independent cases. Gaussian estimating equations also have the advantage that the estimators remain well-defined when outcome measurements are missing or irregularly spaced (Lipsitz et al., 2000).

For fixed $\boldsymbol{\beta}$, the maximizers of the Gaussian pseudo-likelihood are equivalent to the solution of the pseudo-score equations $\mathbf{U}_4(\boldsymbol{\beta}, \boldsymbol{\theta})$, where

$$U_{4i}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \text{tr}\left(\frac{\partial \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \theta_i}\left[(\mathbf{Y}^r - \mathbf{X}\boldsymbol{\beta})(\mathbf{Y}^r - \mathbf{X}\boldsymbol{\beta})' - \boldsymbol{\Sigma}(\boldsymbol{\beta}, \boldsymbol{\theta})\right]\right) = 0 \tag{6.20}$$

for $i = 1, ..., 4$. It is clear that (6.20) is unbiased when $\boldsymbol{\Sigma}$ is correctly modeled. Let $\hat{\boldsymbol{\theta}}_I$ be the solution of (6.20) with $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}_I$. An estimator for the variance of $\hat{\boldsymbol{\beta}}_I$ can then be calculated as

$$C(\hat{\boldsymbol{\beta}}_I) = (\mathbf{X}'\mathbf{D}_r\mathbf{D}_r\mathbf{X})^{-1}\mathbf{X}'\mathbf{D}_r\boldsymbol{\Sigma}_r\left(\hat{\boldsymbol{\beta}}_I, \hat{\boldsymbol{\theta}}_I\right)\mathbf{D}_r\mathbf{X}(\mathbf{X}'\mathbf{D}_r\mathbf{D}_r\mathbf{X})^{-1}. \tag{6.21}$$

Instead of using $\hat{\boldsymbol{\beta}}_I$, one could instead define estimators for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ as the joint solution of the estimating equations given in (6.18) and (6.20) with $\mathbf{W} = \boldsymbol{\Sigma}^{-1}$; denote these $\hat{\boldsymbol{\beta}}_P$ and $\hat{\boldsymbol{\theta}}_P$.[5] The estimators can be obtained by iterating between solution of (6.18) for fixed $\boldsymbol{\theta}$ and (6.20) for fixed $\boldsymbol{\beta}$. The covariance of $\hat{\boldsymbol{\beta}}_P$ can then be estimated using:

$$\mathbf{C}(\hat{\boldsymbol{\beta}}_P) = \left(\mathbf{X}'\hat{\mathbf{D}}_r\hat{\boldsymbol{\Sigma}}^{-1}\hat{\mathbf{D}}_r\mathbf{X}\right)^{-1}, \tag{6.22}$$

where $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}\left(\hat{\boldsymbol{\beta}}_P, \hat{\boldsymbol{\theta}}_P\right)$ and $\hat{\mathbf{D}} = \mathbf{D}\left(\mathbf{X}\hat{\boldsymbol{\beta}}_P\right)$.

[5] An interesting question for future research is whether $\hat{\boldsymbol{\theta}}$ is asymptotically consistent for $\boldsymbol{\theta}$ (as $n \to \infty$) under the AR1 serial dependence structure that I have described.

Breslow and Clayton (1993) suggested using nuisance parameter estimators based the restricted Gaussian likelihood rather than the full Gaussian likelihood. It seems reasonable to consider this alternative in the present context, given the small sample sizes typical of single-case research and in keeping with the use of restricted likelihood for estimation of linear mixed models in Chapter 3. This restricted Gaussian likelihood leads to use of the estimating equation $\mathbf{U}_5(\boldsymbol{\beta}, \boldsymbol{\theta}) = 0$, where

$$U_{5i}(\boldsymbol{\beta}, \boldsymbol{\theta}) = U_{4i}(\boldsymbol{\beta}, \boldsymbol{\theta}) - \text{tr}\left[\frac{\partial \boldsymbol{\Sigma}^{-1}(\boldsymbol{\beta}, \boldsymbol{\theta})}{\partial \theta_i}\mathbf{D}\mathbf{X}\left(\mathbf{X}'\mathbf{D}\boldsymbol{\Sigma}^{-1}\mathbf{D}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{D}\right], \tag{6.23}$$

in place of (6.20). Define the estimators $\hat{\boldsymbol{\beta}}_R$ and $\hat{\boldsymbol{\theta}}_R$ as the solution to the estimating equations given in (6.18) and (6.23). The variance of $\hat{\boldsymbol{\beta}}_R$ can be estimated using (6.22) with $\hat{\boldsymbol{\Sigma}}$ and $\hat{\mathbf{D}}$ evaluated at $\hat{\boldsymbol{\beta}}_R, \hat{\boldsymbol{\theta}}_R$. Alternately, the variance of $\hat{\boldsymbol{\beta}}_I$ can be estimated by evaluating (6.21) with $\hat{\boldsymbol{\beta}}_R, \hat{\boldsymbol{\theta}}_R$.

6.3.3. Small-sample performance

I have described two methods for estimating effect sizes from of serially correlated measurements of free-operant behavior. The first is to estimate the marginal mean regression as if the data were independent, then estimate the dependence structure of the data for purposes of variance estimation alone. The second involves iteratively solving two estimating equations in order to refine the estimate of the marginal mean regression. Due to the computational intensity required to solve the estimating equation for the nuisance parameters, I do not evaluate the latter procedure directly. Instead, I focus on the performance of estimators for the variance of the independence estimator $\hat{\boldsymbol{\beta}}_I$, which can be

calculated based on only a single evaluation of the estimating equation for the nuisance parameters.

I conducted several simulation studies to evaluate various approaches to nuisance parameter estimation when the data are based on event counting, continuous recording, or momentary time sampling procedures. I used a data generating model with a constant marginal mean structure in order to focus on the effects of varying levels of latent variability, serial dependence, and sample size. The simulation designs and results are described fully in Appendix C. Here I highlight the main results.

For the log-linear model of event counting data, an exact covariance structure can be derived based on the posited latent error structure. Two possible approaches to estimating the parameters of the covariance structure are to use the Gaussian pseudo-likelihood score equation in (6.20) or to use the restricted version in (6.23). Based on simulations reported in Appendix C.5, the restricted pseudo-likelihood equation produces badly biased variance estimates, and should not be used. The full pseudo-likelihood score equation produces a better variance estimator (the FML estimator), though one that tends to be biased downward except for relatively long series lengths. Figure 6.9 plots the average relative bias of the FML variance estimator versus sample size, for varying levels of latent variability σ^2 and autocorrelation ρ. At the low level of autocorrelation, the FML variance estimator has reasonably small bias for $n \geq 16$; however, for $\rho = 0.5$, a series of length $n = 24$ is required in order to a variance estimate that is close to unbiased. For very high levels of autocorrelation, the true variance is underestimated even for $n = 24$. The bias of the FML variance estimator is driven by biases in the estimates of σ^2 and ρ, both of which tend to be underestimated. Thus, fairly large sample sizes will be required for

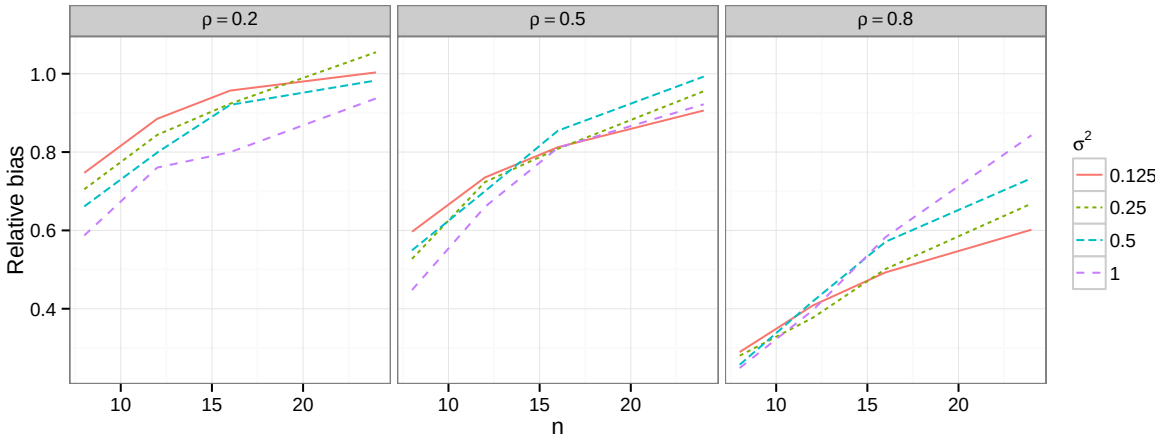


Figure 6.9. Average relative bias of the FML variance estimator based on serial dependence model for event-counting data.

accurate variance estimation, even in a model with the simplest possible marginal mean structure.

For logit-linear models of continuous recording and momentary time sampling data, estimation relies on an approximation for the covariance structure, rather than exact analytic expressions. One would expect the accuracy of variance estimation to be affected by the quality of the approximation, but the actual relationship appears to be more complicated. Based on simulations reported in Appendix C.4, FML estimators based on first-order approximation appears to be preferable for both continuous recording and momentary time sampling data. Figure 6.10 plots the average relative bias of the FML variance estimators based on 1st- and 2nd-order approximations to the covariance, for varying levels of σ^2 and ρ. On average, the first-order approximation has a downward bias, whereas the second-order approximation has an upward bias for larger sample sizes. The 1st-order approximation has lower root mean-squared error than the second-order approximation, and so may be preferred on that basis. The RML estimating equations

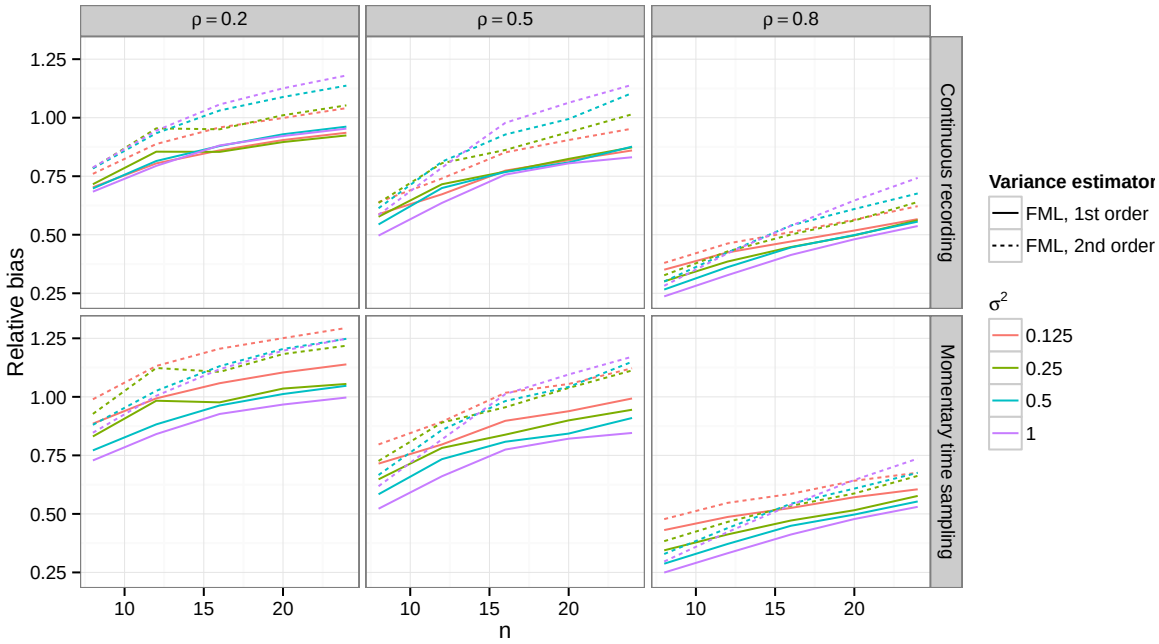


Figure 6.10. Average relative bias of the 1st- and 2nd-order FML variance estimators based on a serial dependence model for continuous recording or momentary time sampling data.

based on first- and second-order approximations both produce badly biased variance estimators; the RML equations should therefore not be used in this context. Just as with event counting data, fairly large sample sizes will be required to recover adequate variance estimates, but even these will be approximate at best. For the logit-linear model, it may be that a different approach to approximating the covariance matrix could lead to better variance estimators; this should be investigated in future work.

6.4. Applications

This section presents two applications of the methods examined in this chapter to actual single-case studies. Each study used a multiple baseline design with several cases,

but I perform a case-by-case analysis, in keeping with the conventional logic of single-case designs.

6.4.1. Ross & Horner (2009)

Ross and Horner (2009) used a multiple baseline design across individuals to evaluate the effect of a school-wide bullying prevention program on six students (two in each of three schools) who had been previously identified as engaging in high rates of aggression towards other students.[6] For each case, investigators collected repeated measurements before, during, and after implementation of the school-wide program, though for analytic purposes I do not use the measurements during the four- or five-day acquisition phase. The outcome measure is the number of incidents of physical and verbal aggression engaged in by each case, measured via direct observation during lunch recess every weekday. I obtained the raw data, including calendar dates of each measurement, directly from the primary authors. Figure 6.11 plots the raw data from the study.

Compared to typical single-case designs, this study used a comparatively large number of measurements per phase. Baseline phases lasted between 17 and 56 calendar days, during which between 12 and 31 measurements were collected on each case. Full implementation phases continued for between 17 and 52 calendar days, during which between 12 and 27 measurements were collected on each case. This is therefore an example where the serial dependence models described in Section 6.3 might be expected to perform adequately, though some of the phases are shorter than would be ideal. Before turning to

[6] In Section 2.2, I identified this study as an example of a design where the minimum natural level of treatment assignment is the school, rather than the individual case. This feature of the study has implications for identifying design-comparable effect sizes. However, case-level effect sizes can still be described if this feature is ignored; I do in the present analysis for purposes of simplicity.

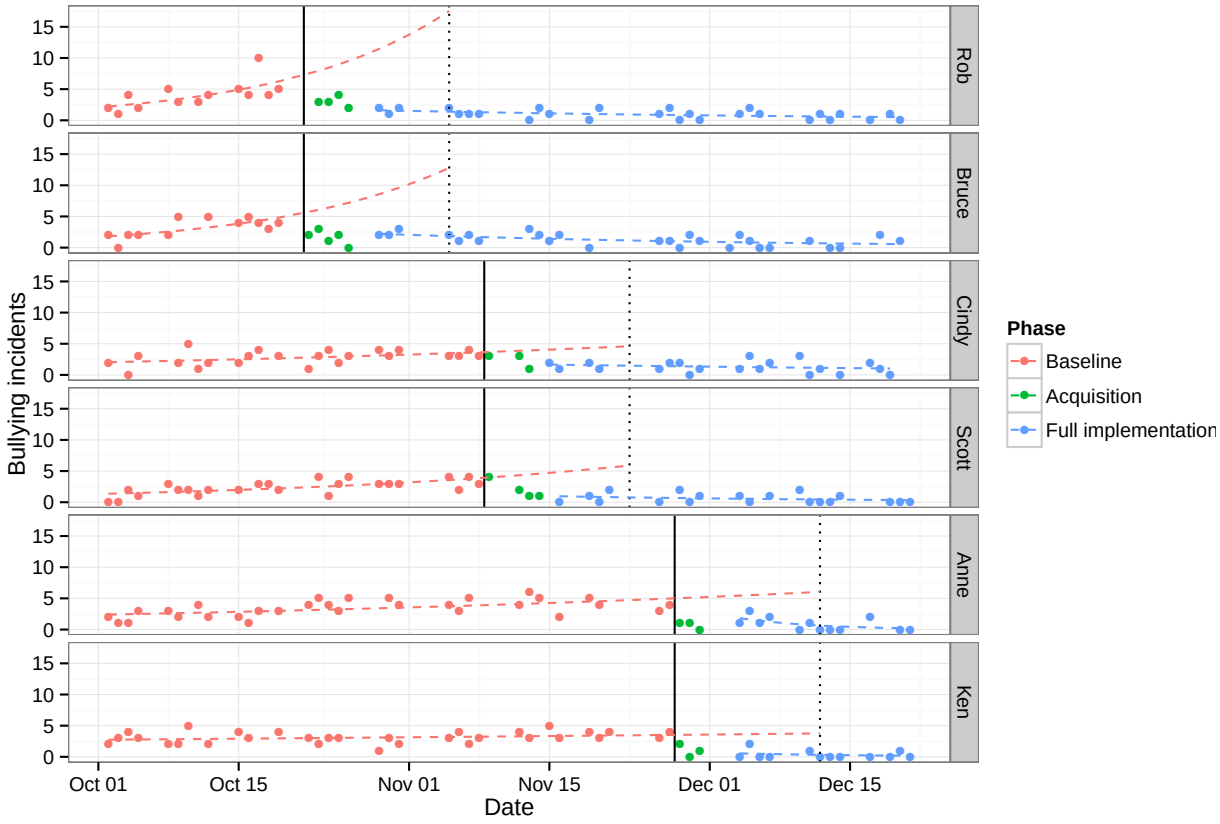


Figure 6.11. Data from multiple baseline design reported by Ross and Horner (2009). Solid vertical lines indicate the introduction of the intervention; dotted vertical lines indicate the point at which the effect size is assessed. Dashed lines depict a log-linear regression fit for each phase.

serial dependence, I report an analysis based on the assumption that repeated measurements are independent.

I focus on a specific target effect size that measures the effect of the intervention two weeks (14 calendar days) after the beginning of the acquisition phase. This target date is indicated by a dotted vertical line in Figure 6.11. Two weeks after the beginning of acquisition, the intervention has been fully implemented for at least one full school week. I use the log-incidence ratio as the effect size metric, which can be interpreted as the log of the ratio of the expected number of bullying incidents if the intervention is in effect to

the expected number of bullying incidents if the intervention had not been introduced, both for a given case and at the specified point in time.

Let Y_{ij} denote the j^{th} outcome measurements from case i, let $\zeta_{ij} = \mathrm{E}(Y_{ij})$, and let d_{ij} denote the calendar date when Y_{ij} was measured, all for $i = 1, ..., 6$ and $j = 1, ..., n_i$ (excluding data from acquisition phases). Let $\tilde{d}_i$ be the calendar date of the first observation during the acquisition phase for case i. I assume that the measurements follow the model

$$\ln(\zeta_{ij}) = \beta_{0i} + \beta_{1i}\left(d_{ij} - \tilde{d}_i\right) + \beta_{2i} I\left(d_{ij} \geq \tilde{d}_i\right) + \beta_{3i}\left(d_{ij} - \tilde{d}_i\right) \times I\left(d_{ij} \geq \tilde{d}_i\right), \tag{6.24}$$

where $I()$ is the usual indicator function. This model accounts for natural changes in the incidence of bullying by case i through the inclusion of log-linear time trends during the baseline phase; it further allows that the log-linear trend in incidence may be affected by intervention. Collect the covariates for the j^{th} measurement on case i in the vector $\mathbf{x}_{ij}$, and let $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \beta_{2i}, \beta_{3i})'$. The target effect size for case i is then $\omega_i^{\zeta} = \beta_{2i} + 14\beta_{3i}$. Note also that the log-incidence ratio can be used to calculate a proportionate reduction in incidence as $\exp\left(\omega_i^{\zeta}\right) - 1$.

I will examine two different models for the variance of Y_{ij}, each of which assumes that the variance of the outcomes is proportional to the mean. The first model assumes that the over- or under-dispersion of the outcomes is constant across phases:

$$\mathrm{Var}(Y_{ij}) = \gamma_i \zeta_{ij}. \tag{6.25}$$

This model is analogous to using the pooled standard deviation for estimating a standardized mean difference between groups, and so I label it as "pooled dispersion." The

second model allow for the possibility that the intervention may affect the dispersion of the outcome (for instance, by changing the mean duration of each bullying incident):

$$\text{Var}\left(Y_{ij}\right) = \left[\gamma_{0i} I\left(d_{ij} < \tilde{d}_i\right) + \gamma_{1i} I\left(d_{ij} \geq \tilde{d}_i\right)\right] \zeta_{ij}. \tag{6.26}$$

In the latter, "separate dispersion" model, γ_{0i} measures dispersion during the baseline phase and γ_{1i} measures dispersion during the full implementation phase, both for case i. The assumptions of these variance models affect how the model-based variance of the target effect size is estimated, though not the point estimates. For completeness, I also report robust variance estimates, as given in (6.7), even though simulation evidence suggests that these may be less precise than the model-based estimates except under gross violations of the variance model.

Table 6.2 reports effect size estimates and standard errors, calculated based on the methods described in Section 6.1.1 and 6.1.3, respectively. The dotted lines in Figure depict the fit of the log-linear regression. Point-estimates for the log-incidence ratio range from -1.13 to -2.54. These effect size estimates correspond to proportionate reductions in the incidence of bullying behavior of between 68% and 92%, two weeks after the start of the intervention. All of the log-incidence ratios are estimated with high precision, regardless of which variance estimator is used. In the model allowing for separate dispersions across phases, there is some indication that dispersion is higher in the full implementation phase, particularly for Scott, Ann, and Ken. As a result, the variance estimates for the "separate dispersion" model are somewhat larger for the final three cases. The robust variance estimates are quite close to the model-based estimates.[7]

[7]Given the similarity of the pooled dispersion estimates, an analyst might consider pooling them across cases (i.e., assuming $\gamma_1 = \cdots = \gamma_6$), or pooling across cases but allowing separate dispersions for each

Table 6.2. Log-incidence ratio effect size estimates for cases in Ross & Horner (2009), assuming independence of repeated measurements.

		Pooled dispersion		Separate dispersion			
Case	Effect size ($\hat{\omega}^{\zeta}$)	V_M	$\hat{\gamma}$	V_M	$\hat{\gamma}_0$	$\hat{\gamma}_1$	V_R
Rob	-2.54	0.26	0.60	0.30	0.71	0.55	0.36
Bruce	-1.94	0.32	0.62	0.31	0.59	0.63	0.22
Cindy	-1.13	0.12	0.59	0.11	0.48	0.71	0.06
Scott	-2.03	0.19	0.68	0.21	0.41	1.02	0.21
Anne	-2.26	0.16	0.66	0.26	0.41	1.38	0.33
Ken	-2.47	0.18	0.52	0.36	0.28	1.22	0.34

Based on a random-effects meta-analysis of the effect size estimates and the separate dispersion model variance estimates, the average log-incidence ratio two weeks after the start of intervention is -2.02, with a 95% confidence interval of (-2.46,-1.46); this average corresponds to a proportionate reduction of between 77% and 92%.[8] The variability of true effect sizes is estimated as $\hat{\tau}^2 = 0.16$ ($I^2 = 40\%$); the Q test for heterogeneity is not statistically significant ($p = 0.16$), though this test has very low power given only six effect size estimates. Note that the random-effects meta-analysis assumes that effect size estimates are independent, which may not be valid given that pairs of cases come from the same school.

The analysis so far has proceeded under the assumption that repeated measures are independent, but this assumption should be scrutinized. Turning now to models that allow for serial dependence, I consider the same mean specification as given in (6.24), but

phase. Both of these approaches lead to variance estimates that are very similar to V_M from the pooled dispersion model.

[8]Using alternative estimates of the effect size variances for inverse-variance weighting leads to very similar results. Fixed-effects meta-analysis also leads to very similar results.

Table 6.3. Effect size estimates, variance estimates, and pseudo-maximum likelihood estimates of nuisance parameters for cases from Ross & Horner (2009), assuming serial dependence of repeated measurements.

	Pooled dispersion					Separate dispersion					
Case	$\hat{\omega}^{\zeta}$	V_M	$\hat{\sigma}^2$	$\hat{\rho}$	$\hat{\gamma}$	$\hat{\omega}^{\zeta}$	V_M	$\hat{\sigma}^2$	$\hat{\rho}$	$\hat{\gamma}_0$	$\hat{\gamma}_1$
Rob	-2.54	0.20	0.12	-0.72	0.28	-2.64	0.13	0.15	-0.60	0.00	0.29
Bruce	-1.94	0.29	0.00	-0.34	0.56	-1.94	0.26	0.00	0.86	0.49	0.58
Cindy	-1.13	0.11	0.00	-0.71	0.53	-1.11	0.09	0.08	-0.45	0.22	0.52
Scott	-2.03	0.18	0.00	-1.00	0.62	-2.03	0.18	0.01	-0.81	0.37	0.90
Anne	-2.26	0.14	0.00	1.00	0.60	-2.24	0.24	0.03	0.70	0.28	1.14
Ken	-2.47	0.17	0.00	-1.00	0.47	-2.46	0.30	0.02	-0.33	0.21	1.00

now allow for a serially correlated latent error term such that

$$\ln\left[\mathrm{E}\left(Y_{ij}|\nu_{ij}\right)\right] = \mathbf{x}_{ij}'\boldsymbol{\beta}_i - \sigma_i^2 + \nu_{ij},$$

where $\nu_{ij} \sim N(0, \sigma_i^2)$ and $(\nu_{i1}, ..., \nu_{in_i})$ follow an AR1 serial dependence model with autocorrelation ρ_i. I consider the same two specifications for the variance model as used previously (pooled dispersion or separate dispersion). I also fit the model assuming that all cases share common nuisance parameters, so that $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \cdots = \boldsymbol{\theta}_6$, where $\boldsymbol{\theta}_i = (\sigma_i^2, \rho_i^2, \gamma_{0i}, \gamma_{1i})$.

Table 6.3 reports estimates of effect sizes $\hat{\omega}^{\zeta}$, variances V_M, and nuisance parameters, all obtained through iterative fitting of the marginal mean estimating equation $\mathbf{U}_3(\boldsymbol{\beta}, \boldsymbol{\theta})$ and the Gaussian full pseudo-likelihood estimating equation $\mathbf{U}_4(\boldsymbol{\beta}, \boldsymbol{\theta})$, as described in Section 6.3.[9] The effect size estimates from the pooled dispersion model are the same to two decimals as the estimates assuming independence of repeated measurements. This is because the latent variability parameters are estimated to be near zero for all but one

[9]The parameter estimates obtained from the initial fit of $\mathbf{U}_4(\boldsymbol{\beta}_I, \boldsymbol{\theta})$ are very similar. In most cases, convergence occurred after two to four iterations.

case, and so the estimated weighting matrix is very close to the identity matrix. The effect size estimates from the separate dispersion model are also very similar to the earlier estimates, and for the same reason. Variance estimates are uniformly smaller than the estimates based on assuming independence, likely driven by the use of full pseudo-likelihood for estimation of the dispersion parameters. Consequently, the effect size estimates and variances reported in Table 6.2 should be preferred, and I do not report a meta-analysis of the estimates from the serial dependence model.

For cases with $\hat{\sigma}^2 \approx 0$, the estimated autocorrelations are not meaningful because the latent errors are negligible. For the first case (Rob), the latent variability estimate is small but the autocorrelation estimate is strongly negative, which leads to estimates of the effect size variance that are smaller than when independence is assumed.[10] The estimated autocorrelation for this case should be interpreted cautiously in light of the simulation evidence, which indicated that the autocorrelation estimates have a negative bias when based on short series. When the nuisance parameters are taken to be common across cases and dispersion is allowed to vary across phases, the nuisance parameter estimates are $\hat{\sigma}^2 = 0.01$, $\hat{\rho} = -0.56$, $\hat{\gamma}_0 = 0.36$, and $\hat{\gamma}_1 = 0.72$; substantively, the latent variability estimate is negligible, and the autocorrelation estimate is not meaningful. Overall, the six cases in this study do not display latent variability of the posited form, which necessarily rules out the possibility of latent serial dependence.

[10]In many other contexts, negative autocorrelation may be implausible, though in this study it does seem possible that the incidence of bullying may be lower following days where the incidence was higher, and vice versa.

6.4.2. Betz, Higbee, & Reagon (2008)

Betz, Higbee, and Reagon (2008) used an unbalanced AB design with replication to evaluate the effect of using a joint activity schedule on the level of engagement of three pairs of autistic children.[11] In this example, each dyad (pair of children) represents a single case. For each dyad, the investigators measured engagement repeatedly during a baseline phase where no joint activity schedule was used, followed by a treatment phase where a teacher prompted the students to use a joint activity schedule to guide interactive play. Prompting was gradually faded over the course of the treatment phase. For analytic purposes, prompting is taken to be a fixed aspect of the intervention, rather than a dyamic consequence of both intervention and the dyad's level of engagement. Further data from maintenance and generalization phases are not used in the present analysis. Figure 6.12 plots the raw data from the study, which were obtained from a figure in the original study.

Outcome measurements were made using a momentary time sampling procedure, recording the presence or absence of engagement every 20 seconds over the course of each 20 minute observation session; the reported measurement on each occasion was then calculated as the proportion of moments where engagement was observed (out of 60 possible). The design included 4, 8, or 10 sessions during baseline, followed by 17, 9, or 23 measurements during each dyad's treatment phase. These phase lengths are shorter than would be desirable for estimation of serial dependence models, but I proceed with such analysis nonetheless. The original article gives no indication of the amount of time between each session (i.e., hourly, daily, bi-weekly), which is unfortunate–without such

[11]The authors describe the design as a "non-concurrent multiple baseline design," but this terminology is misleading because an essential feature of a multiple baseline design is concurrent measurement.

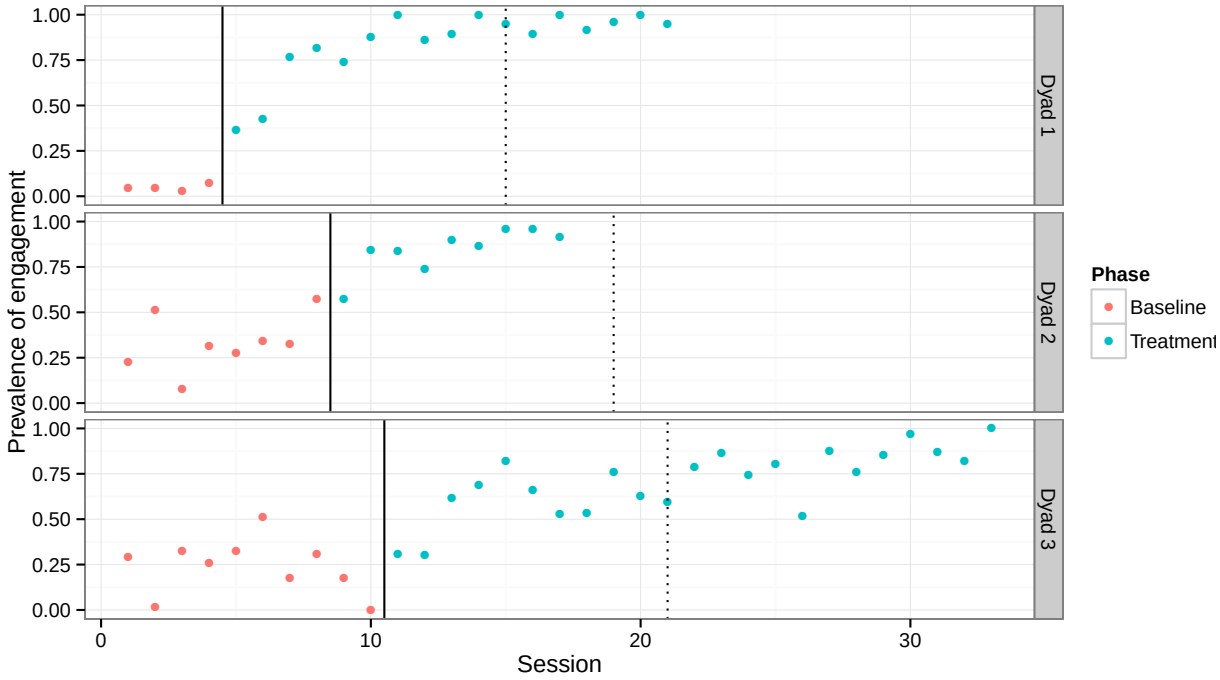


Figure 6.12. Data from an unbalanced design with replication across cases reported by Betz et al. (2008). Solid vertical lines indicate the introduction of the intervention; dotted vertical lines indicate the point at which the effect size is assessed.

information, it is difficult to interpret auto-correlation estimates and to assess the degree of extrapolation on which the effect size estimates are predicated. I focus on a specific target effect size that measures the effect of intervention after 10 sessions; this target time point is indicated by a dotted vertical line in Figure 6.12. The choice of 10 sessions represents a balance between effects that are substantively interesting and what is plausible given the available data. Even 10 sessions is ambitious in the latter regard because the longest baseline phase contains only 10 measurements. The metric of the target effect size is the log-prevalence odds ratio, which can be interpretted as the logged ratio of the odds that the dyad is engaged during treatment to the odds that the dyad would be engaged in the absence of treatment, both for a given case and at the specific point in time.

Table 6.4. Log-prevalence odds ratio effect size estimates for cases in Betz, Higbee, & Reagon (2008), assuming independence of repeated measurements.

Case	Effect size ($\hat{\psi}$)	V_M	$\hat{\gamma}$	V_R
Dyad 1	4.04	44.24	0.07	7.48
Dyad 2	2.52	2.35	0.07	3.17
Dyad 3	2.92	1.76	0.10	2.89

Let Y_{ij} denote the j^{th} outcome measurements from case i and let $\phi_{ij} = \mathrm{E}\left(Y_{ij}\right)$, both for $i = 1, ..., 3$ and $j = 1, ..., n_i$. Let $\tilde{d}_i$ be the session number of the first observation during treatment for case i. I assume that the measurements follow the model

$$\text{logit}\left(\phi_{ij}\right) = \beta_{0i} + \beta_{1i}\left(j - \tilde{d}_i\right) + \beta_{2i} I\left(j \geq \tilde{d}_i\right) + \beta_{3i}\left(j - \tilde{d}_i\right) \times I\left(j \geq \tilde{d}_i\right), \tag{6.27}$$

where $I()$ is the usual indicator function. For dyad i, this model accounts for natural changes in the prevalence of engagement through the inclusion of logit-linear time trends during the baseline phase, and further allows that the trend in prevalence may be affected by intervention. As previously, let $\mathbf{x}_{ij}$ collect the covariates for the j^{th} measurement on case i and let $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \beta_{2i}, \beta_{3i})'$. The target effect size for case i is then $\psi_i = \beta_{2i} + 10\beta_{3i}$. Due to the relatively small number of baseline observations, I only consider one model for the variance of the outcomes, the pooled dispersion model, given by

$$\text{Var}\left(Y_{ij}\right) = \gamma_i \phi_{ij}(1 - \phi_{ij}. \tag{6.28}$$

For completeness, I also report robust variance estimates, as given in (6.7).

Table 6.4 reports effect size estimates and standard errors, calculated based on the methods described in Section 6.1.1 and 6.1.3, respectively. Point-estimates for the log-prevalence odds ratio range from 2.52 to 4.04, which are all large effects, corresponding to many-fold increases in the odds of engagement. However, the effects are estimated with low precision, particularly for Dyad 1, whose estimated effect size of 4.04 has an approximate confidence interval (based on the model-based variance estimate) of $\hat{\psi} \pm 2\sqrt{V_M} = (-9.27, 17.33)$. Such a lack of precision is intuitively consistent with the large extrapolation involved with this dyad, where the baseline trend is extrapolated based on only four data points. By contrast, the robust variance estimate seems inappropriately small. Dyad 2 has an approximate confidence interval (again based on the model-based variance estimates) of (-0.54, 5.58), while dyad 3 has an approximate confidence interval of (0.26, 5.58). Only the case with the longest baseline phase has an effect size estimated with sufficient precision to distinguish it from zero by this rough measure.[12] Based on a fixed-effects meta-analysis using inverses of the model-based variance estimates as weights, the average log-prevalence odds ratio for the three cases is 2.78, with a 95% confidence interval of (0.83, 4.72).

Figure 6.13 depicts the fit of the logit-linear model (6.27) for this study, which is helpful for building intuition and model-checking purposes. The outcome data are plotted on the log-prevalence odds scale, so that fitted trend lines are linear.[13] Note in particular the two observations in the baseline phase for the third dyad that are below -2.5. These observations have high influence. In particular, removing observation $Y_{3,10}$ leads to an

[12] A similar conclusion holds when confidence intervals are based on other variance estimates, such as assuming constant dispersion across cases.

[13] Outcomes equal to zero or one are set to $\text{logit}(1/120) \approx -5$ or $\text{logit}(1 - 1/120) \approx 5$, respectively.

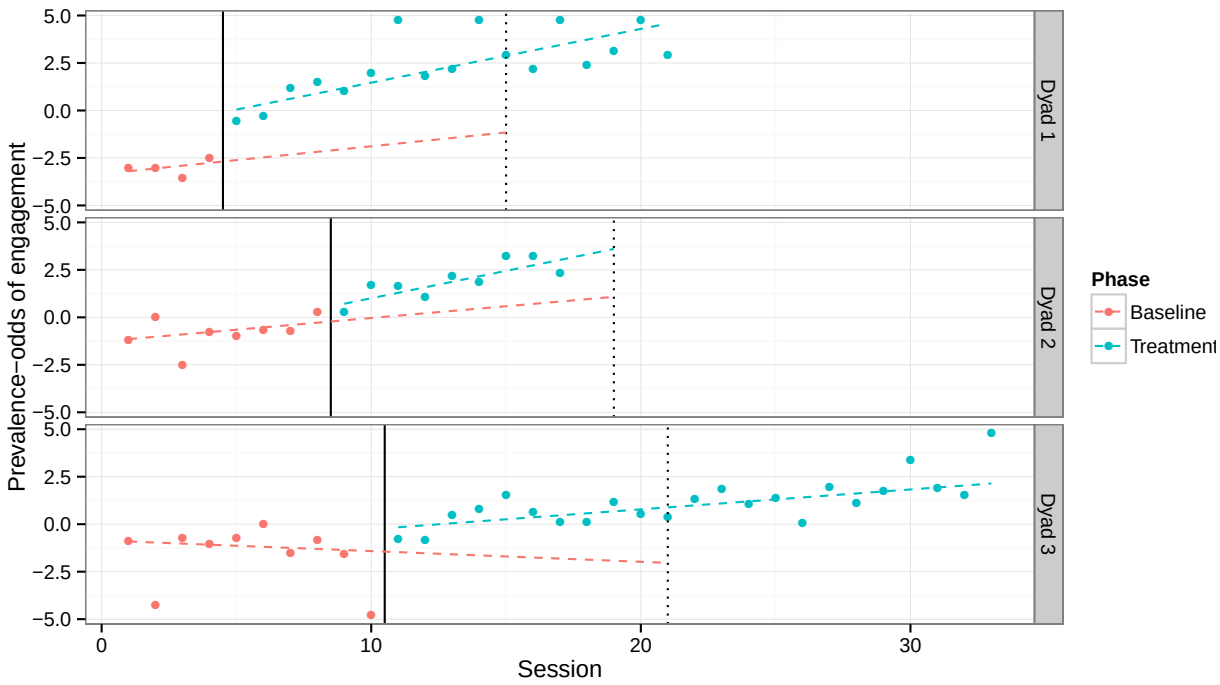


Figure 6.13. Logit-transformed outcomes versus session for Betz et al. (2008), plus fit of model (6.27). Solid vertical lines indicate the introduction of the intervention; dotted vertical lines indicate the point at which the effect size is assessed. Dashed lines depict the estimated logit-linear trends.

effect size estimate of just $\hat{\psi}_3 = 1.41$ with an approximate 95% confidence interval of (-1.44, 4.27).

I now turn to a model that allows for serial dependence. I assume the same marginal mean specification as given in (6.27), but now allow for a serially correlated latent error term such that

$$\text{logit}\left[\text{E}\left(Y_{ij}|\nu_{ij}\right)\right] = f\left(\mathbf{x}_{ij}'\boldsymbol{\beta}_i\right) + \nu_{ij},$$

where $\nu_{ij} \sim N(0, \sigma_i^2)$ and $(\nu_{i1}, ..., \nu_{in_i})$ follow an AR1 serial dependence model with autocorrelation ρ_i. I assume the same pooled same dispersion model as given in (6.28), and also fit the model assuming that all cases share common nuisance parameters, so that

Table 6.5. Effect size estimates, variance estimates, and pseudo-maximum likelihood estimates of nuisance parameters for cases from Betz, Higbee, & Reagon (2008), assuming serial dependence of repeated measurements.

	1st-order approximation					2nd-order approximation				
Case	$\hat{\psi}$	V_M	$\hat{\sigma}^2$	$\hat{\rho}$	$\hat{\gamma}$	$\hat{\psi}$	V_M	$\hat{\sigma}^2$	$\hat{\rho}$	$\hat{\gamma}$
Dyad 1	3.73	25.92	0.18	0.10	0.03	3.69	27.16	0.14	0.14	0.03
Dyad 2	3.08	0.74	0.35	-0.54	0.00	2.99	0.78	0.27	-0.51	0.00
Dyad 3	3.01	1.67	0.11	0.25	0.07	2.99	1.66	0.09	0.26	0.07
Pooled			0.33	-0.12	0.02			0.27	-0.09	0.03

$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 = \cdots = \boldsymbol{\theta}_6$, where $\boldsymbol{\theta}_i = (\sigma_i^2, \rho_i^2, \gamma_i)$. I report estimates based on both first- and second-order approximations to the covariance of the outcomes.

Table 6.3 reports estimates of effect sizes $\hat{\omega}^{\zeta}$, variances V_M, and nuisance parameters, all obtained through iterative fitting of the marginal mean estimating equation $\mathbf{U}_3(\boldsymbol{\beta}, \boldsymbol{\theta})$ and the Gaussian full pseudo-likelihood estimating equation $\mathbf{U}_4(\boldsymbol{\beta}, \boldsymbol{\theta})$, as described in Section 6.3.[14] Effect size estimates, variance estimates, and nuisance parameter estimates from the first- and second-order covariance approximations are all very similar; I discuss the estimates based on the first-order approximation because the simulation results suggest that these may be more stable. The effect size estimates under the serial dependence model are closer to one another and also larger, on average, than the estimates based on assuming independence.[15] The variance estimates are smaller than the estimates based on assuming independence, though the reason for the reduction is not the same across dyads. For dyad 1, the reduction in variance is due to the decrease in the estimated dispersion parameter; the effect size for this dyad is still very imprecisely estimated. For dyad 2,

[14]Convergence occurred between 4 and 12 iterations.

[15]A fixed-effects meta-analysis of the estimates from the first-order approximation produces an estimate for the average log-prevalence odds ratio of 3.01, with a 95% confidence interval of (1.59,4.42).

the reduction in variance is due to a strongly negative estimate of the autocorrelation. For dyad 3, the reduction in variance is slight. due to a modestly positive estimate of autocorrelation.

The final row of Table 6.5 reports estimates based on a model which assumes that nuisance parameters are common across cases. Estimates from both approximations suggest that the data exhibit non-trivial latent variability that is negatively autocorrelated. It is difficult to judge whether negative autocorrelation is plausible due to the lack of information about session timing in the report of the study. Whether the nuisance parameters are pooled or taken as distinct across cases, the variances of the effect size estimates are lower when accounting for possible serial dependence than when independence is assumed, which is rather counter-intuitive. Following the heuristic that larger variance estimates are more conservative, it seems more defensible for the analyst to use the variance estimates based on assuming independence, rather than those based on allowing for serial dependence.

6.5. Discussion

This chapter has examined ways to extend the between-session model for free-operant behavioral measurements to include features that are key for single-case researchers. I first considered models that assume independence of repeated measurements but incorporate simple time trends; such models, and effect sizes defined with respect to them, can be estimated using the framework of quasi-likelihood for generalized linear models. In these models, model-based variance estimators are more accurate than the asymptotically consistent robust variance estimators, so long as the variance model is not drastically

misspecified. The variance model could also be extended to describe other patterns of change over time, while still using the proposed estimating equation approach. This may be useful as a method of describing behaviors where both prevalence and incidence change concurrently, and should be explored in future work.

I then described models where repeated measurements are serially dependent as a result of a latent error process in the parameters of the behavior stream. Serial dependence has been recognized as a crucial feature of statistical models for single-case research (Horner et al., 2012; Wolery et al., 2010), but previous research has been limited to models for continuous, interval-scaled measurements. Incorporating serial dependency involves further complications for data in the form of counts and proportions, including a distinction between the conditional and marginal means of the outcome process. I have proposed an approach to modeling and effect size definition based on marginal means, then investigated an estimation strategy involving linear estimating equations for effect size parameters and Gaussian pseudo-likelihood for the parameters describing variance and serial dependence.

I found using simulation studies that the proposed estimation approach may require larger sample sizes for adequate performance (and particularly for estimation of effect size variances) than are typically available from single-case time series. I can see three possible avenues of work to address this shortcoming. First, one could consider alternative approaches to estimation of the nuisance parameters, such as by introducing penalty terms. It would also be useful to pursue bias corrections or second-order estimating equations for the mean parameters in the models considered in this chapter, so as to maintain

internal coherence with the bias-adjusted effect size estimators proposed in Chapter 5 for the stable phase model.

Second, an alternative to focusing only on improving estimation techniques is to consider introducing exchangeability assumptions at the level of the case. This chapter has focused solely on models for individual cases, in keeping with the ideographic perspective of single-case research, but a natural next step is to introduce assumptions describing variation across cases. As I argued in Chapter 2, such assumptions will in fact be necessary for defining design-comparable effect sizes for studies of free-operant behavior, but they may also bring benefits for estimation of individual effect sizes and nuisance parameters. However, the plausibility of the exchangeability assumptions will need to be carefully articulated and evaluated in the context of single-case studies.

A third avenue of work involves more and better data collection. Guidance about what classes of models can feasibly be estimated for samples of a given size would clearly be useful for the field. Besides just sample size and power analysis, though, improving data quality would be an important step forward. Much wider classes of models can be considered if the analyst has access to lower level data (what I have termed recorded data) rather than only what is available in published graphs. I discuss this further in Section 7.3.

A crucial limitation of the models and estimation methods considered in this chapter arises from the use of the quasi-likelihood framework. I have relied on it because it greatly reduces the scope of the modeling problem: rather than requiring a description of the full distribution of the data, quasi-likelihood relies on assumptions about the first two moments only, and has some useful robustness properties as a result. Moreover, these more

limited assumptions are likely a better representation of the investigator's state of knowledge about the behavior stream process. Thus, quasi-likelihood is highly advantageous when the goal is just to define and estimate a target effect size, but it has the disadvantage of not fully describing the data-generating process. If one's goal is not a reductive summary but a richer description of the behavior stream process, then the quasi-likelihood framework is unsatisfactory. This point is illustrated by the examples presented in Section 6.4, where the reported nuisance parameter estimates are an incomplete description of the data-generating model. Even given accurate nuisance parameter estimates, one would still lack crucial details about the behavior stream process that could inform future simulation work or be used for parametric bootstrapping approaches to variance estimation.

Finally, I note that the models presented in this chapter have focused solely on data generated by event counting, continuous recording, or momentary time sampling procedures. Interval recording is another widely used procedure for behavioral observation in free-operant contexts but, as typically summarized, it does not produce a direct measurement of prevalence or incidence. Models that are superficially similar to those for momentary time sampling could conceivably be applied to interval recording data, but the mean specification would be contingent on details of how the measurements were collected. For example, a linear trend in the mean of 10-second partial interval recording would not be linear if the data were collected using 20-second partial interval recording. Such operational sensitivity makes it very difficult to specify a model that is both conceptually coherent and tractable for estimation. Whether these difficulties can be overcome–or whether a sizable portion of extant single-case research is not amenable to statistical modeling due to flawed measurement operations–remains to be seen.

CHAPTER 7

Future directions

This chapter presents several pieces of work in progress, illustrating some of the projects that I hope to pursue in further work. Some of these future directions lead beyond the domain of single-case research, while others target common research practices in single-case research but are not immediately connected to meta-analysis. In Section 7.1, I describe a method for estimating standardized mean difference effect sizes in longitudinal designs, designed to be insensitive to the serial dependence structure of the repeated measurements. This method is applicable not only to single-case designs, but also to other types of interrupted time series designs, which are beginning to receive more attention in education research (Bloom, 2003; Somers, Zhu, Jacob, & Bloom, 2012).

Chapter 5 described an approach to defining and estimating measurement-comparable effect sizes for quantifying free-operant behavior. One implication of this exercise is that some of the most commonly used procedures for measuring free-operant behavior are deeply flawed, at least as currently analyzed. In Section 7.2, I present further illustrations of the problems with interval recording procedures. These examples demonstrate that–under not implausible circumstances–interval recording can lead to mistaken conclusions about the efficacy of an intervention, regardless of whether one uses a statistical or visual approach to inference. On a related topic, Section 7.3 summarizes some in-progress work aimed at developing new methods of analyzing interval recording data and

new methods of measuring free-operant behavior, both seeking to remedy shortcomings of current practices.

A primary goal of this thesis was to present new approaches to modeling and estimating operationally comparable effect sizes. Having argued that operational comparability–and measurement comparability in particular–is an important criterion in selecting an effect size metric for meta-analysis, I proposed an approach to modeling measurement comparability for outcomes based on free-operant behavior. This new proposal is one of a growing array of effect sizes for single case research, but the operational comparability of many of these other effect sizes has yet to be rigorously examined. Section 7.4 collects notes on several prominent effect sizes, with the aim of understanding their operational comparability. I offer some brief, general concluding thoughts in Section 7.5.

7.1. Robust moment estimation of standardized mean differences

In Chapter 3, I studied methods for estimating design-comparable effect sizes in the family of standardized mean differences, which are defined under one of several different hierarchical models that are linear in terms of the variance component structure. Following the notation of Section 3.3, the target effect size parameters all have the form

$$\delta = \frac{\mathrm{E}(Y_{iJ}|\mathbf{x}_{iJ} = \mathbf{x}^1) - \mathrm{E}(Y_{iJ}|\mathbf{x}_{iJ} = \mathbf{x}^0)}{\sqrt{\mathrm{Var}(Y_{iJ}|\mathbf{x}_{iJ} = \mathbf{x}^0)}} = \frac{\mathbf{p}'\boldsymbol{\beta}}{\sqrt{\mathbf{r}'\boldsymbol{\theta}}}, \tag{7.1}$$

where $\mathbf{x}_{ij}$ is a covariate vector describing the j^{th} measurement occasion for the i^{th} case, $i = 1, ..., m$ and $j = 1, ..., n_i$, $\boldsymbol{\beta}$ is a $p \times 1$ set of fixed effect parameters, $\boldsymbol{\theta}$ is a suitably parameterized $r \times 1$ set of variance component parameters, and and $\mathbf{p}$ and $\mathbf{r}$ are respectively $p \times 1$ and $r \times 1$ vectors of constants. Here, the numerator of the effect size is the

difference in mean outcomes at a fixed point in time J of two groups described by the covariate values $\mathbf{x}^1$ and $\mathbf{x}^0$, and the square of the denominator of the effect size is the total variance of the outcome (across the individual series) also at fixed time J.

I proposed to use restricted maximum likelihood (RML) estimation for the component parameters of a hierarchical model for the data from a single-case design, then to form an effect size estimator by substituting the RML estimates for the corresponding parameters and applying an approximate small-sample bias correction. For models with a single case-level variance component, I found that this estimator had bias and precision comparable to that of an alternative estimator proposed by Hedges et al. (2012a, 2012b, HPS hereafter).

One potential criticism of the adjusted RML estimator for δ is that it is contingent on having the correct model for $\mathbf{y}$. If, for instance, the dependence structure of $\boldsymbol{\epsilon}_i$ is mis-specified, or if $\mathbf{T}$ is incorrectly assumed to have a factor structure, or if the errors $\boldsymbol{\epsilon}_i$ and $\boldsymbol{\eta}_i$ do not follow the posited Gaussian distributions, then the estimates of $\boldsymbol{\theta}$ may be inconsistent even as the number of cases grows large. This in turn will lead to inconsistent estimation of δ. Furthermore, regardless of consistency, if the estimates are based on only a small number of series m, then certain components of $\hat{\boldsymbol{\theta}}$ may have moderate or even severe biases.

All of these model specification threats affect estimation of the variance components. In contrast, RML estimation of the fixed effects is unbiased even if the variance components are poorly estimated, a property that carries through to the numerator of the effect size estimator. Also, sandwich estimates are available for the covariance matrix of the fixed effects that (under general conditions) remain asymptotically consistent despite

mis-specification of the variance structure. It would be desirable to have a method for estimating $\mathbf{r}'\boldsymbol{\theta}$ that is similarly robust to model mis-specification. The effect size estimator proposed by HPS has this robustness quality, in that it uses an exactly unbiased moment estimator of the total variance that does not depend on having good estimates for the other components of the variance structure. Unfortunately, that estimator can only be applied in models with a single random effect for each case.

7.1.1. A robust moment estimator

Here I describe one way in which the estimation approach proposed by HPS could be generalized to handle models with further random effects. Its properties are contingent on the specification of two models. First, one must have the correctly specified marginal mean structure $\mathrm{E}(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$. Second, one must have a correctly specified model for the marginal variance of the repeated measurements. Let

$$\mathrm{Var}(Y_{ij}|\mathbf{x}_{ij}) = \mathbf{g}_{ij}'\boldsymbol{\alpha}, \tag{7.2}$$

where $\mathbf{g}_{ij} = g(\mathbf{x}_{ij})$ is a $t \times 1$ function of the covariate vector $\mathbf{x}_{ij}$ and $\boldsymbol{\alpha}$ is a $t \times 1$ parameter vector. Note that by assumption, $\mathbf{r}'\boldsymbol{\theta} = \mathbf{g}_{HJ}'\boldsymbol{\alpha}$ for some covariate value $\mathbf{g}_{HJ} = g(\mathbf{x}^0)$.

One can construct unbiased estimates of $\mathrm{Var}(Y_{ij}|\mathbf{x}_{ij})$ by taking the sample variance of the outcomes across series with the same values of $\mathbf{x}_{ij}$. For a given value of the covariate $\mathbf{x}_{hj}$ and a given j, let

$$S_{hj}^2 = \frac{1}{m_{hj} - 1} \sum_{i=1}^{m} I(\mathbf{x}_{ij} = \mathbf{x}_{hj}) \left(Y_{ij} - \bar{y}_{hj}\right)^2,$$

where $m_{hj} = \sum_{i=1}^{m} I(\mathbf{x}_{ij} = \mathbf{x}_{hj})$ is the number of series with the target covariate value and

$$\bar{y}_{hj} = \frac{1}{m_{hj}} \sum_{i=1}^{m} Y_{ij} I(\mathbf{x}_{ij} = \mathbf{x}_{hj})$$

is the mean of the outcomes with the target covariate value.

Of course, because the estimator S^2_{hj} applies to the particular covariate value at which the variance in the effect size denominator is sought, one possible estimate of $\mathbf{r}'\boldsymbol{\theta}$ is simply S^2_{HJ}. However, this estimator might be very imprecise because it disregards much of the data. For improved precision, one could use the model for the marginal variance given in (7.2), estimated via weighted least squares. For a given j, suppose that there are k_j unique values of $\mathbf{x}_{ij}$ and a total of $K = \sum_{j=1}^{n} k_j$ unique covariate-by-measurement occasion combinations. Let $\mathbf{G}$ be the $K \times t$ matrix that collects the $\mathbf{g}_{hj}$ corresponding to the unique covariate values, so that

$$\mathbf{G}' = (g(\mathbf{x}_{11})', ..., g(\mathbf{x}_{k_n n})') .$$

Let $\mathbf{S} = (S^2_{11}, ..., S^2_{k_n n})'$ be a $K \times 1$ vector collecting the sample variances at each combination of covariate values, and write $\mathbf{C} = \text{Cov}(\mathbf{S})$. An unbiased estimator of $\mathbf{r}'\boldsymbol{\theta}$ is then given by

$$U = \mathbf{g}'_{HJ}\hat{\boldsymbol{\alpha}} = \mathbf{g}'_{HJ} (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} \mathbf{G}'\mathbf{W}\mathbf{S} \tag{7.3}$$

for a given $K \times K$ weighting matrix $\mathbf{W}$. For arbitrary $\mathbf{W}$, the variance of U is given by

$$\text{Var}(U) = \mathbf{g}'_{HJ} (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} \mathbf{G}'\mathbf{W}\mathbf{C}\mathbf{W}\mathbf{G} (\mathbf{G}'\mathbf{W}\mathbf{G})^{-1} \mathbf{g}_{HJ}, \tag{7.4}$$

The optimal (minimum variance) weighting matrix is given by the inverse of the covariance matrix of $\mathbf{S}$, in which case U has variance $\mathbf{g}'_{HJ}\left(\mathbf{G}'\mathbf{C}^{-1}\mathbf{G}\right)^{-1}\mathbf{g}_{HJ}$. Even though $\mathbf{C}$ depends on the entire covariance structure of $\mathbf{y}$, one can still construct $\mathbf{W}$ based on RML estimators of $\boldsymbol{\theta}$ (or any other estimators) while preserving the unbiasedness of the target variance estimator U.

The estimator U can be written as a quadratic form in $\mathbf{y}$. Let $\mathbf{F}$ be an $mn \times K$ matrix that maps the unique covariate-by-measurement occasion combinations to the observations, with entries $f_{m(i-1)+j,h} = I(\mathbf{x}_{ij} = \mathbf{x}_{hj})$ for $h = 1, ..., K$, $i = 1, ..., m$, and $j = 1, ..., n$. Now observe that the residuals of which the S^2_{hj} are composed can be written as

$$\mathbf{r} = \left[\mathbf{I}_{mn} - \mathbf{F}\left(\mathbf{F}'\mathbf{F}\right)^{-1}\mathbf{F}'\right]\mathbf{y}. \tag{7.5}$$

Note that $\mathbf{F}'\mathbf{F} - \mathbf{I}_K$ is a $K \times K$ diagonal matrix with diagonal entries equal to $m_{11} - 1, ..., m_{k_n n} - 1$. Define the $mn \times mn$ matrices

$$\begin{aligned}\mathbf{B} &= \operatorname{diag}\left[\mathbf{g}'_{HJ}\left(\mathbf{G}'\mathbf{W}\mathbf{G}\right)^{-1}\mathbf{G}'\mathbf{W}\left(\mathbf{F}'\mathbf{F} - \mathbf{I}_K\right)^{-1}\mathbf{F}'\right] \\ \mathbf{A} &= \left[\mathbf{I}_{mn} - \mathbf{F}\left(\mathbf{F}'\mathbf{F}\right)^{-1}\mathbf{F}'\right]\mathbf{B}\left[\mathbf{I}_{mn} - \mathbf{F}\left(\mathbf{F}'\mathbf{F}\right)^{-1}\mathbf{F}'\right] = \left[\mathbf{I}_{mn} - \mathbf{F}\left(\mathbf{F}'\mathbf{F}\right)^{-1}\mathbf{F}'\right]\mathbf{B}.\end{aligned} \tag{7.6}$$

It follows that $U = \mathbf{r}'\mathbf{B}\mathbf{r} = \mathbf{y}'\mathbf{A}\mathbf{y}$. It can be verified that $\mathrm{E}(U) = \operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}) = \mathbf{g}'_{HJ}\boldsymbol{\alpha}$.

Because U is a quadratic form, the usual Box-Satterthwaite approximation can be used to derive an approximate degrees-of-freedom correction:

$$\nu = \frac{U^2}{\operatorname{tr}(\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}\boldsymbol{\Sigma})}, \tag{7.7}$$

and the target effect size estimated by

$$g_U = J(\nu)\frac{\mathbf{p}'\hat{\boldsymbol{\beta}}}{\sqrt{U}} \tag{7.8}$$

for a given, unbiased estimate of $\hat{\boldsymbol{\beta}}$. Of course, this small-sample bias correction is based on the full, Gaussian likelihood posited for $\mathbf{y}$.

7.1.2. Future work

Much work remains to fully develop this approach. Three of the more immediate questions to be answered are:

(1) What form for the marginal variance model given in (7.2) is implied by certain simple random effects specifications? It would be useful to fully work out the algebra for some leading examples, such as for a model with a single random slope.

(2) For some leading examples, how large a loss of efficiency is entailed in using U to estimate the effect size scale, relative to using RML estimators directly?

(3) How would missing observations on some measurement occasions affect the properties of U? It may be that by relying only on the marginal variance model, missing observations pose a greater threat than under the full hierarchical model. It will be important to make a precise statement of the conditions under which missing data is ignorable for estimating U, and consider mitigation strategies for when those conditions are not met.

Beyond these questions, it will also be useful to construct and study examples where the hierarchical model is mis-specified, either in terms of the dependence structure or the error distributions, in order to put the claimed robustness of this approach to the test.

7.2. Problems with partial interval recording

The results and estimation methods presented in Chapter 5 illustrate that partial interval recording data is highly problematic because it is not a direct measurement of any readily interpretable dimension of behavior. Rather, the measurand of a partial interval recording procedure is a function of both prevalence and incidence, as well as the distribution of interim times and the length of the active interval. Results of studies that use partial interval recording are thus difficult to interpret due to their dependence on scientifically irrelevant operational factors.

Still, partial interval recording remains in wide use.[1] One possible justification for its use is that construct validity is secondary to internal validity: so long as the procedure is applied consistently across measurement occasions (i.e., holding the active interval length and session length constant for the duration of the study), the internal validity of the study is preserved. I have encountered this line of argument several times during discussions with applied researchers, and so it is worth illustrating carefully why it is incorrect. In this section, I develop two hypothetical examples demonstrating how using partial interval recording can produce misleading conclusions about whether an intervention has the intended effect. These examples serve as cautionary tales, warning applied researchers of the perils of definitional operationalism and the importance of construct validity.

[1]Some textbooks even recommend it. Kazdin (2011) advises: "Whenever there is doubt as to what assessment strategy should be adopted, an interval approach is almost always applicable" (p. 79).

Texts on behavioral observation sometimes distinguish between between discrete behaviors, in which each instance of the behavior has negligible duration (e.g., hi-fiving) versus state behaviors, in which instances can last for longer periods of time (e.g., hugging). In the former case, interest is in the incidence of the behavior; in the latter, prevalence may be of primary interest. Partial interval recording is used in both contexts, interpreted as a measure either of incidence or of prevalence. I give one example of each context.

7.2.1. Partial interval recording for measuring incidence of a discrete behavior

First, consider a study evaluating the effect of an intervention for reducing the self-injurious behavior of a child with autism. Prior to intervention, the child displays self-injurious behaviors that have very short duration, so that incidence is the primary dimension of interest. Suppose that, prior to intervention, the behaviors follow the alternating renewal process described in Section 5.1.2, with all event durations equal to 0 (so $\mu^B = 0$) and interim time distribution $F_E^B(t)$ given by the following mixture of two gamma distributions:

$$F_E^B(t) = \frac{3}{5} F_\Gamma(t|24, 1) + \frac{2}{5} F_\Gamma(t|8, 6),$$

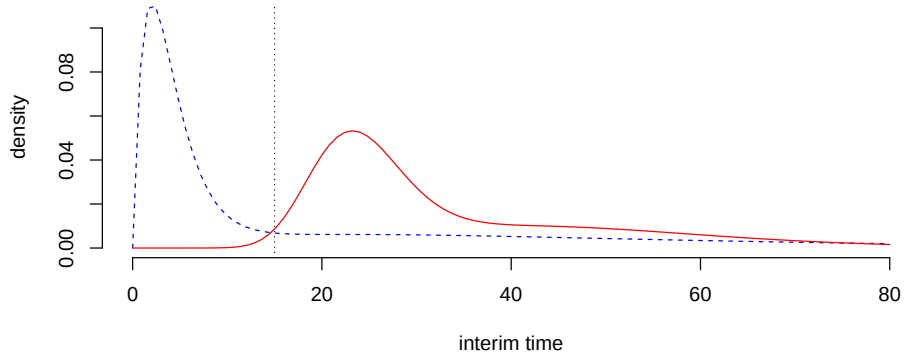
where $F_\Gamma(x|k, \theta)$ denotes the cumulative distribution function of a gamma random variable with shape k and scale θ. Figure 7.1a plots the density of this interim time distribution. Given this distribution, the average interim time between self-injurious behaviors is $\lambda^B = \frac{3}{5} \times 24 + \frac{2}{5} \times 48 = 33.6$ seconds. Further suppose that the intervention causes a change in the distribution of interim times, so that after introducing the treatment, the behaviors follow an alternating renewal process with $\mu^T = 0$ and interim time distribution $F_E^T(t)$

given by

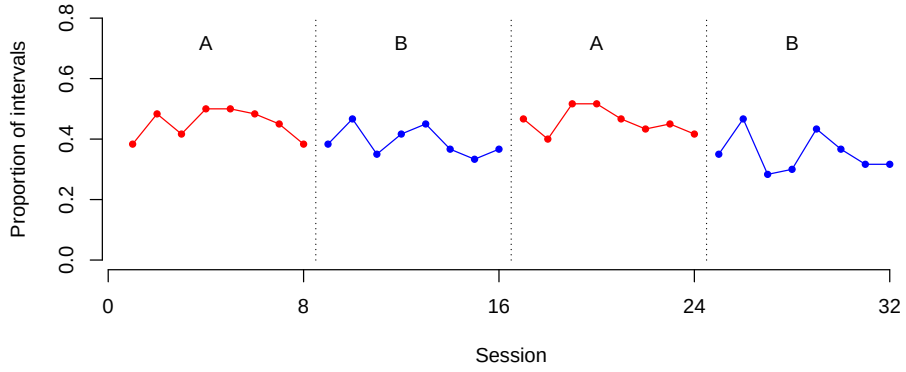
$$F_E^T(t) = \frac{3}{5}F_\Gamma(t|2,2) + \frac{2}{5}F_\Gamma(t|2,24).$$

See Figure 7.1a for a plot of the interim time density after intervention. The average interim time between self-injurious behaviors is now $\lambda^T = \frac{3}{5} \times 4 + \frac{2}{5} \times 48 = 21.6$ seconds. The behaviors are substantially more frequent (going from 2 per minute to 3 per minute), meaning that the intervention does not produce the desired reduction in behavior, and is instead actually harmful.

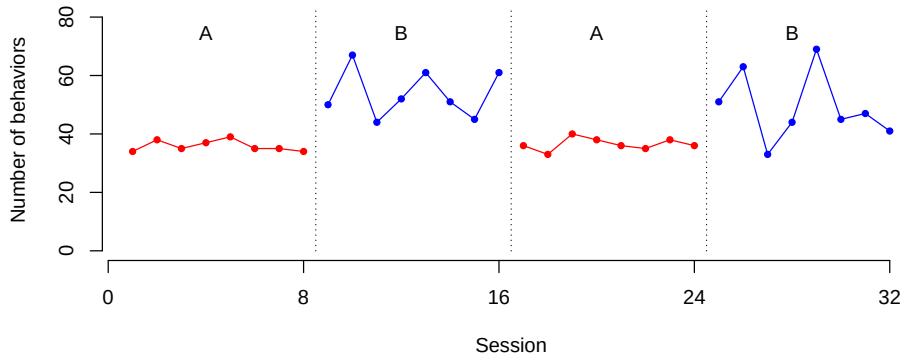
Suppose that the investigator uses an ABAB design with 8 sessions per phase; during each session, he measures self-injurious behavior using partial interval recording with an active interval length of $l = 15$ seconds, 5 seconds of rest time for recording, and a total session length of 20 minutes. Figure 7.1b plots an example of how the results of this study might appear; it was created by simulating behavior stream data according to the distributions F_E^B and F_E^T, then calculating the partial interval recording data according to Equations (5.3) and (**??**). The average proportion of partial intervals in this simulated example is 0.45 during the A (baseline) phases, compared to 0.37 during the B (intervention) phases; the proportion actually *decreases* slightly, even though the true incidence has increased! Of course, this is only one realization. To verify that the decrease is not just a fluke of the particular sample, the expected value of the partial interval data from each phase can be calculated across many repetitions of the study. From the formula given in Table 5.3, the expected proportion of intervals is 0.45 during the A phases compared to 0.38 during the B phases, so the decrease will be observed generally. In fact, if the investigator used more sessions per phase or longer observation times during each session, the decrease in partial intervals would have been even more apparent.



(a) Density of interim time distributions. Solid red line denotes the distribution prior to intervention. Dashed blue line denotes the distribution after intervention. Dotted vertical line indicates the active interval length.



(b) Simulated single-case graph using partial interval recording



(c) Simulated single-case graph using event counting

Figure 7.1. Example of partial interval recording with a discrete behavior

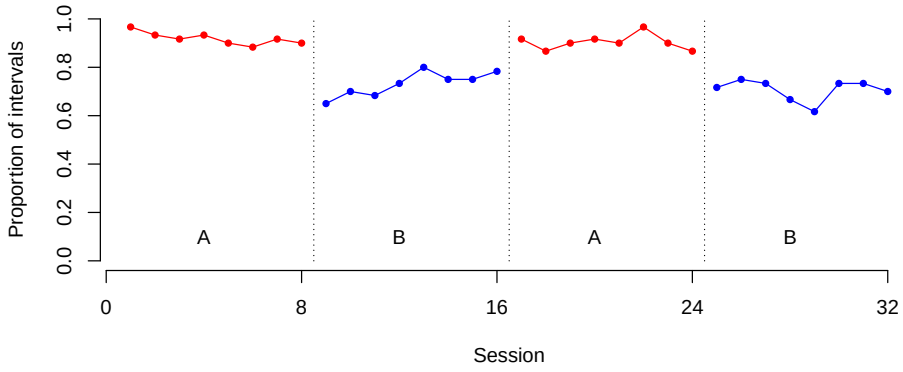
Now imagine that a second researcher also participated in this study, observing each session and using an event counting procedure to record measurements. Figure 7.1c plots the data that she would record, based on the same realization of the underlying behavior stream as used in Figure 7.1b. The increase in frequency is readily apparent; in this simulated example, the average number of events goes from 36.2 during the A phases to 51.5 during the B phase. Over many repetitions of the study, the expected number of events per session is 35.7 during the A phases and 55.6 during the B phases.

7.2.2. Partial interval recording for measuring prevalence of a state behavior

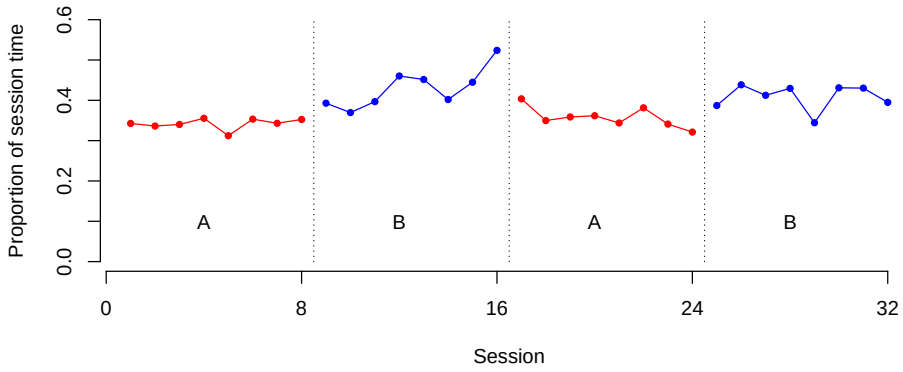
Similarly deceptive results are also possible when using partial interval recording to measure the prevalence of state behaviors.[2] Consider a study evaluating the effect of a particular teaching technique thought to prevent disruptive behavior. A particular child displays disruptive behavior that can last for non-trivial lengths of time, and so the main dimension of interest is prevalence. The investigator uses an ABAB design with eight 20-minute sessions per phase; she measures disruptive behavior using partial interval recording with an active interval length of $l = 15$ seconds and 5 seconds of rest time for recording.

Prior to intervention, the child displays disruptive behaviors that last an average of $\mu^B = 6$ seconds and that follow a gamma distribution with $F_D^B(t) = F_\Gamma(t|2,3)$; the interim time between instances of disruptive behavior also follows a gamma distribution with $F_E^B(t) = F_\Gamma(t|3,4)$, so that the average interim time is $\lambda^B = 12$ seconds. It follows that, on average, the child's prevalence of disruptive behavior is $\mu^B/(\mu^B + \lambda^B) = 0.33$. Suppose that the teaching technique causes an increase in both the average duration

[2]Kraemer (1979) discussed an example similar to the one I present, though her analysis was based on a fixed behavior stream rather than a stochastic process.



(a) Simulated single-case graph using partial interval recording



(b) Simulated single-case graph using continuous recording

Figure 7.2. Example of partial interval recording with a state behavior

of disruptive events and the average interim time. Specifically, when the intervention is applied, $\mu^T = 20$, $F_D^T(t) = F_\Gamma(t|2, 10)$, $\lambda^T = 30$, and $F_E^T(t) = F_\Gamma(t|3, 10)$. On average then, the intervention increases the prevalence of disruptive behavior to $\mu^T/(\mu^T + \lambda^T) = 0.40$.

Figure 7.2a plots an example of how the results of this study might appear. A clear decrease is evident: in this realization of the simulation, the proportion of partial intervals averages 0.91 during the A (baseline) phases versus 0.71 during the B (intervention) phases. More generally, the expected proportion of partial intervals can be calculated as 0.91 during the A phases versus 0.68 during the B phases. Using partial interval

recording may lead the investigator to conclude that the teaching technique reduces the prevalence of the child's disruptive behavior, when in fact the opposite is true. Had the investigator instead used a continuous recording procedure, the results would appear as in Figure 7.2b; in this realization of the simulation, the proportion of session time with disruptive behavior averages 0.35 during A phases and 0.42 during B phases, very close to the true prevalence levels. Using momentary time sampling would have produced similarly unbiased results, though with somewhat lower precision.

7.3. Markov models for partial interval recording and momentary time sampling

Given that the mean of partial interval recording data is very difficult to interpret in terms of the underlying parameters of the behavior stream process, it would be useful to consider other methods for analyzing such data, yet little previous research has done so. To my knowledge, the only existing proposals for more elaborate analysis are due to Ary and Suen (1983) and Suen and Ary (1984, 1986), who proposed procedures for estimating incidence and prevalence from recorded data generated using partial interval recording. However, these procedures lack any articulated, model-based motivation and their empirical performance has been called into question (Rogosa & Ghandour, 1991).

Other approaches have been studied for analyzing momentary time sampling data. For example M. Brown, Solomon, and Stephens (1977) and Griffin and Adams (1983) proposed using a latent alternating Poisson process model of the behavior stream to estimate prevalence and incidence based on momentary time sampling data. In work reported elsewhere (Pustejovsky, 2013), I have followed a similar approach for the analysis

of partial interval recording data. Specifically, I have demonstrated that partial interval recording data can be modeled using a discrete-time Markov chain derived from a latent alternating Poisson process model of the behavior stream. Thus, if the analyst has access to within-session recorded data collected using a partial interval procedure (rather than only reported summary data), full likelihood-based methods can be used to estimate behavioral prevalence and incidence.

I have also proposed a new procedure for recording behavioral observation that involves combining momentary time sampling, partial interval recording, and whole interval recording (Pustejovsky, 2013). This new procedure, called augmented interval recording, also involves a latent alternating Poisson process model, but produces recorded data that has a much simpler dependence structure. Over much of the parameter space, the new procedure yields estimates of prevalence and incidence that are considerably more precise than those based on partial interval recording.

The proposed methods for analyzing recorded partial interval data and augmented interval data have two main limitations, having to do with practical feasibility and sensitivity to modeling assumptions. Regarding feasibility, I suspect that the augmented interval recording procedure will require only marginally more effort than interval recording alone, but this has yet to be verified in the field. Furthermore, estimation of model parameters based on data collected using either procedure will require computations that are considerably more complex than simply taking the mean. Thus, it will be vital to create easy-to-use programs that efficiently automate the required calculations, in order to make the proposed estimation techniques accessible and attractive for researchers in

the field. Regarding sensitivity, the estimation procedures are based on a latent alternating Poisson process model. I use the assumptions of this parametric model due to the mathematical tractability they provide, rather than out of any conviction that they are empirically appropriate. In future research, I will need to study the sensitivity of the proposed methods to violations of the underlying modeling assumptions. I suspect that such violations may be rather difficult to assess, even from the comparatively rich data generated by augmented interval recording.

Several related avenues of further research are required to fully develop these estimation methods and novel recording procedures. First and most immediately, I will need to build a prototype device (such as a smart phone app) so that the augmented interval recording procedure can be tested under realistic conditions. Second, I will need to make further comparisons between the procedures (e.g., momentary time sampling, partial interval recording, and augmented interval recording) and provide practical guidance regarding their use. Third, for the new estimation methods to be usefully applied in the context of single-case designs, I will need to incorporate regression models and perhaps even frailty (random-effects) models for describing session-to-session variation in prevalence and incidence. Finally, it will be useful to consider whether the models for partial interval recording and augmented interval recording can be generalized to accommodate alternating renewal processes other than the alternating Poisson process. In pursuing each of these tasks, the central goal will be to provide better tools for recording and analyzing data based on direct observation of behavior, which improve the precision and interpretability of measurements while maintaining ease of application.

7.4. Other effect size proposals

In Chapters 3 and 5, I have presented a variety of parametric models for single case designs. I showed how these models can be used to express design-comparable and measurement-comparable effect sizes in terms of model components. Such operationally comparable effect sizes add to an expanding literature on methods for summarizing and meta-analyzing single-case designs. The models that I have presented in previous chapters are also useful for understanding the practical interpretation and statistical properties of other effect size proposals in the single-case literature. This section briefly examines the properties of several prominent effect size proposals under the assumptions of some of the simple parametric models that I have described, focusing in particular on their sensitivity to design- and measurement-related study operations.

7.4.1. Within-case standardized mean differences

For a simple two-phase design, Busk and Serlin (1992) proposed using the standardized mean difference between phases as a measure of effect size for a single case. The two main critiques of this effect size are that it deals only with changes in the average level of the outcome (rather than accounting for trends as well) and that its sampling distribution is affected by auto-correlation (Beretvas & Chung, 2008b).

Consider applying the standardized mean difference to the data from a single case in a multiple baseline design, with treatment assigned after time T_i. The standardized mean difference between phases is calculated as

$$d_{BS} = \frac{\bar{y}_i^T - \bar{y}_i^B}{s_i}, \tag{7.9}$$

where (suppressing the dependence of Y_{ij} on T_i)

$$\bar{y}_i^B = \frac{1}{T_i}\sum_{j=1}^{T_i} Y_{ij}, \qquad \bar{y}_i^T = \frac{1}{n-T_i}\sum_{j=T_i+1}^{n} Y_{ij}$$

and

$$s_i^2 = \frac{1}{n-2}\left[\sum_{j=1}^{T_i}\left(Y_{ij} - \bar{y}_i^B\right)^2 + \sum_{j=T_i+1}^{n}\left(Y_{ij} - \bar{y}_i^T\right)^2\right].$$

Suppose that Model MB1 applies, so that there are no time trends, and further suppose that repeated measurements on a given case are independent, so that $\phi = 0$. It then follows that, for a given case i, d_{BS} estimates β_{1i}/σ; if the case is sampled from a larger population of cases, then d_{BS} can be considered an estimate of γ_{10}/σ. This is the average treatment effect, scaled by the variance of the within-case errors. Van den Noortgate and Onghena (2008) note that this parameter is related to the design-comparable effect size d_{AB} by a factor of $\sqrt{1-\rho}$, where $\rho = \tau_0^2/\left(\tau_0^2 + \sigma^2\right)$ can be interpreted as a reliability coefficient for the outcome measure.

It is also interesting to note that if Model MB1 applies, including non-zero first-order auto-correlation, d_{BS} will be biased as an estimate of γ_{10}/σ. This is because s_i^2 will be a biased estimate of σ^2: from a result given in Hedges et al. (2012a, Appendix B),

$$\text{(7.10)} \qquad \mathrm{E}(s_i^2) = \sigma^2\left[1 - \frac{2}{n-2}\left(\frac{1}{T_i}\sum_{j=1}^{T_i-1}\phi^j(T_i-j) + \frac{1}{n-T_i}\sum_{j=1}^{n-T_i-1}\phi^j(n-T_i-j)\right)\right].$$

It can be seen from (7.10) that the bias of s_i^2 depends not only on the degree of auto-correlation, but also on the phase lengths. All else equal, longer phase lengths will lead to less-biased estimates of σ^2 and thus less-biased estimates of γ_{10}/σ. Furthermore, for fixed

phase lengths, the conventional small-sample correction for standardized mean differences (Hedges, 1981) will not apply in an exact sense because s_i^2 will not be χ^2_{n-2}-distributed.

Other researchers have proposed estimation methods that account for serial correlation, which will mitigate (at least to some extent) the problem of using s_i^2 to estimate σ^2. For example, Swaminathan et al. (2010) scale their treatment effect estimate by the standard deviation of the within-case errors, adjusted based on an estimate of the first-order auto-correlation. As noted in Section 2.4.2, these authors define their treatment effect as the difference between the observed outcomes at the mid-point of the treatment phase and the predicted outcome at the same point, based on a linear trend projection from the baseline phase. Under Model MB5, which allows for baseline trends and treatment-by-time trend interactions, their proposed effect size for a given case i has estimand

$$\left(\beta_{1i} + \beta_{3i}\frac{N + T_i}{2}\right)/\sigma.$$

Note that the magnitude of this effect size depends on design features–if the study had been longer, or if the case had been assigned to treatment at a different point, then the target parameter would have been different. As noted in Section 2.4.2, such design-dependence is undesirable because it introduces irrelevant variation and thus reduces the interpretability of the parameter. To illustrate this reduced interpretability, consider how the case-level estimand implied by the Swaminathan et al. (2010) procedure would change if cases were sampled from a larger population of cases. One would have to postulate some mechanism by which new cases would be assigned a treatment assignment time. Then assuming that such assignment would be completely at random, the effect size can be

considered an estimate of

$$\left(\gamma_{10} + \gamma_{30}\frac{N+\bar{T}}{2}\right)/\sigma,$$

where $\bar{T}$ is the mean of the distribution of potential treatment assignment times.

7.4.2. Reliability-corrected standardized mean differences

In the context of cross-sectional experimental designs, Hunter and Schmidt (2004) have proposed using an effect size that divides the usual standardized mean difference by the reliability of the outcome measure ρ. As noted in Section 1.1.2, this reliability-adjustment can be motivated by a particular model for the comparability of different outcome measurement operations. For use with cross-sectional experiments, their procedure typically requires information from an external validation study. However, because single-case designs involve repeated measurements, an internal estimate of the reliability of the outcome is available. The test-retest reliability of the outcome is the correlation between measurements on the same individual at a given time j and at time $j + k$ (in the absence of treatment), $\rho = \text{corr}\,(Y_{ij}, Y_{i,j+k})$, where k is the number of measurement occasions between tests. Under Model MB1 or MB2, the total variance of the outcome is stable, and so $\rho = (\tau_0^2 + \phi^k\sigma^2)/(\tau_0^2 + \sigma^2)$. For sufficiently spaced tests, $\rho \approx \tau_0^2/(\tau_0^2 + \sigma^2)$, as noted by Van den Noortgate and Onghena (2008), and the reliability-adjusted effect size would have estimand γ_{10}/τ_0. So long as there is any within-case variation, this effect sizes will be greater than the design-comparable δ_{AB}.[3]

[3]Of course, the presence of time trends complicates the notion of reliability (Molenaar, 2004; Raudenbush & Liu, 2001). For instance, under Model MB4, the correlation between Y_{ij} and $Y_{i,j+k}$ depends on which measurement occasions are chosen.

7.4.3. Non-overlap statistics

Non-overlap statistics are the most commonly used measures of effect size for meta-analysis of single-case research (Maggin, O'Keeffe, & Johnson, 2011). Though it is sometimes claimed that the interpretation of these statistics does not depend on a particular parametric model (c.f. Parker, Vannest, & Davis, 2011), the implied estimands of many of these statistics do in fact vary depending on the data-generating model. This can be seen by examining the behavior of several non-overlap statistics under simple models for multiple baseline data.

First consider the percentage of non-overlapping data (PND), a popular effect size metric proposed by Scruggs et al. (1987), and intended for use when the data do not display time trends. For a given case, and assuming that the treatment is intended to increase the level of the outcome, the PND is defined as the percentage of data points in the treatment phase that exceed the maximum point of the baseline phase:

$$PND_i = 100\% \times \frac{1}{n - T_i} \sum_{j=T_i+1}^{n} I\left[Y_{ij} > \max\{Y_{i1}, ..., Y_{iT_i}\}\right]. \tag{7.11}$$

Working under assumptions similar to MB1 and assuming no autocorrelation ($\phi = 0$), Allison and Gorman (1994) used simulation to demonstrate that the expectation of PND depends strongly on the length of the baseline phase, making it highly design-dependent.[4] Their simulation results can be verified analytically by noting that, since $\phi = 0$, repeated measurements are independent and identically distributed.[5] Writing $\Phi(\cdot)$ for the cumulative distribution of a standard normal random variate and $\phi(\cdot)$ for the standard normal

[4]They also provide an analytical approximation for the expectation assuming that the treatment effect was null ($\gamma_{10} = 0$), reporting that $E(PND_i) \approx 1 - 2^{-1/T_i}$.

[5]To be precise, this is true conditional on the model parameters for case i.

density, it follows that

$$\begin{aligned}(7.12)\qquad \frac{1}{100\%}\mathrm{E}\left[PND_i\right] &= \mathrm{E}\left[\Pr\left(Y_{1n} > \max\{Y_{i1},...,Y_{iT_i}\}|Y_{i1},...,Y_{iT_i}\right)\right] \\ &= \mathrm{E}\left[1-\Phi\left(\frac{\max\{Y_{i1},...,Y_{iT_i}\}-\beta_{0i}-\beta_{1i}}{\sigma}\right)\right] \\ &= \mathrm{E}\left[1-\Phi\left(Z_{(T_i)}-\frac{\gamma_{10}}{\sigma}\right)\right],\end{aligned}$$

where $Z_{(T_i)}$ is the maximum of T_i independent standard normal random variates. [6] Using the density of $Z_{(T_i)}$ (Severini, 2005, p. 218),

$$(7.13)\qquad \mathrm{E}\left[PND_i\right] = 100\% \times \left[1 - T_i\int_{-\infty}^{\infty}\Phi\left(z-\frac{\gamma_{10}}{\sigma}\right)\left[\Phi(z)\right]^{T_i-1}\phi(z)dz\right],$$

an expression that can be evaluated using numerical integration.[7] Furthermore, if $\gamma_{10}/\sigma = 0$, then (7.13) evaluates to $\mathrm{E}(PND_i) = 100\%/(T_i+1)$.[8] Figure 7.3a plots the expectation of the PND statistic as a function of the within-case standardized mean difference γ_{10}/σ; separate lines are used for treatment assignment times of $T_i = 5, 9, 15$. It can be seen that

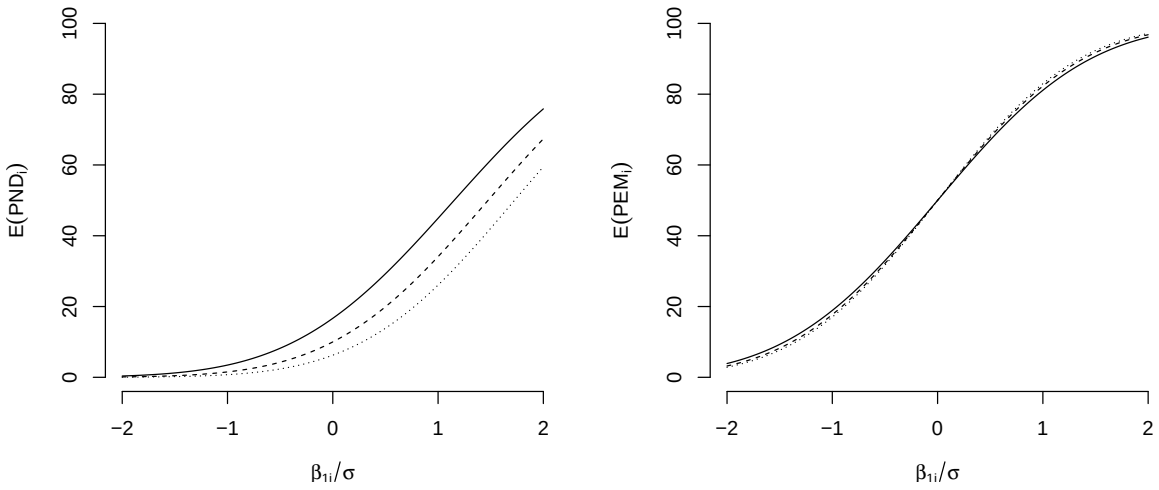
[6] Also,

$$\begin{aligned}\mathrm{Var}\left(\frac{PND_i}{100\%}\right) = \frac{1}{n-T_i}\mathrm{E}\left[\Phi\left(Z_{(T_i)}-\frac{\gamma_{10}}{\sigma}\right)\right] \\ + \frac{n-T_i-1}{n-T_i}\mathrm{E}\left[\Phi^2\left(Z_{(T_i)}-\frac{\gamma_{10}}{\sigma}\right)\right] - \mathrm{E}^2\left[\Phi\left(Z_{(T_i)}-\frac{\gamma_{10}}{\sigma}\right)\right].\end{aligned}$$

[7] This result assumes that the treatment effect is constant across cases. If it is instead assumed that each case is sampled from a larger population of cases and that treatment effects vary as in Model MB2, the expected value of PND in the population would be found by replacing γ_{10} by β_{1i} in (7.13), then taking the expectation over the distribution of β_{1i}. Because $\mathrm{E}(PND_i|\beta_{1i})$ is non-linear with respect to β_{1i}, the population expectation would depend not only on γ_{10}, but also on τ_1^2.

[8] Also, if $\gamma_{10}/\sigma = 0$, then

$$\mathrm{Var}(PND_i) = (100\%)^2 \times \frac{T_i(n+1)}{(n-T_i)(T_i+1)^2(T_i+2)}.$$



(a) E(PND_i) (b) E(PEM_i)

Figure 7.3. Expectations of non-overlap statistics as a function of β_{1i}/σ, for varying values of T_i. Solid lines use $T_i = 5$; dashed lines use $T_i = 9$; dotted lines use $T_i = 15$.

the expected value of PND_i is strongly affected by the length of the baseline phase, even for non-null treatment effects, and is also non-linearly related to γ_{10}/σ.

Another non-overlap effect size, the percentage of data-points exceeding the median (PEM), is closely related to PND and can be similarly analyzed. The PEM statistic, proposed by Ma (2006), is defined as the percentage of data-points in the treatment phase that exceed the median of the baseline phase:

$$PEM_i = 100\% \times \frac{1}{n - T_i} \sum_{j=T_i+1}^{n} I\left[Y_{ij} > \text{median}\{Y_{i1}, ..., Y_{iT_i}\}\right]. \tag{7.14}$$

For simplicity, suppose that T_i is odd. Under Model MB1 and assuming $\phi = 0$, it can be shown that

$$\text{E}\left[PEM_i\right] = 100\% \times \left[1 - T_i \binom{T_i - 1}{\frac{T_i - 1}{2}}\right. \tag{7.15}$$

$$\times \int_{-\infty}^{\infty} \Phi\left(z - \frac{\gamma_{10}}{\sigma}\right) [\Phi(z)]^{(T_i-1)/2} [1 - \Phi(z)]^{(T_i-1)/2} \phi(z) dz \Bigg].$$

(The derivation follows along similar lines to that for PND_i.) Furthermore, if $\gamma_{10}/\sigma = 0$, then $\mathrm{E}(PEM_i) = 100\%/2$.[9] Figure 7.3b plots the expectation of the PEM statistic as a function of the within-case standardized mean difference γ_{10}/σ, for varying baseline phase lengths T_i. It can be seen that the expected value of PEM_i is insensitive to baseline phase length, and converges rapidly to $\Phi(\gamma_{10}/\sigma)$ as T_i increases.

The expected values of the PND and PEM statistics were derived under the assumption that repeated measurements on the same case are (conditionally) independent. Expectations under more general dependence assumptions (such as first-order auto-regression) are considerably more complex because they involve distributions of multivariate normal order statistics, and remain a topic for further investigation. However, the distribution of certain other non-overlap statistics can be obtained even without the independence assumption.

Parker and Vannest (2009) proposed to use the non-overlap of all pairs of data-points in the baseline and the treatment phases (NAP) as a measure of effect size for single-case research.[10] For a treatment expected to improve the level of a continuous outcome, and not displaying trends in baseline or treatment phases, the statistic is defined as the number of pairs in which the observation from the treatment phase is greater than the observation from the baseline phase, as a proportion of all pairs of baseline and treatment

[9] Also, if $\gamma_{10}/\sigma = 0$, then

$$\mathrm{Var}(PEM_i) = (100\%)^2 \times \frac{n+1}{4(n-T_i)(T_i+2)}.$$

[10] The authors note that the NAP statistic is equivalent to the area under a receiver-operator characteristic curve and that it is known in various other contexts as the common language effect size, the probability of superiority, the dominance statistic, and the Mann-Whitney U statistic (Parker & Vannest, 2009).

phase observations:

$$NAP_i = \frac{1}{T_i(n - T_i)} \sum_{j=1}^{T_i} \sum_{k=T_i+1}^{n} I\left[Y_{ik} > Y_{ij}\right]. \tag{7.16}$$

Parker and Vannest (2009) argued that NAP offers improved sensitivity and greater concurrent validity with other measures of effect size, compared to PND, PEM, and other non-overlap statistics.

The expectation of the NAP statistic is determined partially by the level of autocorrelation among repeated measurements. Under the assumptions of Model MB1, and allowing $-1 < \phi < 1$,

$$\begin{aligned} \mathrm{E}(NAP_i) &= \frac{1}{T_i(n - T_i)} \sum_{j=1}^{T_i} \sum_{k=T_i+1}^{n} \Pr\left(Y_{ik} - Y_{ij} > 0\right) \\ &= \frac{1}{T_i(n - T_i)} \sum_{j=1}^{T_i} \sum_{k=T_i+1}^{n} \Phi\left(\frac{\gamma_{10}}{\sigma\sqrt{2(1 - \phi^{|k-j|})}}\right). \end{aligned} \tag{7.17}$$

If $\gamma_{10}/\sigma = 0$, then $\mathrm{E}(NAP_i) = \frac{1}{2}$. Also, if $\phi = 0$ then $\mathrm{E}(NAP_i) = \Phi\left(\gamma_{10}/(\sigma\sqrt{2})\right)$. For a positive treatment effect, the expectation is inflated slightly by ϕ, though not substantially so except when ϕ is close to 1. The increase is also mitigated by longer phase lengths.

7.4.4. Future work

I have analyzed several prominent effect size proposals, focusing mostly on simple models for the multiple baseline design. All of the proposals I have examined are related in some way to the distribution of the within-case errors: several proposals use the within-case variance to scale the treatment effect estimate, while the non-overlap statistics are functions of the full distribution of the within-case error distributions. Thus far, I have

focused entirely on the properties of these statistics when within-case measurement errors are normally distributed, with scale independent of the mean. In future work, I will need to revisit these effect size proposals when applied to different types of measurements, such as continuous recording or event counting. I expect to find that many of the proposals will be very sensitive to the measurement operations of the study, such as the length of the observation session used to collect continuous recording data, the length of the intervals used to collect partial interval recording data. More fundamentally, the interpretation of some proposed effect sizes may even depend on which recording procedure is used.

7.5. Final thoughts

This thesis has proposed new approaches to analyzing single-case research, focusing on models and methods under which operational comparable effect sizes can be defined. I have examined two distinct aspects of operational comparability: design-comparability in Chapters 2 through 4 and measurement-comparability in Chapters 5 and 6. As noted in Section 5.6.2, the general framework for design-comparability could certainly be applied to define effect sizes that are both design- and measurement-comparable. Before proceeding along that path in future work, it is worth first reflecting on some broad lessons drawn from the work presented in this thesis. I offer the following comments largely for purposes of self-critique, though some of them may apply more broadly to other recent proposals for quantitative analysis of single-case research. If other researchers find them useful, so much the better.

The first lesson is that, in order to make true progress, future methodological developments will need to be closely tied to empirical applications. The measurement-comparable

effect sizes for free-operant behavior proposed in Chapter 5 were motivated by an actual research synthesis on a topic that continues to be of interest to the field (c.f. P. L. Morgan, 2006; von Mizener & Williams, 2008). The example gave a sense of the scope across which the proposed effect sizes apply: not to all single-case studies, but to a subset broad enough to encompass the studies included in a full synthesis. Future work that has similar scope, such as effect sizes for measuring changes in restricted-operant behavior, will be useful to the extent that it it aligns with sensible inclusion criteria for the field of application. The same principle holds for the development of design- and measurement-comparable effect sizes, which will only be useful to the extent that a body of empirical research exists where they might reasonably be applied.

A related lesson is that the assessment of new quantitative methods for single-case research should also be closely tied to empirical applications. In the simulation studies presented in this thesis, I have sought to use design parameters that coincide with empirical practice (e.g., the number of cases per study, the length of phases). After all, the primary limitation of any simulation study is lack of generality: if the scope of a simulation is unrealistic, its conclusions will not be relevant to actual applications. However, tying simulations to empirical data proved to be difficult for the models of free-operant behavioral processes described in Chapters 5 and 6, precisely because the proposed models are novel and thus the empirical knowledge base for understanding their parameters does not yet exist. The implication would seem to be that the development of real-life empirical applications should not be limited to methods for which exhaustive simulation evidence already exists, because the latter must be informed by the former. Furthermore,

the development of new analytic methods can have implications for the design stages of research practice, by motivating investigators to collect more or different data.

A final lesson is that, if quantitative analysis of single-case research is to become an established practice, there will need to be greater focus on model fitting and assumption verification. In many of the examples that I have presented, target effect size estimates were sensitive to functional form specification. This feature is a feature common to most if not all interrupted time series designs, and needs to be emphasized more directly. Apart from causal identification assumptions, regular assumptions regarding statistical fit also need to be scrutinized. Though I have sought to follow good analytic practice in the examples I have presented, I have not reported these in any detail. Future work may need to lay out analytic steps and model checks more explicitly in order to establish best practices. Connections between model checking and established conventions of visual analysis should also be explored.

References

Allison, D. B., & Franklin, R. D. (1992). Reflections on visual inspection, response guided experimentation, and Type I error rate in single-case designs. *The Journal of Experimental Education*, *61*(1), 45–51.

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy*, *31*(6), 621–31.

Allison, D. B., & Gorman, B. S. (1994). Make things as simple as possible, but no simpler. A rejoinder to Scruggs and Mastropieri. *Behaviour Research and Therapy*, *32*(8), 885–890. doi: 10.1016/0005-7967(94)90170-8

Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour*, *49*(3/4), 227–267.

Altmann, S. A., & Wagner, S. S. (1970). Estimating rates of behavior from Hansen frequencies. *Primates*, *11*(2), 181–183. doi: 10.1007/BF01731143

Alvero, A. M., Struss, K., & Rappaport, E. (2007). Measuring safety performance: A comparison of whole, partial, and momentary time-sampling recording methods. *Journal of Organizational Behavior Management*, *27*(4), 1–28.

Anglesea, M. M., Hoch, H., & Taylor, B. A. (2008). Reducing rapid eating in teenagers with autism: Use of a pager prompt. *Journal of Applied Behavior Analysis*, *41*(1), 107–111. doi: 10.1901/jaba.2008.41-107

Ary, D., & Suen, H. K. (1983). The use of momentary time sampling to assess both frequency and duration of behavior. *Journal of Behavioral Assessment*, *5*(2), 143–150.

Ayres, K., & Gast, D. L. (2010). Dependent measures and measurement procedures. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 129–165). New York, NY: Routledge.

Bambara, L. M., Koger, F., Katzer, T., & Davenport, T. A. (1995). Embedding choice in the context of daily routines: An experimental case study. *The Journal of the Association for Persons with Severe Handicaps*, *20*(3), 185–195.

Barlow, D. H., & Hersen, M. (1984). *Single Case Experimental Designs: Strategies for Studying Behavior Change*. New York, NY: Pergamon Press, Inc.

Beasley, T. M., Allison, D. B., & Gorman, B. S. (1996). The potentially confounding effects of cyclicity: Identification, prevention, and control. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 279–333). Mahwah, NJ: Lawrence Erlbaum.

Beretvas, S. N., & Chung, H. (2008a). An evaluation of modified R2-change effect size indices for single-subject experimental designs. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 120–128. doi: 10.1080/17489530802446328

Beretvas, S. N., & Chung, H. (2008b). A review of meta-analyses of single-subject experimental designs: Methodological issues and practice. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 129–141. doi: 10.1080/17489530802446302

Betz, A., Higbee, T. S., & Reagon, K. A. (2008). Using joint activity schedules to promote peer engagement in preschoolers with autism. *Journal of Applied Behavior Analysis*, *41*(2), 237–241. doi: 10.1901/jaba.2008.41-237

Bickel, P. J., & Doksum, K. A. (2007). *Mathematical Statistics: Basic Ideas and Selected Topics, Volume I* (2nd ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Biosoft. (2004). *UnGraph for Windows.* Cambridge, Great Britain: Author.

Bloom, H. S. (2003). Using "short" interrupted time-series analysis to measure the impacts of whole-school reforms: With applications to a study of accelerated schools. *Evaluation Review*, *27*(1), 3–49. doi: 10.1177/0193841X02239017

Bloom, H. S. (2005). Randomizing groups to evaluate place-based programs. In H. S. Bloom (Ed.), *Learning more from social experiments: Evolving analytic approaches* (pp. 115–172). New York, NY: Russell Sage Foundation.

Bloom, H. S., Bos, J. M., & Lee, S.-W. (1999). Using cluster random assigment to measure program impacts: Statistical implications for the evaluation of education programs. *Evaluation Review*, *23*(4), 445–469.

Borenstein, M. (2009). Effect sizes for continuous data. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 221–236). New York, NY: Russell Sage Foundation.

Bornstein, M. H. (2002). Measurement variability in infant and maternal behavioral assessment. *Infant Behavior and Development*, *25*(4), 413–432. doi: 10.1016/S0163-6383(02)00143-1

Box, G. E. P., & Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, *70*(349), 70–79.

Bradshaw, C. P., Mitchell, M. M., & Leaf, P. J. (2009). Examining the effects of schoolwide positive behavioral interventions and supports on student outcomes: Results from a randomized controlled effectiveness trial in elementary schools. *Journal of Positive Behavior Interventions*, *12*(3), 133–148. doi: 10.1177/1098300709334798

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, *88*(421), 9–25.

Brown, C. A., & Lilford, R. J. (2006). The stepped wedge trial design: A systematic review. *BMC Medical Research Methodology*, *6*, 54. doi: 10.1186/1471-2288-6-54

Brown, M., Solomon, H., & Stephens, M. A. (1977). Estimation of parameters of zero-one processes by interval sampling. *Operations Research*, *25*(3), 493–505.

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education*, *97*(1), 65–108.

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment*, *10*(3), 229–242.

Busk, P. L., & Serlin, R. C. (1992). Meta-analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case research design and analysis: New directions for psychology and education* (pp. 187–212). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Campbell, D. T. (1957). Factors relevant to the validity of experiments in social settings. *Psychological Bulletin*, *54*(4), 297–312.

Campbell, J. M. (2004). Statistical comparison of four effect sizes for single-subject designs. *Behavior Modification*, *28*(2), 234–46. doi: 10.1177/0145445503259264

Campbell, J. M., & Herzinger, C. V. (2010). Statistics and single subject research methodology. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 417–450). New York, NY: Routledge.

Center, B. A., Skiba, R. J., & Casey, A. (1985). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education*, *19*(4), 387.

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology*, *52*, 685–716. doi: 10.1146/annurev.psych.52.1.685

Chard, D., Ketterlin-Geller, L., & Baker, S. (2009). Repeated reading interventions for students with learning disabilities: Status of the evidence. *Exceptional Children*, *75*(3), 263–281.

Chi, E. M., & Reinsel, G. C. (1989). Models for longitudinal data with random effects and AR(1) errors. *Journal of the American Statistical Association*, *84*(406), 452–459.

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, *19*, 3127–3131.

Chung, Y., Rabe-Hesketh, S., Gelman, A., Liu, J., & Dorie, V. (2013). A non-degenerate estimator for hierarchical variance parameters via penalized likelihood estimation. *Psychometrika*. doi: 10.1007/S11336-013-9328-2

Cole, C. L., & Levinson, T. R. (2002). Effects of within-activity choices on the challenging behavior of children with severe developmental disabilities. *Journal of Positive Behavior Interventions*, *4*(1), 29–37.

Cook, T. D., Shadish, W. R., & Wong, V. C. (2008). Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management*, *27*(4), 724–750. doi: 10.1002/pam.20375

Cooper, H. M. (2009). Hypotheses and problems in research synthesis. In H. M. Cooper, L. V. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 19–35). New York, NY: Russell Sage Foundation.

Cox, D. R. (1962). *Renewal Theory*. Great Britain: Metheun & Co. Ltd.

Cox, D. R. (1981). Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics*, *8*(2), 93–115.

Cox, D. R., & Snell, E. (1968). A general definition of residuals. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, *30*(2), 248–275.

Crainiceanu, C. M., Ruppert, D., & Vogelsang, T. J. (2003). *Some properties of likelihood ratio tests in linear mixed models.* Retrieved from `http://legacy.orie.cornell.edu/davidr/papers/zeroprob_rev01.pdf`

Cronbach, L. J. (1982). *Designing Evaluations of Educational and Social Programs*. San Francisco, CA: Jossey-Bass.

Crosbie, J. (1993). Interrupted time-series analysis with brief single-subject data. *Journal of Consulting and Clinical Psychology*, *61*(6), 966–74.

Curtin, F., Altman, D. G., & Elbourne, D. (2002). Meta-analysis combining parallel and cross-over clinical trials. I: Continuous outcomes. *Statistics in Medicine*, *21*(15), 2131–44. doi: 10.1002/sim.1205

Curtin, F., Elbourne, D., & Altman, D. G. (2002). Meta-analysis combining parallel and cross-over clinical trials. II: Binary outcomes. *Statistics in Medicine*, *21*(15), 2145–59. doi: 10.1002/sim.1206

Czado, C., & Song, P. X. K. (2007). State space mixed models for longitudinal observations with binary and binomial responses. *Statistical Papers*, *49*(4), 691–714. doi: 10.1007/s00362-006-0039-y

Davidian, M., & Giltinan, D. M. (1995). *Nonlinear Models for Repeated Measurement Data*. Boca Raton, FL: Chapman & Hall/CRC.

Davis, R. A., Dunsmuir, W. T. M., & Wang, Y. (2000). On autocorrelation in a Poisson regression model. *Biometrika*, *87*(3), 491–505.

Deeks, J. J. (2002). Issues in the selection of a summary statistic for meta-analysis of clinical trials with binary outcomes. *Statistics in Medicine*, *21*(11), 1575–1600. doi: 10.1002/sim.1188

Dibley, S., & Lim, L. (1999). Providing choice making opportunities within and between daily school routines. *Journal of Behavioral Education*, *9*(2), 117–132.

Dominici, F., Parmigiani, G., Wolpert, R. L., & Hasselblad, V. (1999). Meta-analysis of migraine headache treatments: Combining information from heterogeneous designs. *Journal of the American Statistical Association*, *94*(445), 16–28.

Donner, A., & Klar, N. (2002). Issues in the meta-analysis of cluster randomized trials. *Statistics in Medicine*, *21*(19), 2971–80. doi: 10.1002/sim.1301

Downs, A., Downs, R. C., & Rau, K. (2008). Effects of training and feedback on Discrete Trial Teaching skills and student performance. *Research in Developmental Disabilities*, *29*(3), 235–46. doi: 10.1016/j.ridd.2007.05.001

Dunlap, G., DePerczel, M., Clarke, S., Wilson, D., Wright, S., White, R., & Gomez, A. (1994). Choice making to promote adaptive behavior for students with emotional and behavioral challenges. *Journal of Applied Behavior Analysis*, *27*(3), 505–518.

Dunn, P. K., & Smyth, G. K. (2012). *dglm: Double generalized linear models.* Retrieved from `http://cran.r-project.org/package=dglm`

Dyer, K., Dunlap, G., & Winterling, V. (1990). Effects of choice making on the serious problem behaviors of students with severe handicaps. *Journal of Applied Behavior Analysis*, *23*(4), 515–524.

Edgington, E., & Onghena, P. (2007). *Randomization Tests*. Boca Raton, FL: Chapman & Hall.

Elbourne, D. R., Altman, D. G., Higgins, J. P. T., Curtin, F., Worthington, H. V., & Vail, A. (2002). Meta-analyses involving cross-over trials: Methodological issues. *International Journal of Epidemiology*, *31*(1), 140–9.

Engels, E. A., Schmid, C. H., Terrin, N., Olkin, I., & Lau, J. (2000). Heterogeneity and statistical significance in meta-analysis: An empirical study of 125 meta-analyses. *Statistics in Medicine*, *19*, 1707–1728.

Faith, M. S., Franklin, R. D., Allison, D. B., & Gorman, B. S. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 245–277). Mahwah, NJ: Lawrence Erlbaum.

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods*, *41*(2), 372–84. doi: 10.3758/BRM.41.2.372

Fienberg, S. E. (1972). On the use of Hansen frequencies for estimating rates of behavior. *Primates*, *13*(3), 323–325.

Fisch, G. S. (2001). Evaluating data from behavioral analysis: Visual inspection or statistical models? *Behavioural Processes*, *54*(1-3), 137–154.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2011). *Applied Longitudinal Analysis* (2nd ed.). Hoboken, NJ: John Wiley & Sons.

Fleiss, J., & Berlin, J. A. (2009). Effect sizes for dichotomous outcomes. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 237–253). New York, NY: Russell Sage Foundation.

Franklin, R. D., Allison, D. B., & Gorman, B. S. (1996). *Design and analysis of single-case research*. Mahwah, NJ: Lawrence Erlbaum.

Frea, W. D., Arnold, C. L., & Vittimberga, G. L. (2001). A demonstration of the effects of augmentative communication on the extreme aggressive behavior of a child with autism within an integrated preschool setting. *Journal of Positive Behavior*

Interventions, *3*(4), 194–98.

Frost, C., Clarke, R., & Beacon, H. (1999). Use of hierarchical models for meta-analysis: Experience in the metabolic ward studies of diet and blood cholesterol. *Statistics in Medicine*, *18*(13), 1657–76.

Gabler, N. B., Duan, N., Vohra, S., & Kravitz, R. L. (2011). N-of-1 trials in the medical literature: A systematic review. *Medical Care*, *49*(8), 761–8. doi: 10.1097/MLR .0b013e318215d90d

Gadbury, G. L., & Iyer, H. K. (2000). Unit-treatment interaction and its practical consequences. *Biometrics*, *56*(3), 882–5.

Gadbury, G. L., Iyer, H. K., & Albert, J. M. (2004). Individual treatment effects in randomized trials with binary outcomes. *Journal of Statistical Planning and Inference*, *121*(2), 163–174. doi: 10.1016/S0378-3758(03)00115-0

Gage, N. A., Lewis, T. J., & Stichter, J. P. (2012). Functional behavioral assessment-based interventions for students with or at risk for emotional and/or behavioral disorders in school: A hierarchical linear modeling meta-analysis. *Behavioral Disorders*, *37*(2), 55–77.

Gardenier, N. C., MacDonald, R., & Green, G. (2004). Comparison of direct observational methods for measuring stereotypic behavior in children with autism spectrum disorders. *Research in Developmental Disabilities*, *25*(2), 99–118. doi: 10.1016/j.ridd.2003.05.004

Gast, D. L. (2010). Applied research in education and behavioral sciences. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 1–19). New York, NY: Routledge.

Gast, D. L., & Spriggs, A. D. (2010). Visual analysis of graphic data. In D. L. Gast (Ed.), *Single subject research methodology in behavioral sciences* (pp. 199–233). New York, NY: Routledge.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis*, *1*(3), 515–533.

Gelman, A., & Hill, J. L. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. New York, NY: Cambridge University Press.

Gilmour, A. R., Gogel, B. J., Cullis, B. R., & Thompson, R. (2009). *ASReml User Guide.* Hemel Hempstead, HP1 1ES, UK: VSN International Ltd.

Gilmour, A. R., Thompson, R., & Cullis, B. R. (1995). Average information REML:

An efficient algorithm for variance parameter estimation in linear mixed models. *Biometrics*, *51*(4), 1440–1450.

Gingerich, W. J. (1984). Meta-analysis of applied time-series data. *The Journal of Applied Behavioral Science*, *20*(1), 71–79. doi: 10.1177/002188638402000113

Goldstein, H., & Rasbash, J. (1996). Improved approximations for multilevel models with binary responses. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, *159*(3), 505–513.

Good, I. J. (1961). The frequency count of a Markov chain and the transition to continuous time. *The Annals of Mathematical Statistics*, *32*(1), 41–48.

Gorman, B. S., & Allison, D. B. (1996). Statistical alternatives for single-case designs. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 159–214). Mahwah, NJ: Lawrence Erlbaum.

Gorsuch, R. L. (1983). Three methods for analyzing limited time-series (N of 1) data. *Behavioral Assessment*, *5*(2), 141–154.

Gottman, J. M. (1981). *Time-Series Analysis: A Comprehensive Introduction for Social Scientists*. Cambridge, England: Cambridge University Press.

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology*, *99*(3), 445–476. doi: 10.1037/0022-0663.99.3.445

Greenland, S. (1987). Interpretation and choice of effect measures in epidemiologic analyses. *American Journal of Epidemiology*, *125*(5), 761–768.

Griffin, B., & Adams, R. (1983). A parametric model for estimating prevalence, incidence, and mean bout duration from point sampling. *American Journal of Primatology*, *4*(3), 261–271. doi: 10.1002/ajp.1350040305

Gunter, P. L., Venn, M. L., Patrick, J., Miller, K. A., & Kelly, L. (2003). Efficacy of using momentary time samples to determine on-task behavior of students with emotional/behavioral disorders. *Education and Treatment of Children*, *26*(4), 400–412.

Guyatt, G. H., Keller, J. L., Jaeschke, R., Rosenbloom, D., Adachi, J. D., & Newhouse, M. T. (1990). The n-of-1 randomized controlled trial: Clinical usefulness. *Annals of Internal Medicine*, *112*(4), 293.

Guyatt, G. H., Sackett, D., Adachi, J. D., Roberts, R. S., Chong, J., Rosenbloom, D., & Keller, J. L. (1988). A clinician's guide for conducting randomized trials in

individual patients. *Canadian Medical Association Journal*, *139*(6), 497.

Hall, D., & Severini, T. (1998). Extended generalized estimating equations for clustered data. *Journal of the American Statistical Association*, *93*(444), 1365–1375.

Hammond, D., & Gast, D. L. (2010). Descriptive analysis of single subject research designs: 1983-2007. *Education and Training in Autism and Developmental Disabilities*, *45*(2), 187–202.

Hanley, G. P., Iwata, B. A., & McCord, B. E. (2003). Functional analysis of problem behavior: A review. *Journal of Applied Behavior Analysis*, *36*(2), 147–85. doi: 10.1901/jaba.2003.36-147

Harrop, A., & Daniels, M. (1986). Methods of time sampling: A reappraisal of momentary time sampling and partial interval recording. *Journal of Applied Behavior Analysis*, *19*(1), 73–77.

Harrop, A., Daniels, M., & Foulkes, C. (1990). The use of momentary time sampling and partial interval recording in behavioural research. *Behavioural Psychotherapy*, *18*(2), 121–27. doi: 10.1017/S0141347300018231

Harrop, J. W., & Velicer, W. F. (1985). A comparison of alternative approaches to the analysis of interrupted time-series. *Multivariate Behavioral Research*, *20*, 27–44.

Hart, S. L., & Banda, D. R. (2009). Picture exchange communication system with individuals with developmental disabilities: A meta-analysis of single subject studies. *Remedial and Special Education*, *31*(6), 476–488. doi: 10.1177/0741932509338354

Hartmann, D. P., & Wood, D. D. (1990). Observational methods. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (2nd ed., pp. 107–138). New York, NY: Plenum Press.

Heagerty, P. J., & Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, *88*(4), 973–985.

Heagerty, P. J., & Zeger, S. L. (2000). Marginalized multilevel models and likelihood inference. *Statistical Science*, *15*(1), 1–26.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, *32*(4), 341–370. doi: 10.3102/1076998606298043

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development*

Perspectives, *2*(3), 167–171.

Hedges, L. V. (2011). Effect sizes in three-level cluster-randomized experiments. *Journal of Educational and Behavioral Statistics*, *36*(3), 346–380. doi: 10.3102/1076998610376617

Hedges, L. V., Gurevitch, J., & Curtis, P. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, *80*(4), 1150–1156.

Hedges, L. V., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. Orlando, FL: Academic Press.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012a). *A standardized mean difference effect size for multiple baseline designs.* Evanston, IL.

Hedges, L. V., Pustejovsky, J. E., & Shadish, W. R. (2012b). A standardized mean difference effect size for single case designs. *Research Synthesis Methods*, *3*, 224–239. doi: 10.1002/jrsm.1052

Hedges, L. V., Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*, 39–65. doi: 10.1002/jrsm.5

Hersen, M. (1990). Single-case experimental designs. In A. S. Bellack, M. Hersen, & A. E. Kazdin (Eds.), *International handbook of behavior modification and therapy* (2nd ed., pp. 175–210). New York, NY: Plenum Press.

Hershberger, S. L., Wallace, D. D., Green, S. B., & Marquis, J. G. (1999). Meta-analysis of single-case designs. *Statistical Strategies for Small Sample Research*, 109–132.

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–58. doi: 10.1002/sim.1186

Himle, M. B., Woods, D. W., & Bunaciu, L. (2008). Evaluating the role of contingency in differentially reinforced tic suppression. *Journal of Applied Behavior Analysis*, *41*(2), 285–289. doi: 10.1901/jaba.2008.41-285

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. L., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*(2), 165–179.

Horner, R. H., Sugai, G., Smolkowski, K., Eber, L., Nakasato, J., Todd, A. W., & Esperanza, J. (2009). A randomized, wait-list controlled effectiveness trial assessing school-wide positive behavior support in elementary schools. *Journal of Positive Behavior Interventions*, *11*(3), 133–144. doi: 10.1177/1098300709332067

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children*, *35*(2), 269–290. doi: 10.1353/etc.2012.0011

Horrocks, E., & Higbee, T. S. (2008). An evaluation of a stimulus preference assessment of auditory stimuli for adolescents with developmental disabilities. *Research in Developmental Disabilities*, *29*(1), 11–20. doi: 10.1016/j.ridd.2006.09.003

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment*, *7*(2), 107–118.

Huitema, B. E. (2011). *The Analysis of Covariance and Alternatives: Statistical Methods for Experiments, Quasi-Experiments, and Single-Case Studies*. Hoboken, NJ: John Wiley & Sons.

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods*, *3*(1), 104–116. doi: 10.1037//1082-989X.3.1.104

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement*, *60*(1), 38–58. doi: 10.1177/00131640021970358

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology*, *75*(3), 334–349.

Hunter, J. E., & Schmidt, F. L. (2004). *Methods of Meta-Analysis: Correcting Error and Bias in Research Findings*. Thousand Oaks, CA: Sage Publications, Inc.

Hunter, J. E., Schmidt, F. L., & Le, H. (2006). Implications of direct and indirect range restriction for meta-analysis methods and findings. *Journal of Applied Psychology*, *91*(3), 594–612. doi: 10.1037/0021-9010.91.3.594

Hussey, M. A., & Hughes, J. P. (2007). Design and analysis of stepped wedge cluster randomized trials. *Contemporary Clinical Trials*, *28*(2), 182–91. doi: 10.1016/j.cct.2006.05.007

Jennrich, R. I., & Schluchter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics*, *42*(4), 805–820.

Jenson, W. R., Clark, E., Kircher, J. C., & Kristjansson, S. D. (2007). Statistical reform: Evidence-based practice, meta-analyses, and single subject designs. *Psychology in the Schools*, *44*(5), 483–493.

Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and Tactics of Behavioral Research* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Jolivette, K., Wehby, J. H., Canale, J., & Massey, N. G. (2001). Effects of choice-making opportunities on the behavior of students with emotional and behavioral disorders. *Behavioral Disorders*, *26*(2), 131–145.

Kahng, S., Ingvarsson, E. T., Quigg, A. M., Seckinger, K. E., & Teichman, H. M. (2011). Defining and measuring behavior. In W. W. Fisher, C. C. Piazza, & H. S. Roane (Eds.), *Handbook of applied behavior analysis* (pp. 113–131). New York, NY: Guilford Pres.

Kauermann, G., & Carroll, R. J. (2001). A note on the efficiency of sandwich covariance matrix estimation. *Journal of the American Statistical Association*, *96*(456), 1387–1396.

Kazdin, A. E. (2011). *Single-Case Research Designs: Methods for Clinical and Applied Settings*. New York, NY: Oxford University Press.

Kelly, M. (1977). A review of the observational data-collection and reliability procedures reported in the Journal of Applied Behavior Analysis. *Journal of Applied Behavior Analysis*, *10*(1), 97–101.

Kennedy, C. H. (2004). *Single-Case Designs for Educational Research*. Boston, MA: Allyn & Bacon.

Kenward, M. G., & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983–97.

Kenward, M. G., & Roger, J. H. (2009). An improved approximation to the precision of fixed effects from restricted maximum likelihood. *Computational Statistics & Data Analysis*, *53*(7), 2583–2595. doi: 10.1016/j.csda.2008.12.013

Kern, L., Mantegna, M. E., Vorndran, C. M., Bailin, D., & Hilt, A. (2001). Choice of task sequence to reduce problem behaviors. *Journal of Positive Behavior Interventions*, *3*(1), 3–10.

Kraemer, H. C. (1979). One-zero sampling in the study of primate behavior. *Primates*, *20*(2), 237–244.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf,

D. M., & Shadish, W. R. (2012). Single-case intervention research design standards. *Remedial and Special Education*, *34*(1), 26–38. doi: 10.1177/0741932512452794

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, *15*(2), 124–44. doi: 10.1037/a0017736

Kratochwill, T. R., & Stoiber, K. C. (2002). Evidence-based interventions in school psychology: Conceptual foundations of the Procedural and Coding Manual of Division 16 and the Society for the Study of School Psychology Task Force. *School Psychology Quarterly*, *17*(4), 341–389. doi: 10.1521/scpq.17.4.341.20872

Kreft, I. G. G., & De Leeuw, J. (1998). *Introducing Multilevel Modeling*. Thousand Oaks, CA: Sage Publications, Inc.

Kulkarni, V. G. (2010). *Modeling and Analysis of Stochastic Systems*. Boca Raton, FL: Chapman & Hall/CRC.

Kyse, E. N. (2010). *Analyzing data from single case design studies: A demonstration and comparison of methods*. Unpublished doctoral dissertation, City University of New York.

Kyse, E. N., Rindskopf, D. M., & Shadish, W. R. (2011). *Analyzing data from single-case designs using multilevel models: A primer.*

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*(4), 963–74.

Lambert, M. C., Cartledge, G., Heward, W. L., & Lo, Y.-y. (2006). Effects of response cards on disruptive behavior and academic responding during math lessons by fourth-grade urban students. *Journal of Positive Behavior Interventions*, *8*(2), 88–99.

Larson, E. B., Ellsworth, A. J., & Oas, J. (1993). Randomized clinical trials in single patients during a 2-year period. *Journal of the American Medical Association*, *270*(22), 2708–12.

Laski, K. E., Charlop, M. H., & Schreibman, L. (1988). Training parents to use the natural language paradigm to increase their autistic children's speech. *Journal of Applied Behavior Analysis*, *21*(4), 391–400.

Levin, J. R., & Wampold, B. E. (1999). Generalized single-case randomization tests: Flexible analyses for a variety of situations. *School Psychology Quarterly*, *14*(1), 59–93.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*(1), 13–22.

Lindsey, J. K., & Lambert, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in Medicine*, *17*, 447–469.

Lindstrom, M. J., & Bates, D. M. (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, *83*(404), 1014–1022.

Lipsitz, S. R., & Fitzmaurice, G. M. (2008). Generalized estimating equations for longitudinal data analysis. In G. M. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 43–78). Boca Raton, FL: Chapman & Hall/CRC.

Lipsitz, S. R., Molenberghs, G., Fitzmaurice, G. M., & Ibrahim, J. G. (2000). GEE with Gaussian estimation of the correlations when data are incomplete. *Biometrics*, *56*(2), 528–36.

Longford, N. T. (2000). On estimating standard errors in multilevel analysis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, *49*(3), 389–398.

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. J. (2000). WinBUGS–A Bayesian modelling framework: Concepts, structure, and extensibility. *Statistics and Computing*, *10*, 325–337.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. *Behavior Modification*, *30*(5), 598.

Machalicek, W., O'Reilly, M. F., Beretvas, S. N., Sigafoos, J., Lancioni, G. E., Sorrells, A., ... Rispoli, M. (2008). A review of school-based instructional interventions for students with autism spectrum disorders. *Research in Autism Spectrum Disorders*, *2*(3), 395–416. doi: 10.1016/j.rasd.2007.07.001

MacKinnon, J. G., & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of Econometrics*, *29*, 305–325.

Maggin, D. M., O'Keeffe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985-2009. *Exceptionality*, *19*(2), 109–135. doi: 10.1080/09362835.2011.565725

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V., Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology*, *49*(3), 301–21. doi: 10.1016/j.jsp.2011.03.004

Mann, J., Ten Have, T. R., Plunkett, J. W., & Meisels, S. J. (1991). Time sampling: A methodological critique. *Child Development*, *62*(2), 227–241.

Marquis, J. G., Horner, R. H., Carr, E. G., Turnbull, A. P., Thompson, M., Behrens, G. A., ... Doolabh, A. (2000). A meta-analysis of positive behavior support. In R. Gersten, E. P. Schiller, & S. Vaughan (Eds.), *Contemporary special education research: Syntheses of the knowledge base on critical instructional issues* (pp. 137–178). Mahwah, NJ: Lawrence Erlbaum Associates.

Marschner, I. (2011). glm2: Fitting generalized linear models with convergence problems. *The R Journal*, *3/2*, 12–15.

Marschner, I. C. (2012). *glm2: Fitting Generalized Linear Models.* Retrieved from `http://cran.r-project.org/package=glm2`

Matyas, T. A., & Greenwood, K. M. (1996). Serial dependency in single-case time series. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 215–243). Mahwah, NJ: Lawrence Erlbaum.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics*, *11*(1), 59–67.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models* (2nd ed.). London, UK: Chapman & Hall.

Mendoza, J. L., & Mumford, M. (1987). Corrections for attenuation and range restriction on the predictor. *Journal of Educational Statistics*, *12*(3), 282–293.

Moes, D. R. (1998). Integrating choice-making opportunities within teacher-assigned academic tasks to facilitate the performance of children with autism. *Research and Practice for Persons with Severe Disabilities*, *23*(4), 319–328.

Molenaar, P. C. M. (2004). A manifesto on psychology as idiographic science: Bringing the person back into scientific psychology, this time forever. *Measurement*, *2*(4), 201–218.

Molenberghs, G., & Verbeke, G. (2005). *Models for Discrete Longitudinal Data*. Springer Verlag.

Morgan, P. L. (2006). Increasing task engagement using preference or choice-making. *Remedial and Special Education*, *27*(3), 176–187.

Morgan, P. L., & Sideridis, G. D. (2006). Contrasting the effectiveness of fluency interventions for students with or at risk for learning disabilities: A multilevel random coefficient modeling meta-analysis. *Learning Disabilities Research & Practice*, *21*(4), 191–210. doi: 10.1111/j.1540-5826.2006.00218.x

Morgan, S. L., & Winship, C. (2007). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. New York, NY: Cambridge University Press.

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105–125. doi: 10.1037//1082-989X.7.1.105

Mosteller, F. R., & Boruch, R. F. (2002). Overview and new directions. In F. Mosteller & R. Boruch (Eds.), *Evidence matters: Randomized trials in education research* (pp. 1–14). Washington, DC: Brookings Institution Press.

Mudford, O. C., Taylor, S. A., & Martin, N. T. (2009). Continuous recording and interobserver agreement algorithms reported in the Journal of Applied Behavior Analysis (1995-2005). *Journal of Applied Behavior Analysis*, *42*(1), 165–169. doi: 10.1901/jaba.2009.42-165

Murphy, G., & Goodall, E. (1980). Measurement error in direct observations: A comparison of common recording methods. *Behaviour Research and Therapy*, *18*, 147–150.

Musser, E. H., Bray, M. A., Kehle, T. J., & Jenson, W. R. (2001). Reducing disruptive behaviors in students with serious emotional disturbance. *School Psychology Review*, *30*(2), 294–304.

Neuhaus, J. M., Kalbfleisch, J. D., & Hauck, W. W. (1991). A comparison of cluster-specific and population-averaged approaches for analyzing correlated binary data. *International Statistical Review*, *59*(1), 25–35.

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children*, *71*(2), 137–148.

O'Keeffe, B. V., Slocum, T. A., Burlingame, C., Snyder, K., & Bundock, K. (2012). Comparing results of systematic reviews: Parallel reviews of research on repeated reading. *Education and Treatment of Children*, *35*(2), 333–366. doi: 10.1353/etc.2012.0006

Orr, L. L. (1999). *Social Experiments: Evaluating Public Programs with Experimental Methods*. Thousand Oaks, CA: Sage Publications, Inc.

Parker, R. I., & Vannest, K. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, *40*(4), 357–67. doi: 10.1016/j.beth.2008.10.006

Parker, R. I., Vannest, K., & Brown, L. (2009). The improvement rate difference for single-case research. *Exceptional Children*, *75*(2), 135–150.

Parker, R. I., & Vannest, K. J. (2012). Bottom-up analysis of single-case research designs. *Journal of Behavioral Education*, *21*(3), 254–265. doi: 10.1007/s10864-012-9153-1

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification*, *35*(4), 303–22. doi: 10.1177/0145445511399147

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, *42*(2), 284–99. doi: 10.1016/j.beth.2010.08.006

Patterson, D., & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, *58*(3), 545–554.

Peterson, S. M. P., Caniglia, C., & Royster, A. J. (2001). Application of choice-making intervention for a student with multiply maintained problem behavior. *Focus on Autism and Other Developmental Disabilities*, *16*(4), 240–246.

Pinheiro, J. C., & Bates, D. M. (1996). Unconstrained parameterizations for variance-covariance matrices. *Statistics and Computing*, *6*, 289–296.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-Effects Models in S and S-PLUS*. New York, NY: Springer Verlag.

Pinheiro, J. C., Bates, D. M., DebRoy, S., & Sarkar, D. (2012). *nlme: Linear and Nonlinear Mixed Effects Models*. Retrieved from `http://cran.r-project.org/package=nlme`

Poulson, R. S., Gadbury, G. L., & Allison, D. B. (2012). Treatment heterogeneity and individual qualitative interaction. *The American Statistician*, *66*(1), 16–24. doi: 10.1080/00031305.2012.671724

Powell, J., Martindale, A., & Kulp, S. (1975). An evaluation of time-sample measures of behavior. *Journal of Applied Behavior Analysis*, *4*(4), 463–469.

Powell, S., & Nelson, B. (1997). Effects of choosing academic assignments on a student with attention deficit hyperactivity disorder. *Journal of Applied Behavior Analysis*, *30*(1), 181–3. doi: 10.1901/jaba.1997.30-181

Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: A critical reexamination and new recommendations. *Psychological Methods*, *10*(2), 178–92. doi: 10.1037/1082-989X.10.2.178

Primavera, L. H., Allison, D. B., & Alfonso, V. C. (1996). Measurement of dependent variables. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 41–89). Mahwah, NJ: Lawrence Erlbaum.

Pustejovsky, J. E. (2012). *Converting from d to r to z when the design uses extreme groups, dichotomization, or experimental control.* Evanston, IL.

Pustejovsky, J. E. (2013, April). *Observation procedures and Markov chain models for estimating the prevalence and incidence of a behavior.* Poster presented at the annual meeting of the American Educational Research Association, San Francisco, CA.

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *GLLAMM Manual.* Berkeley, CA. Retrieved from `http://biostats.bepress.com/ucbbiostat/paper160`

Rapp, J. T., Colby-Dirksen, A. M., Michalski, D. N., Carroll, R. A., & Lindenberg, A. M. (2008). Detecting changes in simulated events using partial-interval recording and momentary time sampling. *Behavioral Interventions*, *23*, 237–269.

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

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Inc.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2011). *HLM 7.* Lincolnwood, IL: Scientific Software International.

Raudenbush, S. W., & Liu, X.-F. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, *6*(4), 387–401.

Reschly, A. L., Busch, T. W., Betts, J., Deno, S. L., & Long, J. D. (2009). Curriculum-based measurement oral reading as an indicator of reading achievement: A meta-analysis of the correlational evidence. *Journal of School Psychology*, *47*(6), 427–69. doi: 10.1016/j.jsp.2009.07.001

Rhoda, D. A., Murray, D. M., Andridge, R. R., Pennell, M. L., & Hade, E. M. (2011). Studies with staggered starts: Multiple baseline designs and group-randomized trials. *American Journal of Public Health*, *101*(11), 2164–9. doi:

10.2105/AJPH.2011.300264

Robins, J. M., & Hernán, M. A. (2009). Estimation of the causal effects of time-varying exposures. In G. M. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis* (pp. 553–599). Boca Raton, FL: Chapman & Hall/CRC.

Rogers, L. A., & Graham, S. (2008). A meta-analysis of single subject design writing intervention research. *Journal of Educational Psychology*, *100*(4), 879–906. doi: 10.1037/0022-0663.100.4.879

Rogosa, D., & Ghandour, G. (1991). Statistical models for behavioral observations. *Journal of Educational Statistics*, *16*(3), 157–252.

Romaniuk, C., Miltenberger, R., Conyers, C., Jenner, N., Jurgens, M., & Ringenberg, C. (2002). The influence of activity choice on problem behaviors maintained by escape versus attention. *Journal of Applied Behavior Analysis*, *35*(4), 349–62. doi: 10.1901/jaba.2002.35-349

Rosenbaum, P. R. (2002). *Observational Studies* (2nd ed.). New York, NY: Springer Verlag.

Rosenthal, R. (1994). Parametric measures of effect size. In *The handbook of research synthesis* (pp. 231–244). New York, NY: Russell Sage Foundation.

Ross, S. W., & Horner, R. H. (2009). Bully prevention in positive behavior support. *Journal of Applied Behavior Analysis*, *42*(4), 747–759. doi: 10.1901/jaba.2009.42-747

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. doi: 10.1037/h0037350

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*(1), 34–58.

Rubin, D. B. (1992). Meta-analysis: Literature synthesis or effect-size surface estimation? *Journal of Educational Statistics*, *17*(4), 363–374.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

Saddler, B., Behforooz, B., & Asaro, K. (2008). The effects of sentence-combining instruction on the writing of fourth-grade students with writing difficulties. *The Journal of Special Education*, *42*(2), 79–90. doi: 10.1177/0022466907310371

Sánchez-Meca, J., Marín-Martínez, F., & Chacón-Moscoso, S. (2003). Effect-size indices for dichotomized outcomes in meta-analysis. *Psychological Methods*, *8*(4), 448–67. doi: 10.1037/1082-989X.8.4.448

SAS Institute Inc. (2008). *SAS/STAT(R) 9.2 User's Guide*. Cary, NC: SAS Institute Inc.

Schoenfeld, W. N. (1972). Problems of modern behavior theory. *Integrative Physiological and Behavioral Science*, *7*(1), 33–65.

Schutte, N. S., Malouff, J. M., & Brown, R. F. (2008). Efficacy of an emotion-focused treatment for prolonged fatigue. *Behavior Modification*, *32*(5), 699–713. doi: 10 .1177/0145445508317133

Scruggs, T. E., & Mastropieri, M. A. (2012). PND at 25: Past, present, and future trends in summarizing single-subject research. *Remedial and Special Education*, *34*(1), 9–19. doi: 10.1177/0741932512440730

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research. *Remedial and Special Education*, *8*(2), 24–43.

Severini, T. A. (2005). *Elements of Distribution Theory*. Cambridge, England.

Seybert, S., Dunlap, G., & Ferro, J. (1996). The effects of choice-making on the problem behaviors of high school students with intellectual disabilities. *Journal of Behavioral Education*, *6*(1), 49–65.

Shadish, W. R., Brasil, I. C. C., Illingworth, D. A., White, K. D., Galindo, R., Nagler, E. D., & Rindskopf, D. M. (2009). Using UnGraph to extract data from image files: Verification of reliability and validity. *Behavior Research Methods*, *41*(1), 177–83. doi: 10.3758/BRM.41.1.177

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA: Houghton, Mifflin and Company.

Shadish, W. R., Hedges, L. V., Pustejovsky, J. E., Rindskopf, D. M., Boyajian, J., & Sullivan, K. J. (2012). Analyzing single-case designs: d, G, hierarchical models, Bayesian estimators, generalized additive models, and the hopes and fears of researchers about analyses. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and data-analysis advances.* Washington, DC: American Psychological Association.

Shadish, W. R., & Ragsdale, K. (1996). Random versus nonrandom assignment in

controlled experiments: Do you get the same answer? *Journal of Consulting and Clinical Psychology*, *64*(6), 1290–305.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation*, *113*, 95–109. doi: 10.1002/ev.217

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 188–196. doi: 10.1080/17489530802581603

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*(4), 971–80. doi: 10.3758/s13428-011-0111-y

Shager, H. M., Schindler, H. S., Magnuson, K. A., Duncan, G. J., Yoshikawa, H., & Hart, C. M. D. (2012). Can research design explain variation in Head Start research results? A meta-analysis of cognitive and achievement outcomes. *Educational Evaluation and Policy Analysis*, *35*(1), 76–95. doi: 10.3102/0162373712462453

Shavelson, R. J., & Towne, L. (Eds.). (2002). *Scientific Research in Education*. National Academies Press.

Shogren, K. A., Faggella-Luby, M. N., Bae, S. J., & Wehmeyer, M. L. (2004). The effect of choice-making as an intervention for problem behavior. *Journal of Positive Behavior Interventions*, *6*(4), 228–237.

Singer, J. D., & Willett, J. B. (2003). *Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence*. New York, NY: Oxford University Press.

Slavin, R. E. (2008). What works? Issues in synthesizing educational program evaluations. *Educational Researcher*, *37*(1), 5–14. doi: 10.3102/0013189X08314117

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, *17*(4), 510–50. doi: 10.1037/a0029312

Somers, M.-A., Zhu, P., Jacob, R., & Bloom, H. (2012). *The validity and precision of the comparative interrupted time series design and the difference-in-difference design in educational evaluation.* Retrieved from `https://appam.confex.com/appam/2012/webprogram/ExtendedAbstract/Paper1758/MDRC_CITS&DD_paper_draft_October2012.pdf`

Song, P. X. K., & Tan, M. (2000). Marginal models for longitudinal continuous proportional data. *Biometrics*, *56*(2), 496–502.

StataCorp. (2011). *Stata, Release 12: Longitudinal-Data/Panel-Data Reference Manual*. Stata Press.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, *50*(4), 1171–1177.

Suen, H. K., & Ary, D. (1984). Variables influencing one-zero and instantaneous time sampling outcomes. *Primates*, *25*(1), 89–94.

Suen, H. K., & Ary, D. (1986). A post hoc correction procedure for systematic errors in time-sampling duration estimates. *Journal of Psychopathology and Behavioral Assessment*, *8*(1), 31–38. doi: 10.1007/BF00960870

Sutton, A. J., & Higgins, J. P. T. (2008). Recent developments in meta-analysis. *Statistics in Medicine*, *27*, 625–650.

Swaminathan, H., Horner, R. H., Sugai, G., Smolkowski, K., Spaulding, S. A., & Hedges, L. V. (2010). *Application of generalized least squares regression to measure effect size in single-case research: A technical report.* Washington, DC.

Tapp, J., Wehby, J. H., & Ellis, D. (1995). A multiple option observation system for experimental studies : MOOSES. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 25–31.

Tvete, I. F., Olsen, I. C., Fagerland, M. W., Meland, N., Aldrin, M., Smerud, K. T., & Holden, L. (2012). An approach to combining parallel and cross-over trials with and without run-in periods using individual patient data. *Clinical Trials*, *9*(2), 164–75. doi: 10.1177/1740774511430714

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352–402. doi: 10.1037/a0028446

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly*, *18*(3), 325–346.

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 1–10.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment*

and Intervention, *2*(3), 142–151. doi: 10.1080/17489530802505362

Velicer, W. F. (1994). Time series models of individual substance abusers. In L. M. Collins & L. A. Seitz (Eds.), *Advances in data analysis for prevention intervention research* (pp. 264–301). National Institute on Drug Abuse.

Velicer, W. F., & McDonald, R. P. (1984). Time series analysis without model identification. *Multivariate Behavioral Research*, *19*, 33–47.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

von Mizener, B. H., & Williams, R. L. (2008). The effects of student choices on academic performance. *Journal of Positive Behavior Interventions*, *11*(2), 110–128. doi: 10.1177/1098300708323372

Wang, Y.-G., & Carey, V. J. (2004). Unbiased estimating equations from working correlation models for irregularly timed repeated measures. *Journal of the American Statistical Association*, *99*(467), 845–853. doi: 10.1198/016214504000001178

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika*, *61*(3), 439–447.

West, B. T., Welch, K. B., & Galecki, A. T. (2007). *Linear Mixed Models: A Practical Guide Using Statistical Software*. Boca Raton, FL: Chapman & Hall/CRC.

What Works Clearinghouse. (2012). *Phonological Awareness Training* (Tech. Rep.). Washington, DC: Institute for Education Sciences, U.S. Department of Education. Retrieved from `http://ies.ed.gov/ncee/wwc/pdf/intervention_reports/wwc_pat_060512.pdf`

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta analysis in individual-subject research. *Behavioral Assessment*, *11*(3), 281–296.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*(4), 817–838.

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education*, *44*(1), 18–28. doi: 10.1177/0022466908328009

Zeger, S. L. (1988). A regression model for time series of counts. *Biometrika*, *75*(4), 621–629.

Zeger, S. L., Liang, K.-Y., & Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, *44*(4), 1049–1060.

Zeileis, A. (2004). Econometric computing with HC and HAC covariance matrix estimators. *Journal of Statistical Software*, *11*(10), 1–17.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, *16*(9), 1–16.

Zucker, D. R., Ruthazer, R., & Schmid, C. H. (2010). Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: Methodologic considerations. *Journal of Clinical Epidemiology*, *63*(12), 1312–23. doi: 10.1016/j.jclinepi.2010.04.020

Zucker, D. R., Schmid, C. H., McIntosh, M. W., D'Agostino, R. B., Selker, H. P., & Lau, J. (1997). Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *Journal of Clinical Epidemiology*, *50*(4), 401–410.

Zuluaga, C. A., & Normand, M. P. (2008). An evaluation of the high-probability instruction sequence with and without programmed reinforcement for compliance with high-probability instructions. *Journal of Applied Behavior Analysis*, *41*(3), 453–457. doi: 10.1901/jaba.2008.41-453

APPENDIX A

A twice-adjusted estimator for the standardized mean difference

Chapter 3 described models and estimation methods for a standardized mean difference effect size having the general form

$$\delta_{AB} = \frac{\mathbf{p}'\boldsymbol{\gamma}}{\sqrt{\mathbf{r}'\boldsymbol{\theta}}},$$

where $\boldsymbol{\gamma}$ is a vector of fixed effect parameters and $\boldsymbol{\theta}$ a vector of variance component parameters, both defined for a hierarchical linear model described in Section 3.3, and $\mathbf{p}$ and $\mathbf{r}$ are constant vectors of appropriate length. The restricted maximum likelihood (RML) estimator of δ_{AB} is calculated by substituting RML estimates for the corresponding component parameters of the effect size. I described an adjusted estimator of the effect size based on approximating the sampling distribution of the RML estimator by a non-central t distribution. The development proceeded under the assumption that RML estimator of the squared denominator of the effect size was unbiased, so that only a degrees-of-freedom adjustment was needed for the RML estimator. In this Appendix, I present a further adjustment to the RML estimator, based on using an approximation to the bias of the squared denominator $\mathbf{r}'\boldsymbol{\theta}$. The approximation to the bias of $\hat{\boldsymbol{\theta}}$ was given by Cox and Snell (1968) and applied in a similar context by Kenward and Roger (2009).

The development proceeds as follows: I first derive an expression for the first-order bias of the restricted maximum likelihood estimators of the variance components $\boldsymbol{\theta}$ in a

mixed-effects linear model, as defined in Chapter 3. I then demonstrate that the bias is unchanged for linear re-parameterization of the variance components and show that the bias correction term ξ does not depend on the statistic it corrects. Based on these properties, the final section presents the twice-adjusted effect size estimator.

A.1. Bias of RML estimators of variance components

Define the joint null cumulants of the restricted maximum likelihood given in (3.31) as

$$\kappa_{s,tu} = \mathrm{E}\left[\frac{\partial l_R}{\partial \theta_s}\frac{\partial^2 l_R}{\partial \theta_t \partial \theta_u}\right], \qquad \kappa_{stu} = \mathrm{E}\left[\frac{\partial^3 l_R}{\partial \theta_s \partial \theta_t \partial \theta_u}\right].$$

Cox and Snell (1968, equation 20) show that

$$\mathrm{E}(\hat{\theta}_w - \theta_w) = \frac{1}{2}\sum_{s=1}^{r}\sum_{t=1}^{r}\sum_{u=1}^{r}\left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]^{-1}_{st}\left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]^{-1}_{wu}(\kappa_{s,tu} + \kappa_{t,su} + \kappa_{stu}) + O(m^{-1}), \tag{A.1}$$

for $w = 1, ..., r$. The third derivatives of l_R, along with their expected values, are needed for computing the approximate biases of the parameters. These are as follows:

$$\begin{aligned}\frac{\partial^3 l_R}{\partial \theta_s \partial \theta_t \partial \theta_u} &= \frac{1}{2}\mathbf{y}'\mathbf{Q}\left[\dddot{\mathbf{V}}_{stu} + [6]\dot{\mathbf{V}}_{\nu_1}\mathbf{Q}\dot{\mathbf{V}}_{\nu_2}\mathbf{Q}\dot{\mathbf{V}}_{\nu_3} - [3]\ddot{\mathbf{V}}_{\nu_1\nu_2}\mathbf{Q}\dot{\mathbf{V}}_{\nu_3} - [3]\ddot{\mathbf{V}}_{\nu_1}\mathbf{Q}\dot{\mathbf{V}}_{\nu_2\nu_3}\right]\mathbf{Q}\mathbf{y} \\ &\quad - \frac{1}{2}\mathrm{tr}\left(\mathbf{Q}\ddot{\mathbf{V}}_{stu} + \mathbf{Q}\dot{\mathbf{V}}_s\mathbf{Q}\dot{\mathbf{V}}_t\mathbf{Q}\dot{\mathbf{V}}_u + \mathbf{Q}\dot{\mathbf{V}}_t\mathbf{Q}\dot{\mathbf{V}}_s\mathbf{Q}\dot{\mathbf{V}}_u - [3]\mathbf{Q}\dot{\mathbf{V}}_{\nu_1}\mathbf{Q}\ddot{\mathbf{V}}_{\nu_2\nu_3}\right) \\ \mathrm{E}\left(\frac{\partial^3 l_R}{\partial \theta_s \partial \theta_t \partial \theta_u}\right) &= \kappa_{stu} = 2\mathrm{tr}\left(\mathbf{Q}\dot{\mathbf{V}}_s\mathbf{Q}\dot{\mathbf{V}}_t\mathbf{Q}\dot{\mathbf{V}}_u\right) - \frac{1}{2}\mathrm{tr}\left([3]\mathbf{Q}\dot{\mathbf{V}}_{\nu_1}\mathbf{Q}\ddot{\mathbf{V}}_{\nu_2\nu_3}\right),\end{aligned}$$

where $[6]\dot{\mathbf{V}}_{\nu_1}\mathbf{Q}\dot{\mathbf{V}}_{\nu_2}\mathbf{Q}\dot{\mathbf{V}}_{\nu_3}$ indicates the sum over the 6 permutations of $\nu_1, \nu_2, \nu_3 = s, t, u$. The mixed third-degree null cumulants are

$$\kappa_{s,tu} = \kappa_{s,ut} = \frac{1}{2}\mathrm{tr}(\mathbf{Q}\dot{\mathbf{V}}_s\mathbf{Q}\ddot{\mathbf{V}}_{tu}) - \mathrm{tr}(\mathbf{Q}\dot{\mathbf{V}}_s\mathbf{Q}\dot{\mathbf{V}}_t\mathbf{Q}\dot{\mathbf{V}}_u).$$

This allows the following simplification:

$$\kappa_{s,tu} + \kappa_{t,su} + \kappa_{stu} = -\frac{1}{2}\text{tr}(\mathbf{Q}\dot{\mathbf{V}}_u\mathbf{Q}\ddot{\mathbf{V}}_{st}).$$

Let $\left[\mathbf{C}_u^{\boldsymbol{\theta}}\right]_{s,t} = \text{tr}(\mathbf{Q}\dot{\mathbf{V}}_u\mathbf{Q}\ddot{\mathbf{V}}_{st})$ for $s,t = 1,...,r$, $c_u^{\boldsymbol{\theta}} = \text{tr}\left(\left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]^{-1}\mathbf{C}_u^{\boldsymbol{\theta}}\right)$ and $\mathbf{c}^{\boldsymbol{\theta}} = \left(c_1^{\boldsymbol{\theta}},...,c_r^{\boldsymbol{\theta}}\right)'$, where $\mathcal{I}_E^{\boldsymbol{\theta}}$ is the expected information matrix parameterized by $\boldsymbol{\theta}$. With this notation established, (A.1) can be expressed as

$$\begin{aligned}\text{E}(\hat{\theta}_w - \theta_w) &= -\frac{1}{4}\sum_{s=1}^{r}\sum_{t=1}^{r}\sum_{u=1}^{r}\left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]_{st}^{-1}\left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]_{wu}^{-1}\text{tr}(\mathbf{Q}\dot{\mathbf{V}}_u\mathbf{Q}\ddot{\mathbf{V}}_{st}) + O(m^{-1}) \\ &= -\frac{1}{4}\sum_{u=1}^{r}\left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]_{wu}^{-1}\text{tr}\left(\left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]^{-1}\mathbf{C}_u^{\boldsymbol{\theta}}\right) + O(m^{-1}).\end{aligned}$$

The bias of the full vector of variance components is therefore

$$\text{E}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) = -\left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]^{-1}\mathbf{c}^{\boldsymbol{\theta}}/4 + O\left(m^{-1}\right). \tag{A.2}$$

Finally, define $\mathbf{b}^{\boldsymbol{\theta}} = -\left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]^{-1}\mathbf{c}^{\boldsymbol{\theta}}/4$ to be the approximate bias of the RML variance component estimates. In the next section, I show that this expression is invariant to linear re-parameterization of the variance components.

A.2. Properties of the approximate bias correction

For use in a correction to the RML effect size estimator, it will be important that the approximate bias correction term is invariant to linear re-parameterization and does not depend on the parameter that it is intended to correct. Recall the re-parameterization described in Section 3.3.5, in which $\boldsymbol{\psi} = g(\boldsymbol{\theta})$, with first entry $\psi_1 = \mathbf{r}'\boldsymbol{\theta}$ and remaining entries $\boldsymbol{\psi}_*$ that create a one-to-one mapping from $\boldsymbol{\theta}$; let $h = g^{-1}$. Define $\mathbf{c}^{\boldsymbol{\psi}}$ and $\mathbf{C}_u^{\boldsymbol{\psi}}$,

$u = 1, ..., r$, just as the analogous terms involving $\boldsymbol{\theta}$. The invariance property can then be expressed as:

$$b_1^{\boldsymbol{\psi}} = \mathbf{r}'\mathbf{b}^{\boldsymbol{\theta}}. \tag{A.3}$$

Though this property of joint-null cumulants might be considered general knowledge, I have not been able to locate a proof. Therefore, I offer the following demonstration of (A.3). To begin, note that because $\psi_1 = \sum_{i=1}^{r} r_i\theta_i$ and $\boldsymbol{\theta} = h(\boldsymbol{\psi})$, it follows that

$$\sum_{i=1}^{r} r_i \frac{\partial h_i}{\partial \psi_s} = I(s = 1)$$

for $s = 1, ..., r$ and that

$$\sum_{i=1}^{r} r_i \frac{\partial^2 h_i}{\partial \psi_s \partial \psi_t} = 0 \tag{A.4}$$

for $s, t = 1, ..., r$. Let $\mathbf{H}_j = \partial^2 h_j / \partial \boldsymbol{\psi}' \partial \boldsymbol{\psi}$. Now observe that the $(s,t)^{th}$ entry of $\mathbf{C}_u^{\boldsymbol{\psi}}$ is

$$\operatorname{tr}\left(\mathbf{Q}\frac{\partial \mathbf{V}}{\partial \psi_u}\mathbf{Q}\frac{\partial^2 \mathbf{V}}{\partial \psi_s \partial \psi_t}\right) = \sum_{i,j,k} \operatorname{tr}(\mathbf{Q}\dot{\mathbf{V}}_i\mathbf{Q}\ddot{\mathbf{V}}_{jk})\frac{\partial h_i}{\partial \psi_u}\frac{\partial h_j}{\partial \psi_s}\frac{\partial h_k}{\partial \psi_t} + \sum_{i,j} \operatorname{tr}(\mathbf{Q}\dot{\mathbf{V}}_i\mathbf{Q}\dot{\mathbf{V}}_j)\frac{\partial h_i}{\partial \psi_u}\frac{\partial^2 h_j}{\partial \psi_s \partial \psi_t},$$

and so

$$\begin{aligned}
\mathbf{C}_u^{\boldsymbol{\psi}} &= \sum_{i,j,k} \left[\mathbf{C}_i^{\boldsymbol{\theta}}\right]_{j,k} \frac{\partial h_i}{\partial \psi_u}\frac{\partial h_j}{\partial \boldsymbol{\psi}'}\frac{\partial h_k}{\partial \boldsymbol{\psi}} + \sum_{i,j} \left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]_{i,j} \frac{\partial h_i}{\partial \psi_u}\mathbf{H}_j \\
c_u^{\boldsymbol{\psi}} &= \operatorname{tr}\left(\left[\mathcal{I}_E^{\boldsymbol{\psi}}\right]^{-1}\mathbf{C}_u^{\boldsymbol{\psi}}\right) = \sum_i c_i^{\boldsymbol{\psi}}\frac{\partial h_i}{\partial \psi_u} + \sum_{i,j}\left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]_{i,j}\frac{\partial h_i}{\partial \psi_u}\operatorname{tr}\left(\left[\mathcal{I}_E^{\boldsymbol{\psi}}\right]^{-1}\mathbf{H}_j\right) \\
\mathbf{c}^{\boldsymbol{\psi}} &= (\nabla h)'\mathbf{c}^{\boldsymbol{\theta}} + \sum_j (\nabla h)'\left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]_{\cdot,j} \operatorname{tr}\left(\left[\mathcal{I}_E^{\boldsymbol{\psi}}\right]^{-1}\mathbf{H}_j\right).
\end{aligned}$$

The relationship between $\left[\mathcal{I}^{\boldsymbol{\psi}}\right]^{-1}$ and $\left[\mathcal{I}^{\boldsymbol{\theta}}\right]^{-1}$ given in (3.43) then leads to

$$\left[\mathcal{I}_E^{\boldsymbol{\psi}}\right]^{-1}\mathbf{c}^{\boldsymbol{\psi}} = (\nabla g)\left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]^{-1}\mathbf{c}^{\boldsymbol{\theta}} + \sum_j (\nabla g)_{\cdot,j}\mathrm{tr}\left(\left[\mathcal{I}_E^{\boldsymbol{\psi}}\right]^{-1}\mathbf{H}_j\right).$$

Since $(\nabla g)_{1,j} = r_j$, it follows that

$$\left[\left(\mathcal{I}_E^{\boldsymbol{\psi}}\right)^{-1}\mathbf{c}^{\boldsymbol{\psi}}\right]_1 = \mathbf{r}'\left[\mathcal{I}_E^{\boldsymbol{\theta}}\right]^{-1}\mathbf{c}^{\boldsymbol{\theta}} + \mathrm{tr}\left(\left[\mathcal{I}_E^{\boldsymbol{\psi}}\right]^{-1}\sum_j \mathbf{r}_j\mathbf{H}_j\right),$$

the second term of which is zero by (A.4). Division by -4 then gives (A.3). Finally, note that this invariance property holds if one takes $c_u = \mathrm{tr}\left(\mathcal{I}^{-1}\mathbf{C}_u\right)$, regardless of which information matrix is used. However, it is required that the expected information matrix be used to compute the bias correction $\mathbf{b} = -\left[\mathcal{I}_E\right]^{-1}\mathbf{c}/4$.

Next, I show that $\mathbf{r}'\mathbf{b}^{\boldsymbol{\theta}}$ is proportional to $\mathbf{r}'\hat{\boldsymbol{\theta}}$, so that the bias correction does not depend on the parameter it is intended to correct. Because of the linear invariance property (A.3), it is sufficient to show that $b_1^{\boldsymbol{\psi}} \propto \psi_1$. Observe that

$$\begin{aligned}
\left[\mathbf{C}_1^{\boldsymbol{\psi}}\right]_{1,1} &= 0 & \left[\mathbf{C}_u^{\boldsymbol{\psi}}\right]_{1,1} &= 0\\
\left[\mathbf{C}_1^{\boldsymbol{\psi}}\right]_{t,1} &= \mathrm{tr}\left(\mathbf{R}\dot{\mathbf{W}}_t\right)/\hat{\psi}_1^2 & \left[\mathbf{C}_u^{\boldsymbol{\psi}}\right]_{t,1} &= \mathrm{tr}\left(\mathbf{R}\dot{\mathbf{W}}_u\mathbf{R}\dot{\mathbf{W}}_t\right)/\hat{\psi}_1\\
\left[\mathbf{C}_1^{\boldsymbol{\psi}}\right]_{s,t} &= \mathrm{tr}\left(\mathbf{R}\ddot{\mathbf{W}}_{st}\right)/\hat{\psi}_1 & \left[\mathbf{C}_u^{\boldsymbol{\psi}}\right]_{s,t} &= \mathrm{tr}\left(\mathbf{R}\dot{\mathbf{W}}_u\mathbf{R}\ddot{\mathbf{W}}_{st}\right)
\end{aligned}$$

for $s,t,u = 2,...,r$. To evaluate $c_u^{\boldsymbol{\psi}}$, the full inverse of the expected information matrix is needed; this can be written as:

$$\left[\mathcal{I}_E^{\boldsymbol{\psi}}\right]^{-1} = \begin{bmatrix} \frac{2\hat{\psi}_1^2}{N-p-\mathbf{k}_E'\mathbf{L}_E^{-1}\mathbf{k}_E} & \frac{-2\hat{\psi}\mathbf{k}_E'\mathbf{L}_E^{-1}}{N-p-\mathbf{k}_E'\mathbf{L}_E^{-1}\mathbf{k}_E} \\ \frac{-2\hat{\psi}\mathbf{L}_E^{-1}\mathbf{k}_E}{N-p-\mathbf{k}_E'\mathbf{L}_E^{-1}\mathbf{k}_E} & 2\left(\mathbf{L}_E - \frac{\mathbf{k}_E\mathbf{k}_E'}{N-p}\right)^{-1} \end{bmatrix}, \tag{A.5}$$

where $\mathbf{k}_E = \left[\text{tr}\left(\mathbf{R}\dot{\mathbf{W}}_2\right), ..., \text{tr}\left(\mathbf{R}\dot{\mathbf{W}}_r\right)\right]'$ and $\mathbf{L}_E$ is a matrix with $(s-1, t-1)^{th}$ entry $\text{tr}\left(\mathbf{R}\dot{\mathbf{W}}_s\mathbf{R}\dot{\mathbf{W}}_t\right)$ for $s, t = 2, ..., r$.

Now observe that $\mathbf{k}'_E\mathbf{L}_E^{-1}\left[\mathbf{C}_u^{\boldsymbol{\psi}}\right]'_{1,t=2,...,r} = \text{tr}\left(\mathbf{R}\dot{\mathbf{W}}_u\right)$. Therefore

$$c_1^{\boldsymbol{\psi}} = \text{tr}\left(\left[\mathcal{I}_E^{\boldsymbol{\psi}}\right]^{-1}\mathbf{C}_1^{\boldsymbol{\psi}}\right) = \frac{2}{\hat{\psi}_1}\left[\text{tr}\left(\left(\mathbf{L}_E - \frac{\mathbf{k}_E\mathbf{k}'_E}{N-p}\right)\mathbf{M}_E\right) - \frac{2\mathbf{k}'_E\mathbf{L}_E^{-1}\mathbf{k}_E}{N-p-\mathbf{k}'_E\mathbf{L}_E^{-1}\mathbf{k}_E}\right]$$

$$c_u^{\boldsymbol{\psi}} = \text{tr}\left(\left[\mathcal{I}_E^{\boldsymbol{\psi}}\right]^{-1}\mathbf{C}_u^{\boldsymbol{\psi}}\right) = 2\left[\text{tr}\left(\left(\mathbf{L}_E - \frac{\mathbf{k}_E\mathbf{k}'_E}{N-p}\right)\mathbf{O}_u\right) - \frac{2\text{tr}\left(\mathbf{R}\dot{\mathbf{W}}_u\right)}{N-p-\mathbf{k}'_E\mathbf{L}_E^{-1}\mathbf{k}_E}\right]$$

where $\mathbf{M}_E$ has $(s-1, t-1)$ entry $\text{tr}\left(\mathbf{R}\ddot{\mathbf{W}}_{st}\right)$ and $\mathbf{O}_u$ has $(s-1, t-1)$ entry $\text{tr}\left(\mathbf{R}\dot{\mathbf{W}}_u\mathbf{R}\ddot{\mathbf{W}}_{st}\right)$ for $s, t, u = 2, ..., r$. It can be seen that $c_1^{\boldsymbol{\psi}}$ is proportional to $1/\hat{\psi}_1$ and $c_u^{\boldsymbol{\psi}}$ is constant with respect to $\hat{\psi}_1$. Writing $\mathbf{c}_*^{\boldsymbol{\psi}} = (c_2^{\boldsymbol{\psi}}, ..., c_r^{\boldsymbol{\psi}})$ and using (A.5) once more,

$$-4b_1^{\psi} = \left(\mathcal{I}_E^{\boldsymbol{\psi}}\right)^{-1}_{1,\cdot}\mathbf{c}^{\boldsymbol{\psi}} = \frac{2\hat{\psi}_1^2}{N-p-\mathbf{k}'_E\mathbf{L}_E^{-1}\mathbf{k}_E}c_1^{\boldsymbol{\psi}} - \frac{2\hat{\psi}\mathbf{k}'_E\mathbf{L}_E^{-1}\mathbf{c}_*^{\boldsymbol{\psi}}}{N-p-\mathbf{k}'_E\mathbf{L}_E^{-1}\mathbf{k}_E} \propto \hat{\psi}_1.$$

Thus, the RML estimator $\hat{\psi}_1$ has bias that is approximately multiplicative.

A.3. A twice-adjusted effect size estimator

Using the expression for the approximate bias of $\hat{\boldsymbol{\theta}}$ from Cox and Snell (1968), I now define a twice-adjusted estimator for the effect size parameter. Define the constant

$$\xi = 1 - \frac{\mathbf{r}'\mathbf{b}^{\boldsymbol{\theta}}}{4\mathbf{r}'\hat{\boldsymbol{\theta}}}, \tag{A.6}$$

so that $\text{E}\left(\mathbf{r}'\hat{\boldsymbol{\theta}}\right) \approx \xi\mathbf{r}'\boldsymbol{\theta}$. Note that ξ does not involve $\mathbf{r}'\hat{\boldsymbol{\theta}}$ because $\mathbf{r}'\mathbf{b}^{\boldsymbol{\theta}}$ is proportional to $\mathbf{r}'\hat{\boldsymbol{\theta}}$. The degrees of freedom ν defined in Chapter 3 must now be adapted to incorporate the bias-correction term. Again from a theorem in Hedges (2007), the distribution of

$\left(\sqrt{\xi}\,\hat{\delta}_{AB}/\kappa\right)$ can be approximated by a non-central t distribution with ν^* degrees of freedom and non-centrality parameter $\left(\sqrt{\xi}\,\delta_{AB}/\kappa\right)$, where

$$\nu^* = \frac{2\xi^2(\mathbf{r}'\hat{\boldsymbol{\theta}})^2}{\mathbf{r}'\mathbf{C}(\hat{\boldsymbol{\theta}})\mathbf{r}}. \tag{A.7}$$

It follows further that a bias-corrected effect size estimator is given by

$$g^*_{AB} = J(\nu^*) \times \sqrt{\xi} \times \hat{\delta}_{AB}, \tag{A.8}$$

where $J(x) = 1 - 3/(4x - 1)$, and that g^*_{AB} has approximate variance

$$\mathrm{Var}(g_{AB}) \approx J(\nu^*)^2 \left[\frac{\nu^*\kappa^2}{\nu^* - 2} + \xi\,\delta^2_{AB}\left(\frac{\nu^*}{\nu^* - 2} - \frac{1}{J(\nu^*)^2}\right)\right]. \tag{A.9}$$

As with the singly-adjusted estimator, substituting g^*_{AB} for δ_{AB} produces an estimate of the variance of g^*_{AB}.

APPENDIX B

Distribution theory for direct observation recording procedures

B.1. Expectation of interval recording data

The expectations of the interval recording procedures can be derived by conditioning on the state of the alternating renewal process (ARP) at the beginning of each interval. First consider that the residual interim time at time-point t can be expressed as $B(t) = \sum_{u=0}^{N(t)} (D_u + E_u) - t$. In an equilibrium ARP, the conditional distribution of the residual interim time, given that $Y(t) = 0$, is

$$\Pr\left(B(t) \leq x | Y(t) = 0\right) = \frac{1}{\lambda} \int_0^x \tilde{F}_E(t) dt \tag{B.1}$$

(Kulkarni, 2010, Thm. 9.17). It follows that the expected value of a recorded datum generated by partial interval recording is

$$\begin{aligned} \mathrm{E}(U_k) &= \sum_{a=0}^{1} \Pr\left[0 < \int_0^{l^-} Y\left(l + \frac{(k-1)L}{K}\right) dt \middle| Y\left(\frac{(k-1)L}{K}\right) = a\right] \\ &\qquad \times \Pr\left[Y\left(\frac{(k-1)L}{K}\right) = a\right] \\ &= \phi + (1-\phi) \Pr\left[B\left(\frac{(k-1)L}{K}\right) < l \middle| Y\left(\frac{(k-1)L}{K}\right) = 0\right] \\ &= \phi + \zeta \int_0^{l^-} \tilde{F}_E(t) dt. \end{aligned} \tag{B.2}$$

The the reported datum Y^P is the mean of the recorded data, and thus has expectation equal to (B.2), as given in Table 5.3.

The expectation of a reported datum generated by whole interval recording can be derived from the fact that the procedure is equivalent to applying partial interval recording to the absence of the behavior rather than its presence. If $\mathrm{E}(U_k; F_D(\mu), F_E(\lambda))$ denotes the expectation of U_k when $D_1 \sim F_D(\mu)$ and $E_1 \sim F_E(\lambda)$, then

$$\begin{aligned}\mathrm{E}(W_k; F_D(\mu), F_E(\lambda)) &= 1 - \mathrm{E}(U_k; F_E(\lambda), F_D(\mu)) \\ &= 1 - \left[(1-\phi) + \zeta \int_0^{l^-} \tilde{F}_D(t)dt\right] \\ &= \phi - \zeta \int_0^{l^-} \tilde{F}_D(t)dt.\end{aligned} \tag{B.3}$$

Further, $\mathrm{E}\left(Y^W\right) = \mathrm{E}(W_k)$, as given in Table 5.3.

B.2. Bounds for the bias of a partial interval recording datum

Although a partial interval recording datum is biased as a measure of prevalence, its bias can be bounded under certain assumptions about the event durations and interim times. Assume that $\mu_L^* \leq \mu \leq \mu_U^*$ and $F_E(l) \leq p^*$ for known μ_L^*, μ_U^*, and p^*. Let $B = \int_0^{l^-} \tilde{F}_E(t)dt$, so that $\mathrm{E}\left(Y^P\right) = \frac{\mu+B}{\mu+\lambda}$. It follows that the proportionate bias of Y^P is bounded by

$$\frac{B}{\mu_U^* + B} \leq \frac{\mathrm{E}\left(Y^P\right) - \phi}{\mathrm{E}\left(Y^P\right)} \leq \frac{B}{\mu_L^* + B}.$$

Now, since $F_E(l) \leq p^*$, it follows that $(1-p^*) \leq \tilde{F}_E(t) \leq 1$ for all $t \leq l$, and further that $(1-p^*)l \leq B \leq l$. Combining the two inequalities yields

$$\frac{(1-p^*)l}{\mu_U^* + (1-p^*)l} \leq \frac{\mathrm{E}\left(Y^P\right) - \phi}{\mathrm{E}\left(Y^P\right)} \leq \frac{l}{\mu_L^* + l}, \tag{B.4}$$

as given in (5.4).

A bound for the bias of a partial interval recording datum as a measure of incidence can be constructed under assumptions about the maximum average event duration and the distribution of interim times, following a similar argument as above. Assume that $\mu \leq \mu_U^*$ and $F_E(l) \leq p^*$ for known μ_U^* and p^*. With B as previously defined, the proportionate bias of Y^P is bounded by

$$\frac{B-1}{B} \leq \frac{\mathrm{E}\left(Y^P\right) - \zeta}{\mathrm{E}\left(Y^P\right)} \leq \frac{\mu_U^* + B - 1}{\mu_U^* + B}.$$

Since $(1-p^*)l \leq B \leq l$, it follows that

$$\frac{(1-p^*)l - 1}{(1-p^*)l} \leq \frac{\mathrm{E}\left(Y^P\right) - \zeta}{\mathrm{E}\left(Y^P\right)} \leq \frac{\mu_U^* + l - 1}{\mu_U^* + l}, \tag{B.5}$$

as given in (5.5).

B.3. Bounds for the log-interim ratio

Estimable bounds for the log-interim ratio ω^λ can be constructed under assumptions about the behavior stream process in both baseline and treatment phases. Assume that the intervention does not change the average event duration, so that $\mu^0 = \mu^1 = \mu$. Further suppose that the interim times in each treatment condition follow exponential distributions, so that $F_E\left(t|\lambda^t\right) = 1 - \exp\left(-t/\lambda^t\right)$, $t = 0, 1$. Denote the expected value of

a partial interval recording datum made under condition t as

$$\pi_t^P = \mathrm{E}\left(\bar{y}_t^P\right) = 1 - \frac{\lambda^t \exp\left(-l/\lambda^t\right)}{\mu + \lambda^t}.$$

Note that for fixed π_t^P,

$$\mu = f_{\pi_t^P}(\lambda^t) = \frac{\lambda^t \left[\exp\left(-l/\lambda^t\right) - 1 + \pi_t^P\right]}{1 - \pi_t^P}, \tag{B.6}$$

with

$$\frac{\partial f_{\pi_t^P}}{\partial \lambda^t} = \frac{\exp\left(-l/\lambda^t\right)}{1 - \pi_t^P} + \frac{l \exp\left(-l/\lambda^t\right)}{\lambda^t(1 - \pi_t^P)} - 1 = \frac{(\mu + l)\lambda^t + l\mu}{\left(\lambda^t\right)^2}.$$

Because μ must be greater than zero, $f_{\pi_t^P}(\lambda^t)$ takes values in $\left(\frac{-l}{\log(1-\pi_t^P)}, \infty\right)$ and is strictly increasing over its domain. Define $g\left(\cdot, \pi_t^P\right) = f_{\pi_t^P}^{-1}$, so $g\left(\mu, \pi_t^P\right) = \lambda^t$. The log-interim ratio is therefore $\omega^\lambda = \ln g\left(\mu, \pi_0^P\right) - \ln g\left(\mu, \pi_1^P\right)$. Observe that

$$\frac{\partial \omega^\lambda}{\partial \mu} = \frac{\partial g\left(\mu, \pi_0^P\right)/\partial \mu}{g\left(\mu, \pi_0^P\right)} - \frac{\partial g\left(\mu, \pi_1^P\right)/\partial \mu}{g\left(\mu, \pi_1^P\right)} = \frac{\lambda^0}{l\mu + (\mu + l)\lambda^0} - \frac{\lambda^1}{l\mu + (\mu + l)\lambda^1}. \tag{B.7}$$

Also note that

$$\frac{\partial g\left(\mu, \pi_t^P\right)}{\partial \pi_t^P} = \frac{-\lambda^t(\mu + \lambda^t)^2}{[l\mu + (\mu + l)\lambda^t]\exp(-l/\lambda^t)} < 0 \tag{B.8}$$

and

$$\lim_{\mu \to 0} g\left(\mu, \pi_t^P\right) = \frac{l}{-\ln\left(1 - \pi_t^P\right)},$$

$$\lim_{\mu \to \infty} \frac{g\left(\mu, \pi_t^P\right)}{\mu} = \lim_{\mu \to \infty} \frac{1 - \pi_t^P}{\exp\left[-l/g\left(\mu, \pi_t^P\right)\right] - 1 + \pi_t^P} = \frac{1 - \pi_t^P}{\pi_t^P}.$$

Now suppose that $\pi_0^P > \pi_1^P$. It follows from (B.8) that $\lambda^0 < \lambda^1$ and further that $\partial\omega^\lambda/\partial\mu$ from (B.7) is strictly decreasing in μ. Therefore

$$\sup_{\mu>0} \omega^\lambda\left(\mu, \pi_0^P, \pi_1^P\right) = \lim_{\mu\to 0}\left[\ln g\left(\mu, \pi_0^P\right) - \ln g\left(\mu, \pi_1^P\right)\right] = \mathrm{cll}\left(\pi_1^P\right) - \mathrm{cll}\left(\pi_0^P\right)$$

$$\inf_{\mu>0} \omega^\lambda\left(\mu, \pi_0^P, \pi_1^P\right) = \lim_{\mu\to \infty}\left[\ln g\left(\mu, \pi_0^P\right) - \ln g\left(\mu, \pi_1^P\right)\right] = \mathrm{logit}\left(\pi_1^P\right) - \mathrm{logit}\left(\pi_0^P\right),$$

by which (5.18) follows. If $\pi_0^P \leq \pi_1^P$ then, by a similar argument,

$$\mathrm{cll}\left(\pi_1^P\right) - \mathrm{cll}\left(\pi_0^P\right) < \omega^\lambda < \mathrm{logit}\left(\pi_1^P\right) - \mathrm{logit}\left(\pi_0^P\right),$$

as given in (5.19). It is rather interesting that the bounds do not depend on the interval length l.

B.4. Moments under an alternating poisson process

Table 6.1 reports the variance of the reported datum from five types of recording procedures used for observation of free-operant behavior, under the assumptions of the alternating poisson process (APP). Here I provide derivations of these expressions. Throughout this section, I assume that the event durations are exponentially distributed with mean μ and that the interim times are also exponentially distributed with mean λ. As in Chapter 5, I denote the prevalence $\phi = \frac{\mu}{\mu+\lambda}$ and the incidence $\zeta = \frac{1}{\mu+\lambda}$; for ease of notation, I also employ an alternate parameterization where $\rho = \frac{1}{\mu} + \frac{1}{\lambda} = \frac{\zeta}{\phi(1-\phi)}$.

Under the assumptions of the alternating poisson process, $\{Y(t), t \geq 0\}$ is a continuous time Markov chain, having the property that

$$\text{(B.9)} \qquad Pr(Y(s+t) = 1 | Y(s) = a, Y(r) : 0 \leq r < s) = Pr(Y(s+t) = 1 | Y(s) = a)$$
$$= Pr(Y(t) = 1 | Y(0) = a)$$

for $a \in \{0, 1\}$ and $s, t \geq 0$ (Kulkarni, 2010, Thm. 6.1). Denote the transition probabilities of this continuous time Markov chain by

$$\text{(B.10)} \qquad p_0(t) = Pr(Y(t) = 1 | Y(0) = 0) = \phi \left(1 - e^{-\rho t}\right)$$

$$\text{(B.11)} \qquad p_1(t) = Pr(Y(t) = 1 | Y(0) = 1) = (1 - \phi) e^{-\rho t} + \phi$$

(*ibid.*, Equation 6.17, p. 207).

B.4.1. Event counting

Because the APP is a special case of the alternating renewal process, the expectation of a reported datum generated by an event counting procedure is equal to the incidence times the session length: $\mathrm{E}\left(Y^E\right) = \zeta L$. The variance and higher moments of Y^E can be evaluated from the Laplace transform of its probability generating function. I proceed as far as possible in general terms, before specializing the expressions to the case of the alternating poisson process. Letting $G_u = D_u + E_u$ for $u = 1, 2, 3, ...$, the Laplace transform of G_1 is $f_G^*(s) = f_D^*(s) f_E^*(s)$. From Cox (1962, Eq. 3.2.6), the Laplace transform of the

probability generating function is therefore

$$G^*(s,\upsilon) = \frac{1}{s} + \frac{\zeta(\upsilon-1)[1-f_G^*(s)]}{s^2[1-\upsilon f_G^*(s)]}.$$

For a given session length L, the r^{th} factorial moment of Y^E is given by the r^{th} derivative of $G(s,\upsilon)$ with respect to υ, evaluated at $s=L$, $\upsilon=1$:

$$E\left(\frac{Y^E!}{(Y^E-r)!}\right) = G^{(r)}(L,1).$$

Thus, for $r = 1, 2, 3, ...,$

$$G^{*(r)}(s,\upsilon) = \frac{\zeta r!\left[1-f_G^*(s)\right]^2\left[f_G^*(s)\right]^{r-1}}{s^2\left[1-\upsilon f_G^*(s)\right]^{r+1}}, \qquad G^{*(r)}(s,1) = \frac{\zeta r!\left[f_G^*(s)\right]^{r-1}}{s^2\left[1-f_G^*(s)\right]^{r-1}}$$

and the Laplace transform of the variance (the second cumulant) of Y^E is therefore

$$\kappa_2^*(s) = G^{*(2)}(s,1) + \frac{\zeta}{s^2} - \frac{2\zeta^2}{s^3} = \frac{2\zeta f_D^*(s) f_E^*(s)}{s^2\left[1-f_D^*(s)f_E^*(s)\right]} + \frac{\zeta}{s^2} - \frac{2\zeta^2}{s^3}, \tag{B.12}$$

which can be evaluated for specific cases.

For the alternating poisson process, the Laplace transforms of D_1 and E_1 are

$$f_D^*(s) = \frac{1}{\mu s+1}, \qquad f_E^*(s) = \frac{1}{\lambda s+1}.$$

From (B.12), the Laplace transform of the variance (the second cumulant) is therefore

$$\begin{aligned}\kappa_2^*(s) &= \frac{2\zeta^2}{s^3\left(s+\frac{\zeta}{\phi(1-\phi)}\right)} + \frac{\zeta}{s^2} - \frac{2\zeta^2}{s^3} \\ &= \frac{\zeta}{s^2} + 2\zeta\phi(1-\phi)\left(\frac{\phi(1-\phi)}{\zeta s} - \frac{\phi(1-\phi)}{\zeta\left(s+\frac{\zeta}{\phi(1-\phi)}\right)} - \frac{1}{s^2}\right),\end{aligned}$$

It follows that

$$\text{(B.13)}\quad \text{Var}(Y^E) = \kappa_2(L) = \zeta L\left[\phi^2 + (1-\phi)^2\right] + 2\phi^2(1-\phi)^2\left[1 - \exp\left(\frac{-\zeta L}{\phi(1-\phi)}\right)\right].$$

For long session lengths, $\text{Var}\left(Y^E\right)$ will be approximately proportional to its mean.[1]

B.4.2. Continuous recording

The reported datum from a continuous recording procedure provides an unbiased estimate of prevalence:$\text{E}\left(Y^C\right) = \phi$. The variance of Y^C can be derived directly by using the properties of continuous time Markov chains:

$$\begin{aligned}\text{(B.14)}\qquad \text{Var}\left(Y^C\right) &= \text{E}\left[(Y^C)^2\right] - \phi_i^2 \\ &= \frac{1}{L^2}\text{E}\left[\int_0^{L^-}\int_0^{L^-} Y(s)Y(t)dsdt\right] - \phi^2 \\ &= \frac{1}{L^2}\int_0^{L^-}\int_0^{L^-} \text{E}\left[Y(s)Y(t)\right]dsdt - \phi^2 \\ &= \frac{2}{L^2}\int_0^{L^-}\int_0^{t} \Pr[Y(s)=1]\Pr[Y(t)=1|Y(s)=1]dsdt - \phi^2\end{aligned}$$

[1]Rogosa and Ghandour (1991, p. 184) proposed to approximate (B.13) using:

$$\text{Var}(Y^E) \approx \zeta L\left[\phi^2 + (1-\phi)^2\right] + 2\phi^2(1-\phi)^2.$$

$$
\begin{aligned}
&= \frac{2}{L^2}\int_0^{L^-}\int_0^t \phi p_1(t-s)dsdt - \phi^2 \\
&= \frac{2\phi}{L^2}\int_0^{L^-}\int_0^t \left[(1-\phi)e^{-\rho(t-s)} + \phi\right] dsdt - \phi^2 \\
&= \frac{2\phi(1-\phi)}{\rho L}\left(1 - \frac{1-e^{-\rho L}}{\rho L}\right) \\
&= \frac{2\phi^2(1-\phi)^2}{\zeta L}\left(1 - \frac{\phi(1-\phi)\left[1-\exp\left(\frac{-\zeta L}{\phi(1-\phi)}\right)\right]}{\zeta L}\right).
\end{aligned}
$$

Note that if $\frac{\zeta L}{\phi(1-\phi)}$ is large, then $\mathrm{Var}\left(Y^C\right)$ will be approximately proportional to $\phi^2(1-\phi)^2$.[2]

B.4.3. Momentary time sampling

The moments of the reported datum from a momentary time sampling procedure can be derived by considering the recorded data $X_1, ..., X_K$. First, observe that $\mathrm{E}(X_k) = \mathrm{E}[Y(kL/K)] = \phi$, and so the reported datum provides an unbiased estimate of prevalence: $\mathrm{E}\left(Y^M\right) = \phi$. To find the variance of Y^M, note that for $k > j$,

$$
\begin{aligned}
\mathrm{Cov}(X_j, X_k) &= \Pr(X_j = 1)\left[\Pr(X_k = 1 | X_j = 1) - \Pr(X_k = 1)\right] \\
&= \phi\left[p_1((k-j)L/K) - \phi\right] \\
&= \phi(1-\phi)e^{-\rho(k-j)L/K} = \phi(1-\phi)\exp\left(\frac{-\zeta(k-j)L}{\phi(1-\phi)K}\right).
\end{aligned}
$$

[2]Rogosa and Ghandour (1991, p. 230) gave the following approximate expression for the variance of Y^C:

$$
\mathrm{Var}(Y^C) \approx \frac{2\phi(1-\phi)}{\rho^2 L^2}\left(\rho L + \frac{\phi(1-\phi)(1-2\phi)}{[(1-\phi)^2 + \phi^2]^2}\right).
$$

Therefore

$$\text{(B.15)} \qquad \begin{aligned} \mathrm{Var}(Y^M) &= \frac{1}{K^2}\left[\sum_{k=1}^{K}\mathrm{Var}(X_k) + 2\sum_{k=1}^{K-1}(K-k)\mathrm{Cov}(X_1, X_{1+k})\right] \\ &= \frac{\phi(1-\phi)}{K}\left[1 + \frac{2}{K}\sum_{k=1}^{K-1}(K-k)\exp\left(\frac{-\zeta k L}{\phi(1-\phi)K}\right)\right] \end{aligned}$$

A similar approach could be followed to derive the variance of Y^M under a non-exponential equilibrium ARPs, though the transition probabilities $p_a(t)$ might involve considerably more complex expressions.[3]

Using a result from Good (1961), Rogosa and Ghandour (1991, Eq. 5.1) offered the following approximation to (B.15):

$$\text{(B.16)} \qquad \mathrm{Var}(Y^M) \approx \frac{\phi(1-\phi)}{K}\left[\frac{2}{1-\exp\left(\frac{-\zeta L}{\phi(1-\phi)K}\right)} - 1\right].$$

This approximation is the limit of $K\mathrm{Var}\left(Y^M\right)$ as K and L increase in fixed proportion, that is, as momentary time samples are taken at fixed intervals for an increasingly long session.

B.4.4. Partial interval recording

The expectation of a reported datum generated by a partial interval recording procedure can be found by evaluating (B.2) using the specific form of the cumulative distribution function F_E. Recall that l denotes the length of the active portion each interval. For the

[3]Simple expressions for the Laplace transforms of the transition probabilities are available (see Kulkarni, 2010, Thm. 9.14), but these are generally difficult to invert analytically.

alternating poisson process, $\tilde{F}_E(t) = \exp(l/\lambda)$, and so

$$\text{(B.17)} \qquad \mathrm{E}\left(Y^P\right) = \mathrm{E}(U_k) = 1 - (1-\phi)\exp\left(\frac{-\zeta l}{(1-\phi)}\right) = \frac{\mu + \lambda\left(1 - e^{-l/\lambda}\right)}{\mu + \lambda}.$$

To find the variance of the Y^P, I first evaluate $E(U_h U_k)$ for $1 \leq h < k \leq K$. Let $t_0 = (h-1)L/K$ denote the beginning of the h^{th} interval, $t_1 = t_0 + l$ denote the end of the active portion of the h^{th} interval, and $t_2 = (k-1)L/K$ denote the beginning of the k^{th} interval.

I will first need to find $\Pr\left(Y\left(t_1\right) = 1 \middle| U_h = 1\right)$. Observe that

$$\begin{aligned}
\Pr\left(Y(t_1) = 1, U_h = 1 | Y(t_0) = 1\right) &= \Pr\left(Y(t_1) = 1 | Y(t_0) = 1\right) \\
&= p_1(l) = (1-\phi)e^{-\rho l} + \phi, \\
\Pr\left(Y(t_1) = 1, U_h = 1 | Y(t_0) = 0\right) &= \Pr(Y(t_1) = 1 \cap Y(s) = 1, t_0 \leq s < t_1 | Y(t_0) = 0) \\
&= \int_0^{l^-} p_1\left(l - t\right) f_E(t) dt \\
&= \phi\left(1 - e^{-\rho l}\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
\Pr\left(Y(t_1) = 1 \cap U_h = 1\right) &= \phi \Pr\left(Y(t_1) = 1, U_h = 1 | Y(t_0) = 1\right) \\
&\quad + (1-\phi)\Pr\left(Y(t_1) = 1, U_h = 1 | Y(t_0) = 0\right) = \phi \\
\Pr\left(Y(t_1) = 1 | U_h = 1\right) &= \frac{\phi}{1 - (1-\phi)e^{-\rho\phi l}}.
\end{aligned}$$

Conditioning on $Y(t_1)$,

$$\begin{aligned}\Pr\left(Y(t_2)=1|U_h=1\right) &= \Pr\left(Y(t_1)=0|U_h=1\right)p_0(t_2-t_1)\\ &\quad + \Pr\left(Y(t_1)=1|U_h=1\right)p_1(t_2-t_1)\\ &= \phi + \frac{\phi(1-\phi)e^{-\rho(t_2-t_1)-\rho\phi l}}{1-(1-\phi)e^{-\rho\phi l}}.\end{aligned}$$

Now conditioning on $Y(t_2)$,

$$\begin{aligned}\Pr\left(U_k=1|U_h=1\right) &= \sum_{a=0}^{1}\Pr\left(Y(t_2)=a|X_h=1\right)Pr\left(U_k=1|Y(t_2)=a\right)\\ &= 1-e^{-\rho\phi l}+e^{-\rho\phi;}\Pr\left(Y(t_2)=1|U_h=1\right)\\ &= 1-(1-\phi_i)e^{-\rho\phi l}+\frac{\phi(1-\phi)e^{-\rho(t_2-t_1)-2\rho\phi l}}{1-(1-\phi)e^{-\rho\phi l}}.\end{aligned}$$

It therefore follows that

$$\begin{aligned}\mathrm{Cov}(U_h,U_k) &= \Pr(U_h=1)\left[\Pr\left(U_k=1|U_h=1\right)-\Pr(U_k=1)\right]\\ &= \phi(1-\phi)\exp\left[-\rho(k-h)L/K-(2\phi-1)\rho l\right].\end{aligned}$$

Thus,

$$\begin{aligned}\mathrm{Var}\left(Y^P\right) &= \frac{1}{K}\mathrm{Var}(U_1)+\frac{2}{K^2}\sum_{k=1}^{K-1}(K-k)\mathrm{Cov}(U_1,U_{k+1})\\ &= \frac{1}{K}\left[1-(1-\phi)e^{-\rho\phi l}\right](1-\phi)e^{-\rho\phi l}\\ &\qquad\times\left[1+\frac{2\phi e^{(1-\phi)\rho l}}{K\left[1-(1-\phi)e^{-\rho\phi l}\right]}\sum_{k=1}^{K-1}(K-k)\exp\left(\frac{-\rho kL}{K}\right)\right].\end{aligned} \tag{B.18}$$

Note that the derivation of this expression depends strongly on the independence of increments in the alternating poisson process; this will inhibit generalizations to alternating renewal processes using event duration and interim time distributions other than exponentials.

Finally, observe that if the interval length L/K is held fixed, then as the number of intervals K increases,

(B.19)
$$\mathrm{Var}\left(Y^P\right) \approx \frac{1}{K}\left[1-(1-\phi)e^{-\rho\phi l}\right](1-\phi)e^{-\rho\phi l}\left(1+\frac{2\phi e^{-\rho L/K+\rho(1-\phi)l}}{\left[1-(1-\phi)e^{-\rho\phi l}\right](1-e^{-\rho L/K})}\right).$$

B.4.5. Whole interval recording

The expectation of a reported datum generated by whole interval recording can be found directly from (B.3) to be

(B.20)
$$\mathrm{E}\left(Y^W\right) = \mathrm{E}(W_1) = \phi\exp\left(\frac{-\zeta l}{\phi}\right) = \frac{\mu e^{-l/\mu}}{\mu+\lambda}.$$

The variance of Y^W can be found by using the fact that whole interval recording is equivalent to partial interval recording for the absence of a behavior, which implies that $\mathrm{Var}\left(Y^W;\mu,\lambda\right) = \mathrm{Var}\left(Y^P;\lambda,\mu\right)$. Thus,

(B.21)
$$\mathrm{Var}\left(Y^W\right) = \frac{1}{K}\phi e^{-\rho(1-\phi)l}\left(1-\phi e^{-\rho(1-\phi)l}\right) \times\left[1+\frac{2(1-\phi)e^{\phi\rho l}}{K\left[1-\phi e^{-\rho(1-\phi)l}\right]}\sum_{k=1}^{K-1}(K-k)\exp\left(\frac{-\rho k L}{K}\right)\right].$$

For fixed interval length L/K and large K,

$$\text{(B.22)}\quad \operatorname{Var}\left(Y^W\right) \approx \frac{1}{K}\phi e^{-\rho(1-\phi)l}\left(1-\phi e^{-\rho(1-\phi)l}\right)\left(1+\frac{2(1-\phi)e^{-\rho L/K+\rho\phi l}}{\left(1-\phi e^{-\rho(1-\phi)l}\right)\left(1-e^{-\rho L/K}\right)}\right).$$

APPENDIX C

Equilibrium alternating renewal process simulations

This appendix collects several sets of simulation results regarding proposed effect size estimators for behavioral observation data, as described in Chapters 5 and 6. All of the studies involved simulating behavior stream data based on an alternating renewal process, then calculating reported data points based on the realized behavior stream.

Recall that the alternating renewal process entails only first-moment assumptions about the event duration and interim time distributions. For purposes of simulating the behavior stream, it was necessary to choose specific parametric forms for these distributions. In the simulations described in the following sections, I considered three different possibilities. First, I simulated an alternating poisson process in which event durations and interim times are both exponentially distributed, with rates $1/\mu$ and $1/\lambda$, respectively; these simulations are labeled "Exp-Exp." Second, I used gamma-distributed event durations and interim times; specifically, I assumed that event durations are distributed as $\Gamma(3, \mu/3)$ and interim times are distributed as $\Gamma(3, \lambda/3)$, where $\Gamma(k, \theta)$ is a gamma distribution with shape k and scale θ. These simulations are labeled "Γ_3-Γ_3." Finally, I assumed that event durations have constant duration μ and that interim times follow a $\Gamma(3, \lambda/3)$ distribution; these simulations are labeled "Const-Γ_3." For each of the simulations, I used within-phase sample sizes of $n = 4, 8, 12$, corresponding to short, medium, and long phase lengths. I set the total session length equal to $L = 1$, so that values of incidence are to be interpreted as the frequency of behaviors per session. For momentary

time sampling and interval recording procedures, I used $K = 30$ intervals per session; also, I used $l = L/K = 1/30$ as the active interval length for the interval recording procedures, which corresponds to no rest period in between active intervals.

C.1. Basic effect size estimators

In Section 5.3, I proposed several estimators for the log-incidence ratio, log-prevalence ratio, and log-prevalence odds ratio for use under the assumption that the behavior stream is stable within phases. Since all of the estimators are differences between transformations (non-linear functions) of within-phase moments and observations are assumed to be independent across phases, it suffices to evaluate the bias and precision of the transformed within-phase moments, rather than simulating the estimators directly. For instance, rather than simulating continuous recording data with a baseline prevalence ϕ^0 and treatment-phase prevalence $\phi^1 = \omega^\phi \phi^0$ for various values of ϕ^0 and ω^ϕ, I simulate event-counting data from a single phase with prevalence ϕ, for various values of ϕ, and evaluate the bias of the log-mean prevalence: Bias $\left[\ln\left(\hat{y}^C\right)\right] = \mathrm{E}\left(\ln \hat{y}^C\right) - \ln \phi$. For given values of ϕ^0, ϕ^1, the bias of the prevalence ratio estimator F^C is then equal to the difference in the biases of the component estimators for the log-mean prevalence.

Figure C.1 plots the bias of two estimators for the log-incidence based on event-counting data, varying the true incidence between $\zeta = 2$ and $\zeta = 40$ (on the horizontal axis) and the true prevalence between $\phi = 0.05$ and $\phi = 0.50$ (line colors correspond to different values of prevalence). The columns of the lattice correspond to different values of the within-phase sample size n, while the rows of the lattice correspond to different event duration and interim time distributions in the ARP used to simulate the behavior

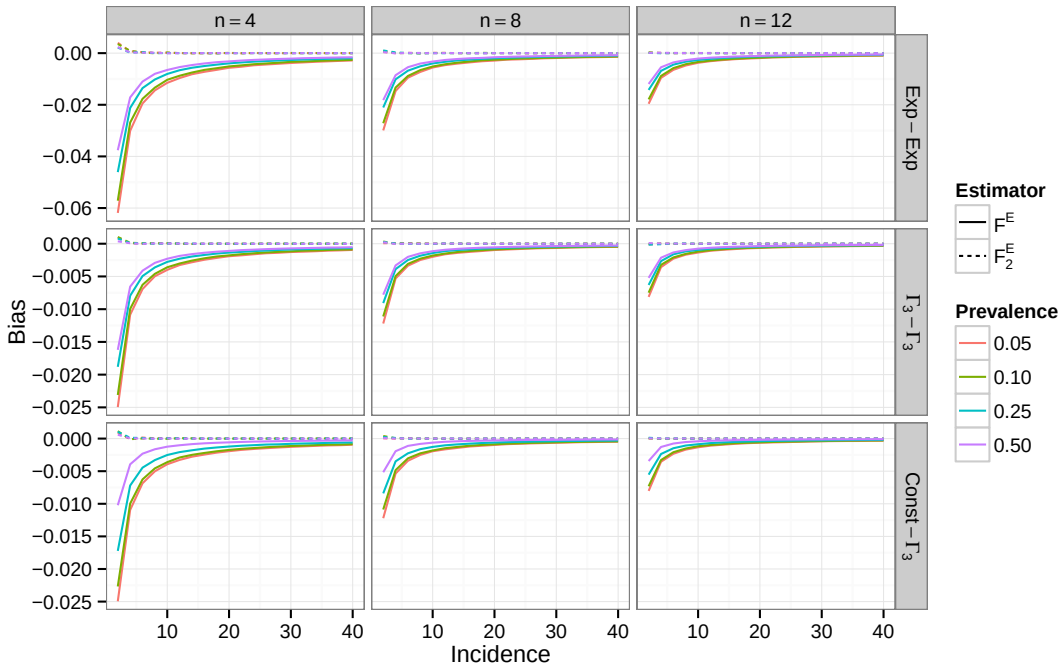


Figure C.1. Bias of log-incidence estimators using event-counting data, as a function of incidence ζ, prevalence ϕ, within-phase sample size n, and distributional assumptions.

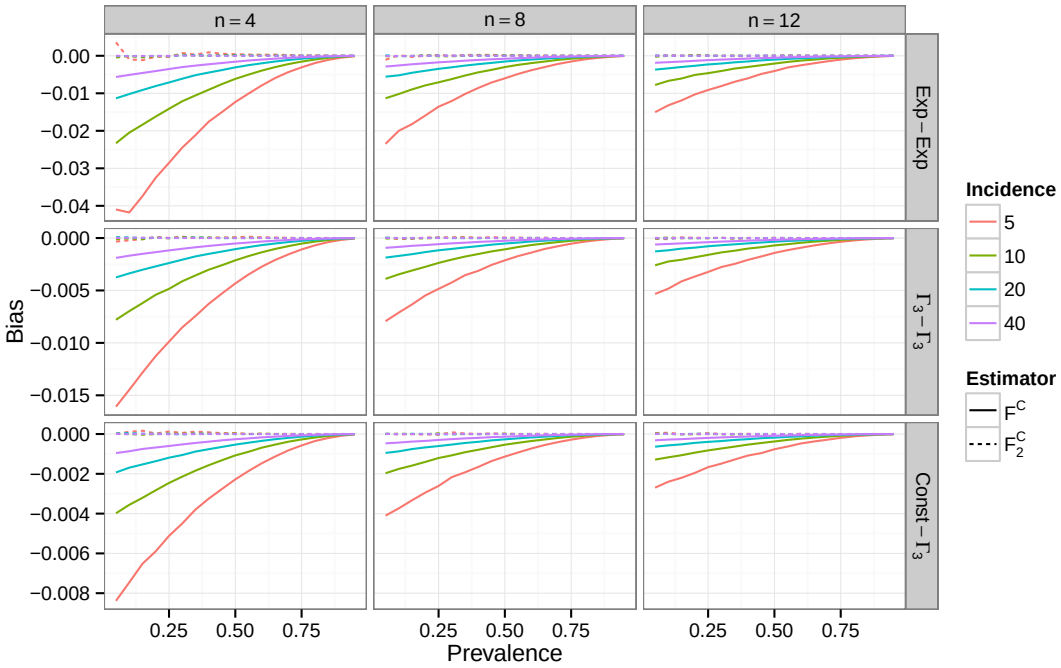
stream.[1] The solid lines correspond to the moment estimator used to construct F^E, while the dashed lines correspond to the bias-corrected estimator used to construct F_2^E. In both cases, I used $c^E = 1/(2n)$ to correct observed zero-mean outcomes. It can be seen that the biases of the moment estimator are quite small except when the true incidence is low and the sample size is small. The bias-corrected estimator is very close to unbiased. The two estimators have practically identical mean-squared error (not shown), leading me to prefer the bias-corrected estimator. The approximate variance estimator V_L^E given in (5.8) is also close to unbiased for the actual variance of the bias-corrected effect size estimator L_2^E.

[1]Note that the vertical scale of the graphs varies across rows of the lattice.

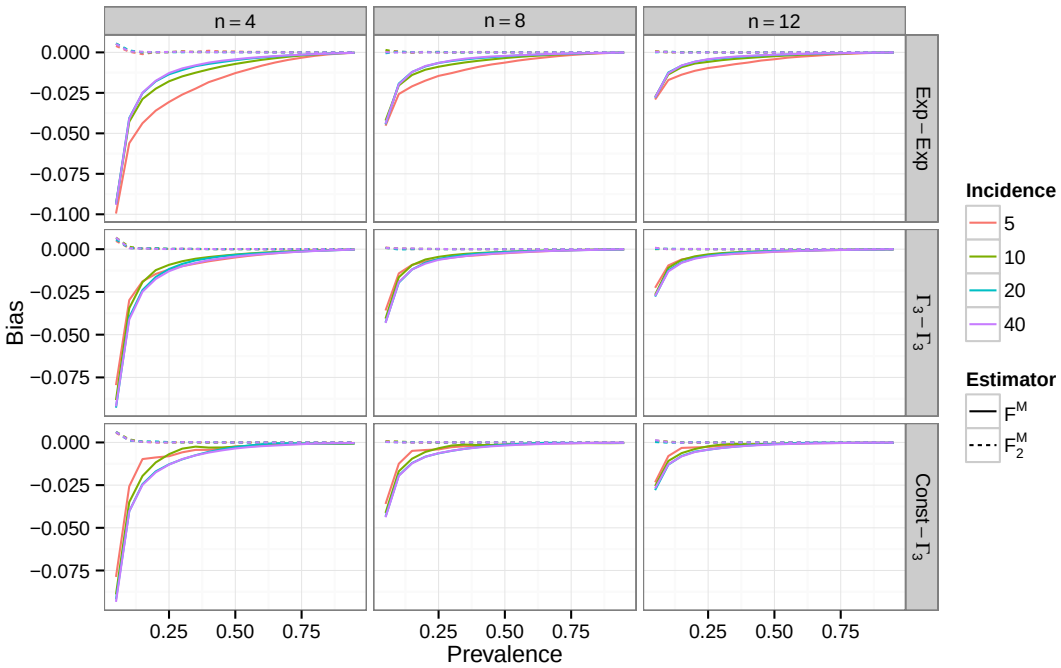
Figures C.2a and C.2b plot the bias of estimators for the log-prevalence based on continuous recording data and momentary time sampling data, respectively. The true prevalence varies between $\phi = 0.05$ and $\phi = 0.95$ (on the horizontal axis), while the true incidence varies between $\zeta = 5$ and $\zeta = 40$ (in these figures, line colors correspond to different values of incidence); within-phase sample size n and distributional assumptions are varied on columns and rows of the lattice, respectively.[2] For continuous recording data, I used $c^C = 1/(2\zeta Ln)$ to correct observed zero-mean outcomes (which were very rare); for momentary time sampling data, I used $c^M = 1/(2Kn)$. The bias of the moment estimator is fairly small in each case, except when the true prevalence is close to zero and the sample size is small. Note that the magnitude of the bias depends on incidence and on the distributional assumptions of the ARP, with larger biases in the Exp-Exp model than in the $\Gamma_3 - \Gamma_3$ or Const$-\Gamma_3$ models; other possible distributional assumptions could lead to larger biases in the moment estimator. For both types of data, the bias-corrected estimator is very close to unbiased. Just as with event counting data, the estimators have comparable mean-squared error, leading me to prefer the bias-corrected estimator.

The approximate variance estimator performs adequately for response ratios based on continuous recording or momentary time sampling data. Averaging across the parameter space, the approximate variance estimator V_L^C is close to unbiased for the actual variance of the bias-corrected effect size estimator L_2^C based on continuous recording data. A similar pattern holds for momentary time sampling data, though the variance estimator V_L^M tends to over-estimate the actual variance of L_2^M by as much as 10% when the true prevalence is very low.

[2]Note that the vertical scale of the graphs varies across rows of the lattice.

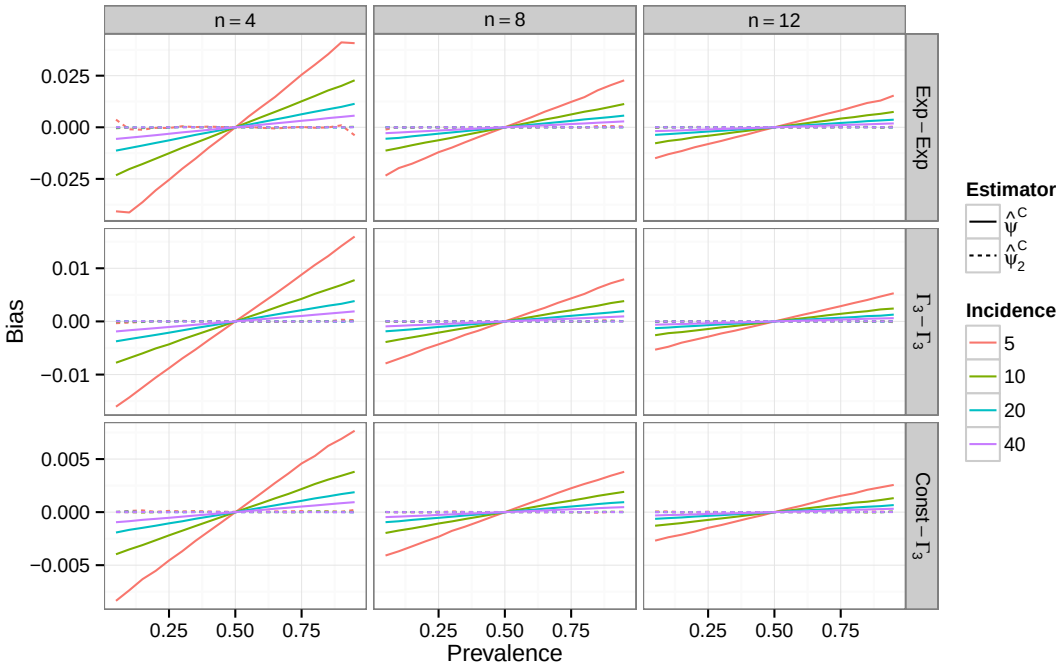


(a) Continuous recording data

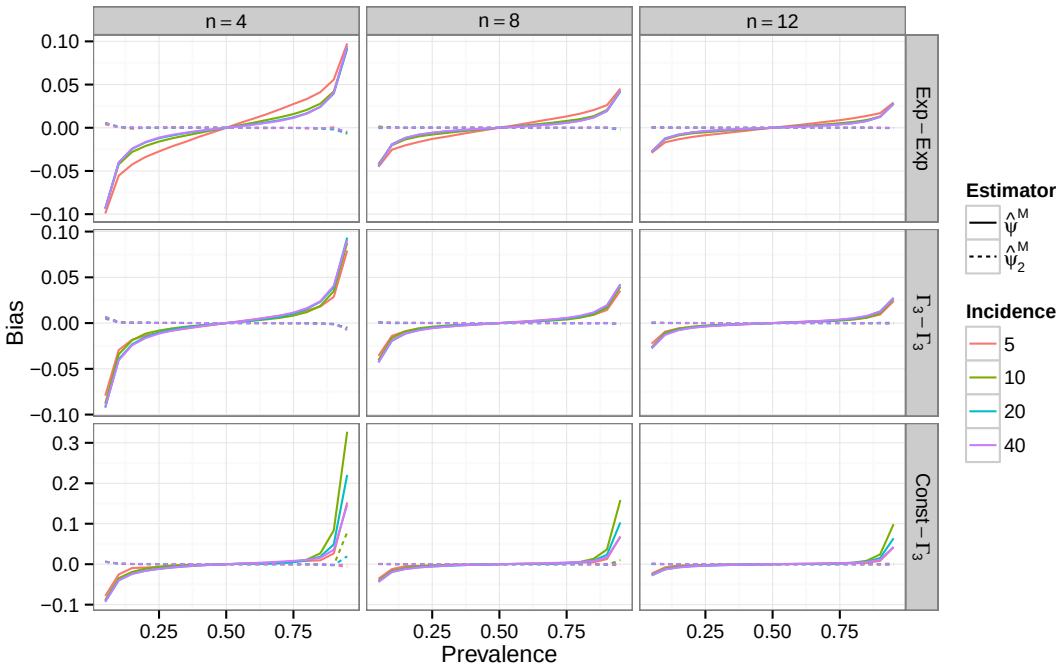


(b) Momentary time sampling

Figure C.2. Bias of log-prevalence estimators using (a) continuous recording data and (b) momentary time sampling data, as a function of prevalence ϕ, incidence ζ, within-phase sample size n, and distributional assumptions.



(a) Continuous recording data



(b) Momentary time sampling

Figure C.3. Bias of log-prevalence odds estimators using (a) continuous recording data and (b) momentary time sampling data, as a function of prevalence ϕ, incidence ζ, within-phase sample size n, and distributional assumptions.

Figures C.3a and C.3b plot the bias of estimators for the log-prevalence odds based on continuous recording data and momentary time sampling data, respectively; these figures are constructed in the same fashion as Figure C.2. The bias of the moment estimator based on continuous recording data is fairly small, except when the true incidence is low and the sample size is small, and is approximately proportional to $\phi - 0.5$. The moment estimator based on momentary time sampling data has comparatively larger biases when the true prevalence is close to zero or one. For both types of data, the bias-corrected estimator is very close to unbiased and has mean-squared error that is slightly smaller than the moment estimator.

C.2. Prevalence trend models

The next set of simulations examines the bias of proposed estimators that account for linear trends in log-prevalence odds, based on data collected using continuous recording or momentary time sampling. Rather than simulating full single-case designs, I generate data from a single phase and examine the bias in the estimated level and trend regression coefficients. I also examine the bias in the predicted value of the log-prevalence odds at a fixed point, extrapolating the estimated trend by one quarter of the length of the data series. Together, these analyses give an indication of the bias in estimates of log-prevalence odds ratios, which are differences between linear combinations of the regression coefficients, estimated based on data from separate phases.

In order to fully describe the data generating model, I need to specify not only a model for prevalence, but also a model for incidence and parametric forms for the event duration and interim time distributions F_D and F_E. I consider a linear model for prevalence. For

a series of length n, define

$$t_j = (2j - n - 1)/(n - 1), \tag{C.1}$$

so that the time trend is centered at the mid-point of the series, and the first and last observations in the series have values $t_1 = -1$ and $t_n = 1$, respectively. Let ϕ^* be a given value of prevalence and let ζ^* be a given value of incidence, both at the mid-point of the series. I use the following model for prevalence:

$$\text{logit}\,(\phi_j) = \beta_0 + \beta_1 t_j, \tag{C.2}$$

$j = 1, ..., n$, where $\beta_0 = \text{logit}(\phi^*)$ and β_1 measures the linear change in the log-odds. A further target parameter is the log-prevalence odds at one quarter of the series length beyond the final observation; denote this estimand $\eta_J = \beta_0 + 1.5 \times \beta_1$.

I examine two models for the incidence. In the first model, indicated by $I = 0$, I assume that the mean event duration is held constant as prevalence changes. In the second model, indicated by $I = 1$, I assume that incidence is held constant as prevalence changes. These models can be summarized as

$$\ln \zeta_j = \ln \zeta^* + (1 - I)\,(\ln \phi_j - \ln \phi^*) \tag{C.3}$$

for $j = 1, ..., n$. Finally, for the event duration and interim time distributions, I examine a model in which both F_D and F_E are exponential (i.e., an alternating Poisson process; this is labeled $G = $ Exp-Exp in the results) and a model in which both F_D and F_E follows gamma distributions with shape parameters equal to 3 (this is labeled $G = \Gamma_3$-Γ_3 in the

Table C.1. Simulation design for prevalence trend model

Parameter	Definition	# Levels	Levels
ϕ^*	Prevalence at $t_j = 0$	9	0.1 (0.1) 0.9
ζ^*	Incidence at $t_j = 0$	3	5, 10, 20
β_1	Change in log-prevalence odds	3	0.0, 0.5, 1.0
I	Indicator of incidence model	2	0, 1
n	Series length	3	4, 8, 12
G	Generating distributions	2	Exp-Exp, Γ_3-Γ_3

results). The reported datum from occasion j is simulated by applying measurement procedure r to a realization of a behavior stream that follows an ARP with specified values of prevalence, incidence, and generating distributions:

$$Y_j^r \sim M_r\left(ARP\left[\phi_j, \zeta_j, F_D, F_E\right]\right), \tag{C.4}$$

where $M_r()$ indicates the application of measurement procedure r to the behavior stream, $r \in \{C, M\}$. Table C.1 summarizes the design of the simulation, which is a $9 \times 3 \times 3 \times 2 \times 3 \times 2$ factorial.

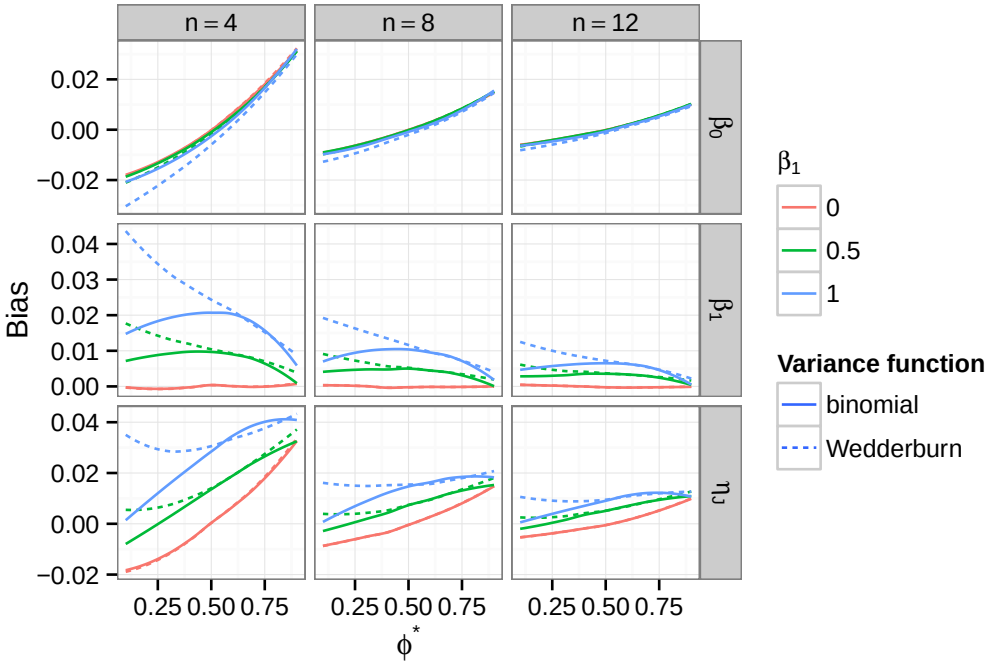
For each combination of factor levels, I simulate 20,000 series and calculate the following statistics. First, I estimate the generalized linear model given in (C.2). For continuous recording data, I compare two different variance functions; one is based on the binomial variance function, $V(x) = x(1-x)$, and the other is based on the Wedderburn variance function, $V(x) = x^2(1-x)^2$. For momentary time sampling data, I consider only the binomial variance function. I calculate the bias and root mean-squared error (rmse) of the estimated regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$, and of the predicted log-prevalence odds η_J. For each estimator, I calculate two different variance estimators, the model-based variance given in (6.8) and the heteroskedasticity-robust estimator given in (6.7).

C.2.1. Bias and efficiency of log-prevalence odds estimators

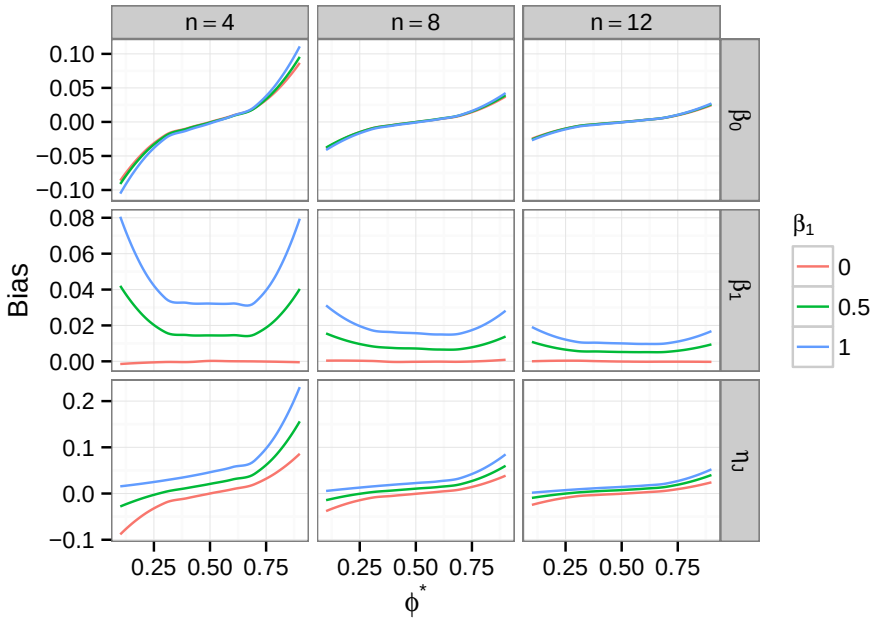
Figures C.4a and C.4b display the average bias of the estimated regression coefficients and predicted log-prevalence odds when based on data from continuous recording and momentary time sampling, respectively; the average bias is calculated across levels of the mid-point incidence ζ^*, the incidence model I, and the generating distributions G.[3] For continuous recording data, the estimator of β_0 has a small-sample bias that is approximately proportional to the true value, independent of β_1 and regardless of which variance function is used; the bias is fairly small, even for the smallest sample size considered. The estimator of the slope β_1 has more pronounced, positive bias at the smallest sample size, also approximately proportional to the true value; here, use of the Wedderburn variance function produces greater bias than the binomial variance. For momentary time sampling data, the estimators of β_0 and β_1 have biases that are approximately proportional to the true values, but generally larger in magnitude than those based on continuous recording data.

For both types of data, the biases are more pronounced when the generating process is more variable. Among the factor levels considered, the biases of $\hat{\beta}_0$ and $\hat{\beta}_1$ are more pronounced when the incidence is $\zeta^* = 5$ and when the generating distribution is $G =$ Exp-Exp. The biases are also more pronounced when the incidence is allowed to vary with prevalence ($I = 0$). In order to provide a sense of the most extreme levels of bias, Figures C.5a and C.5b plot the biases of each estimator, based on continuous recording data and

[3]In each figure, the upper row of the lattice displays the bias of $\hat{\beta}_0$, the middle row displays the bias of $\hat{\beta}_1$, and the lower row displays the bias of $\hat{\eta}_J$. The columns of the lattice correspond to different values of the sample size n; the x-axis of each panel corresponds to ϕ^*, and different colors correspond to different true values of β_1. In Figure C.4a, the solid lines represent the bias of estimators based on the binomial variance function and the dashed lines represent the bias of estimators based on the Wedderburn variance function.

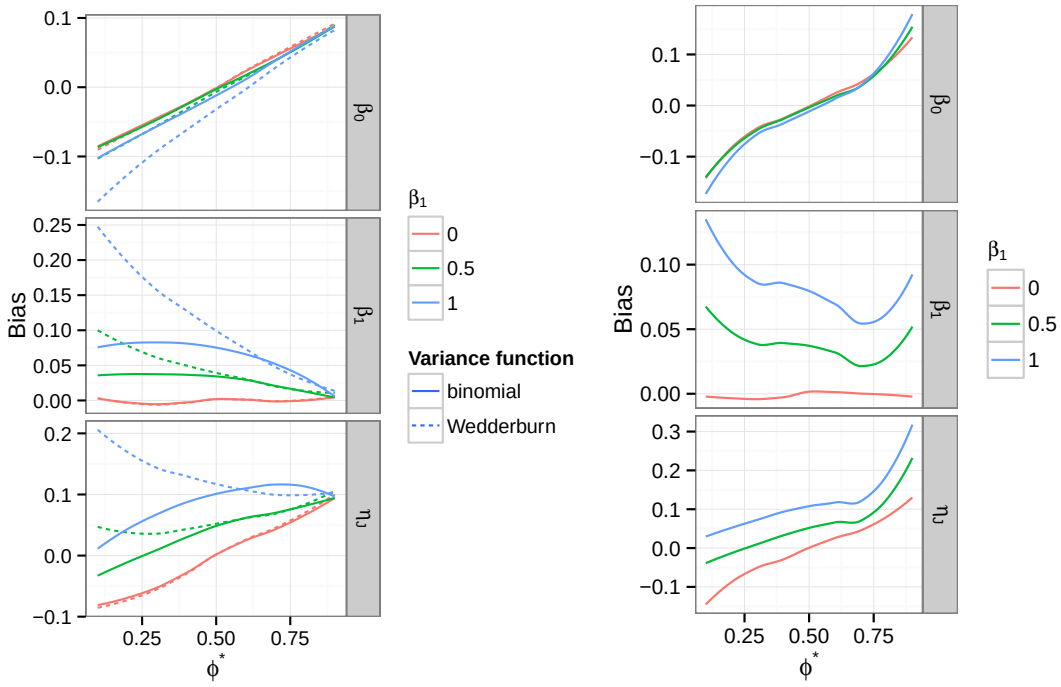


(a) Continuous recording data



(b) Momentary time sampling

Figure C.4. Average bias of level, slope, and log-prevalence odds estimators based on (a) continuous recording data and (b) momentary time sampling data, as a function of mid-point prevalence ϕ^*, slope coefficient β_1, and within-phase sample size n.



(a) Continuous recording (b) Momentary time sampling

Figure C.5. Maximal bias of level, slope, and log-prevalence odds estimators based on (a) continuous recording data and (b) momentary time sampling data, for $n = 4$, $G =$ Exp-Exp, $I = 0$, and $\zeta^* = 5$.

momentary time sampling data, respectively, when $n = 4$, $\zeta^* = 5$, $G =$ Exp-Exp, and $I = 0$. The biases are quite pronounced, particularly those for β_1, where the estimator tends to be upwardly biased by as much as 10%.

For modeling continuous recording data, it would seem that the binomial variance function is preferable to the Wedderburn variance function if bias is the only criteria, though the differences in bias between the two are mostly small. If precision is also taken into account, the differences are smaller still. Figure C.6 plots the average rmse of $\hat{\beta}_0$, $\hat{\beta}_1$, and $\hat{\eta}_J$ versus ϕ^* when the estimators are based on each variance function; the average is taken across levels of β_1, mid-point incidence ζ^*, the incidence model I, and the generating

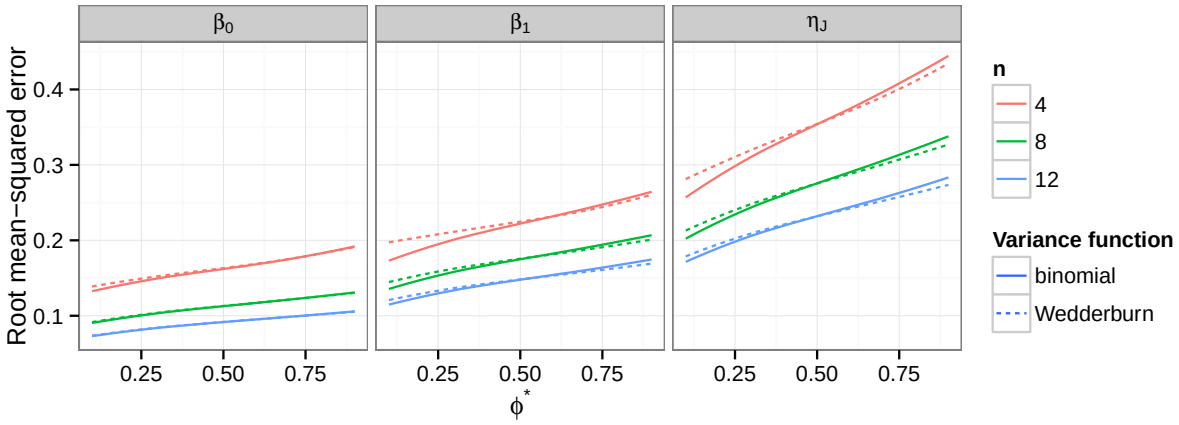


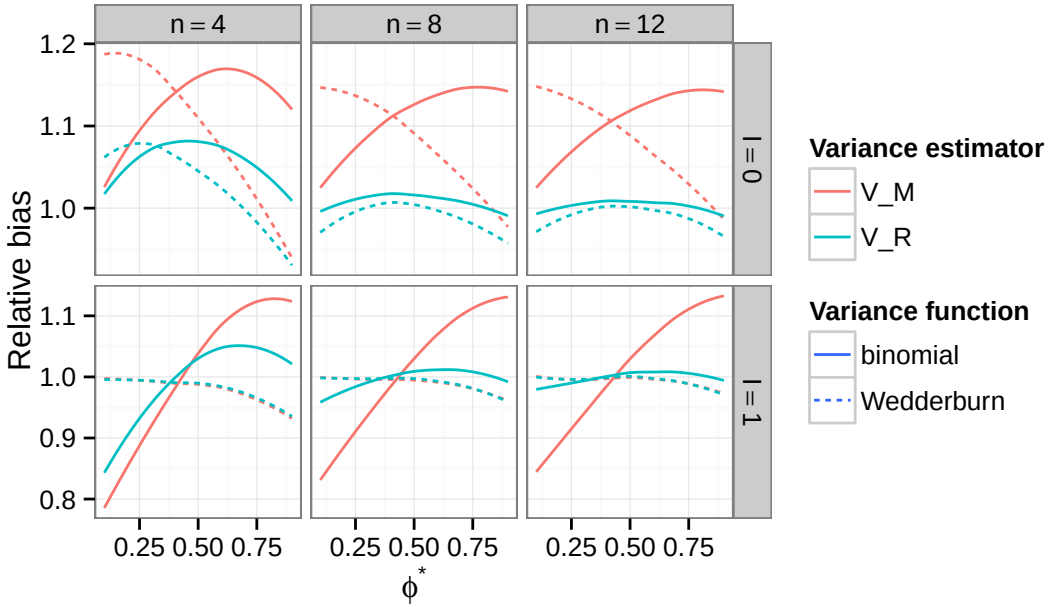
Figure C.6. Average root mean-squared error of level, slope, and log-prevalence odds estimators based on continuous recording data.

distributions G. The average differences in rmse between the two variance functions are negligible.

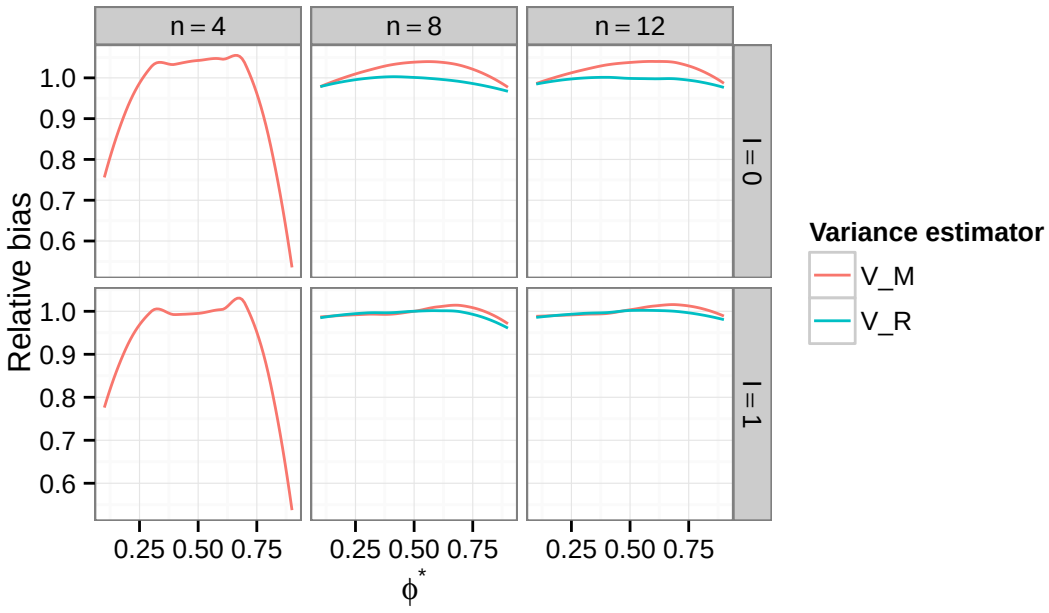
C.2.2. Bias and efficiency of variance estimators

For ease of presentation, I examine the performance of the variance estimators for the log-prevalence odds η_J, rather than for the regression coefficient estimators separately. Figures C.7a and C.7b depict the average relative bias of the model-based and robust variance estimators ($V_M(\hat{\eta}_J)$ and $V_R(\hat{\eta}_J)$) based on continuous recording and momentary time sampling data, averaging over β_1, ζ^*, and G.[4] The average relative bias is plotted versus ϕ^*, with the columns of the lattice corresponding to sample size n and the rows of the lattice corresponding to the incidence model I. For $n = 4$, the robust variance estimator based on momentary time sampling is not depicted because its relative bias is orders of magnitude larger than that of the model-based estimator. At most combinations of factor

[4]For variance estimator V corresponding to the statistic $\hat{\eta}_J$, I calculate the relative bias as $\mathrm{E}\left[V(\hat{\eta}_J)\right]/\mathrm{Var}(\hat{\eta}_J)$.



(a) Continuous recording



(b) Momentary time sampling

Figure C.7. Average relative bias of model-based and robust variance estimators for log-prevalence odds based on (a) continuous recording data and (b) momentary time sampling data.

levels, the robust variance estimator is less biased than the model-based estimator, except for momentary time sampling when $n = 4$. Note in particular that V_R remains close to unbiased when the analytic model for the variance differs to a greater degree from the data-generating model (viz., when $I = 0$).

The results in Figure C.7a also reflect on the choice between variance functions for modeling continuous recording data. Recall that the assumption indicated by I affects the degree to which the assumed variance structure is mis-specified. When $I = 1$ (bottom row), the incidence is approximately constant and so the variance model may be considered approximately correct. Note that in this case, V_M is approximately unbiased when based on the Wedderburn variance function, whereas its bias depends on ϕ^* when based on the binomial variance function; this is true even at the largest sample size considered. This suggests that the Wedderburn variance function may be useful because it provides more accurate variance estimation, even if the resulting effect size estimator has slightly larger variance. Of course, when $I = 0$ and the variance model is more severely mis-specified, both variance functions lead to over-estimation of the true variance.

Figure C.8 compares the relative root mean-squared error of the two variance estimators, for data based on continuous recording (left panel) and momentary time sampling (right panel).[5] The average is taken across the levels of ϕ^*, β_1, ζ^*, I, and G. For $n = 4$, the robust variance estimator based on momentary time sampling is again omitted because its relative bias is orders of magnitude larger than that of the model-based estimator. For both types of data, the robust variance estimator is inefficient relative to the model-based variance estimator. Although the robust variance estimator is asymptotically consistent,

[5]For variance estimator V corresponding to a statistic $\hat{\eta}_J$, I calculate the relative root mean-squared error as $\sqrt{\mathrm{E}^2\left[V(\hat{\eta}_J) - \mathrm{Var}(\hat{\eta}_J)\right]}/\mathrm{Var}(\hat{\eta}_J)$.

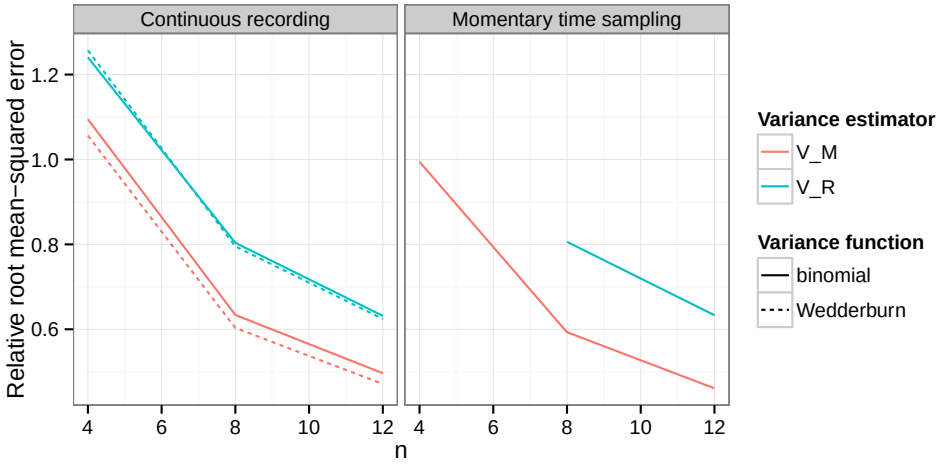


Figure C.8. Average root mean-squared error of model-based and robust variance estimators for log-prevalence odds.

it appears that its greater sampling variability swamps its reduced bias. Conversely, though the model-based variance estimator is not necessarily asymptotically consistent, its bias and sampling variability appear to be small enough that its use is recommended.

C.3. Incidence trend models

The next set of simulations examines the bias of proposed estimators for log-linear trends in incidence, based on data collected using event counting. Just as in previous sections, I generate data from a single phase and examine the bias in the estimated level and trend regression coefficients, as well as the bias in the predicted value of the log-incidence one quarter of the length of the data series beyond the final observation. I assume a linear model for the log-incidence, in which

$$\log\left(\zeta_j\right) = \beta_0 + \beta_1 t_j, \tag{C.5}$$

Table C.2. Simulation design for incidence trend model

Parameter	Definition	# Levels	Levels
ζ^*	Incidence at $t_j = 0$	8	5 (5) 40
ϕ^*	Prevalence at $t_j = 0$	2	0.1, 0.3
β_1	Change in log-incidence	3	0.0, 0.5, 1.0
n	Series length	3	4, 8, 12
G	Generating distributions	3	Exp-Exp, Γ_3-Γ_3, Const-Γ_3

$j = 1, ..., n$, where t_j is defined as in (C.1), $\beta_0 = \log(\zeta^*)$ for ζ^* measured at the mid-point of the series, and β_1 measures the linear change in the log-incidence. I assume that mean event duration is held constant as incidence changes, so that $\phi_j = \phi^*\zeta_j/\zeta^*$, $j = 1, ..., n$, for ϕ^* measured at the mid-point of the series. I examine three alternatives for event duration and interim time distributions: an alternating Poisson process in which both F_D and F_E follow exponential distributions (G = Exp-Exp), a process in which both F_D and F_E follows gamma distributions with shape parameters equal to 3 ($G = \Gamma_3$-Γ_3), and a process in which event duration is constant and F_E follows a gamma distribution with shape parameter 3 (G = Const-Γ_3). The reported datum from occasion j is simulated by applying event counting to a realization of a behavior stream following an ARP process with specified values of prevalence, incidence, and generating distributions:

$$Y_j^E \sim M_E\left(ARP\left[\phi_j, \zeta_j, F_D, F_E\right]\right). \tag{C.6}$$

Table C.2 summarizes the design of the simulation, which is a $8 \times 2 \times 3 \times 3 \times 3$ factorial.

For each combination of factor levels, I simulate 20,000 series. I calculate the bias and rmse of the estimated regression coefficients, $\hat{\beta}_0$ and $\hat{\beta}_1$, as well as an extrapolated log-incidence $\hat{\eta}_J = \hat{\beta}_0 + 1.5\hat{\beta}_1$. For each of these statistics, I calculate two different variance

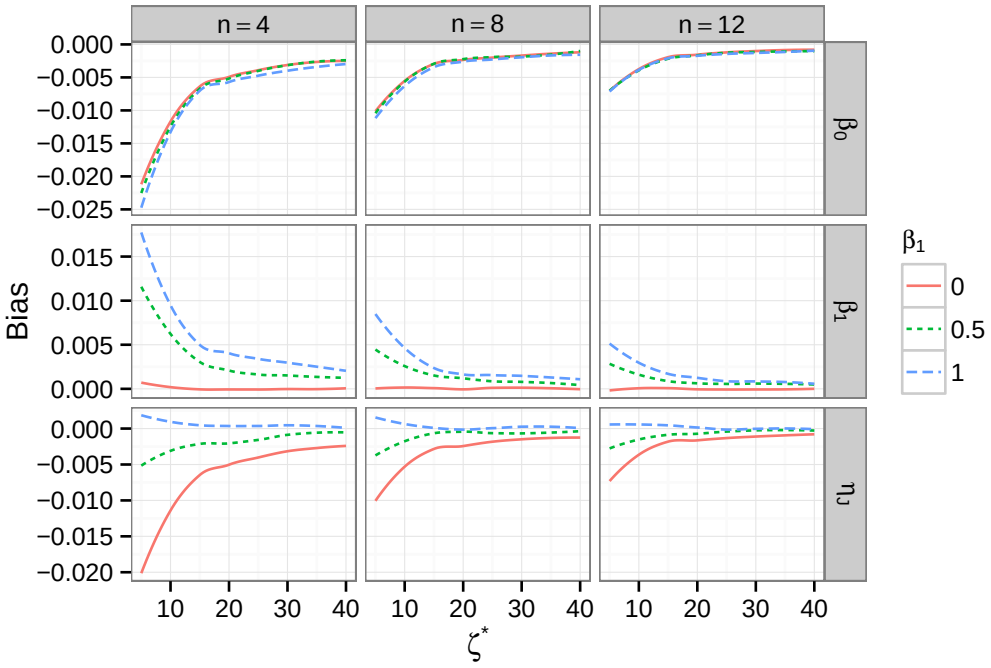


Figure C.9. Average bias of level and slope estimators as a function of mid-point prevalence ζ^*, slope coefficient β_1, and within-phase sample size n.

estimators, the model-based variance given in (6.8) and the heteroskedasticity-robust estimator given in (6.7).

C.3.1. Bias of log-incidence estimators

Figure C.9 displays the average bias of the estimated regression coefficients and predicted log-incidence $\hat{\eta}_J$; the average bias is calculated across levels of the mid-point prevalence ϕ^* and the generating distributions G. The estimator of β_0 has a small, negative bias that decreases as ζ^* increases and is approximately independent of β_1. The estimator of the slope β_1 has a small, positive bias, which also decreases as ζ^* increases. These biases are partially off-setting, yielding an estimator for η_J that has bias of less than 1% when the mid-point prevalence level is greater than 10 incidents per session. Just as with

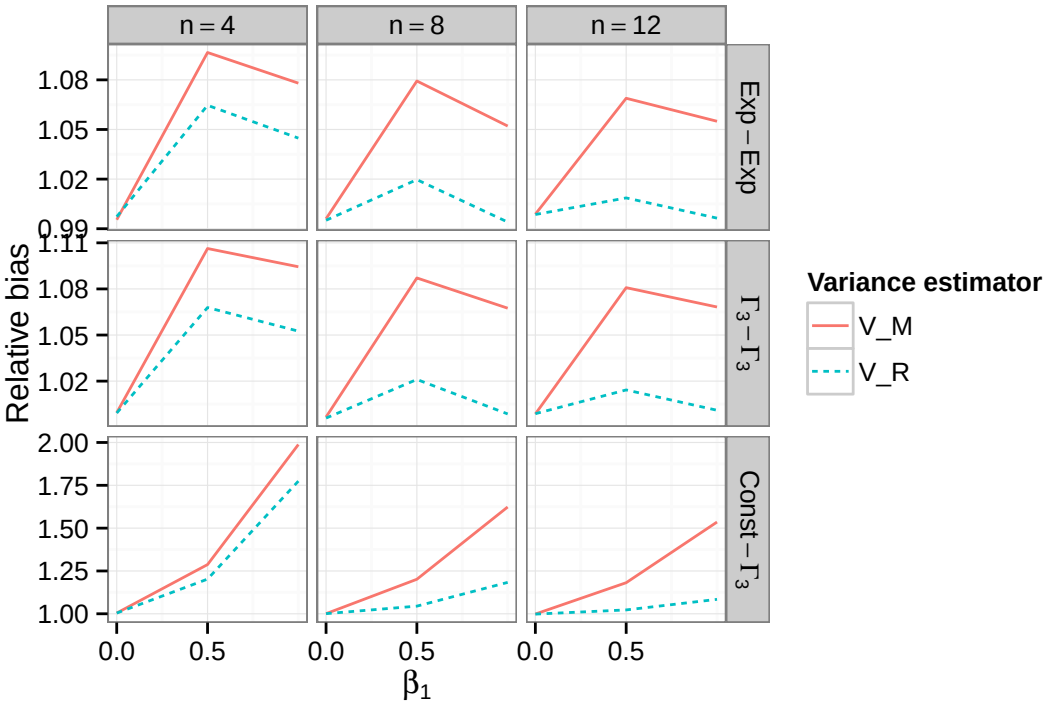


Figure C.10. Average relative bias of model-based and robust variance estimators for log-incidence odds based on event counting data.

continuous recording and momentary time sampling data, the biases are more pronounced when the generating process is more variable. Among the factor levels considered, the biases of $\hat{\beta}_0$ and $\hat{\beta}_1$ are more pronounced at the smaller level of prevalence $\phi^* = 0.1$ and when the generating distribution is $G =$ Exp-Exp.

C.3.2. Bias and efficiency of variance estimators

For ease of presentation, I examine the performance of the variance estimators for the log-incidence η_J, rather than for the regression coefficient estimators separately. Figure C.10 depicts the average relative bias of the model-based and robust variance estimators ($V_M(\hat{\eta}_J)$ and $V_R(\hat{\eta}_J)$), averaging over ζ^*, ϕ^*, and G. The average relative bias is plotted

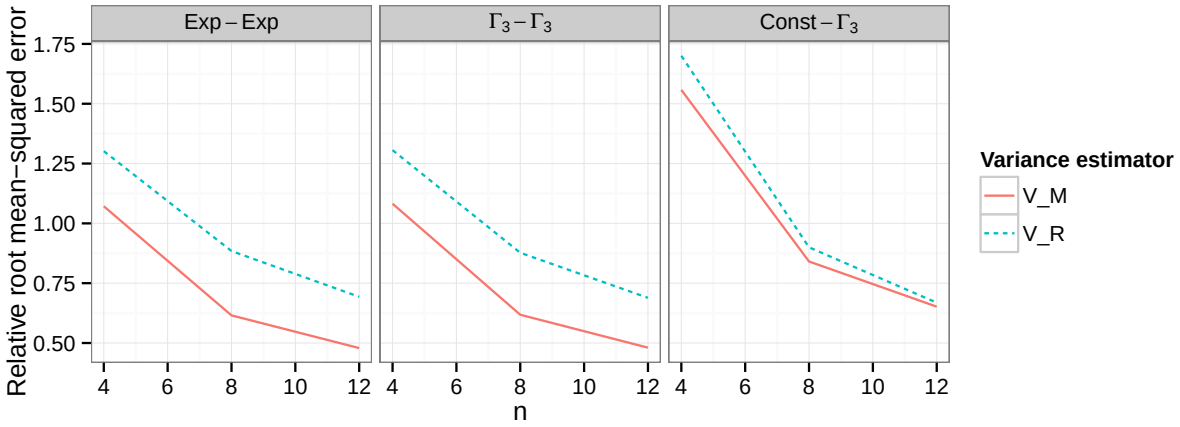


Figure C.11. Average root mean-squared error of model-based and robust variance estimators for log-prevalence odds.

versus β_1, for increasing sample size n. The relative bias of both variance estimators depends on the generating distribution and on β_1. For $G \in \{\text{Exp} - \text{Exp}, \Gamma_3 - \Gamma_3\}$, the model-based estimator has a moderate, positive relative bias while the robust estimator is closer to unbiased. When $G = \text{Const} - \Gamma_3$, the variance model is more severely mis-specified, which creates large biases in the model-based estimator; the robust estimator is also affected, but its bias remains smaller.

Figure C.11 compares the relative root mean-squared error of the two variance estimators. The average is taken across the levels of ζ^*, β_1, and ϕ^*. Despite its smaller biases, the robust variance estimator remains inefficient relative to the model-based variance estimator. Although the robust variance estimator is asymptotically consistent, it appears that its increased sampling variability makes it too large for it to be of use. Conversely, though the model-based variance estimator is not necessarily asymptotically consistent, its bias and sampling variability appear to be small enough that its use is recommended, at least when the variance model is not drastically mis-specified.

C.4. Prevalence dependence models

This section describes a set of simulations involving models in which the latent values of prevalence are serially dependent from session to session. I examine the performance of log-prevalence odds estimators based on continuous recording or momentary time sampling procedures, both of which produce direct measures of prevalence. As in previous simulations, I generate data from a single phase rather than simulating a full single-case design.

I assume that the marginal distribution of the reported data is stable, so that $\beta = \text{logit}\left[\text{E}\left(Y_j^r\right)\right] = \text{logit}\left(\pi^r\right)$ for $r \in \{C, M\}$. I assume that incidence is constant from session to session and that dependence between sessions arises due to serial correlation in values of prevalence. Specifically,

$$\begin{aligned}
\left(Y_j^r | \nu_j\right) &\sim M_r\left(ARP\left[\phi_j, \zeta_j\right]\right) \\
\zeta_j &= \zeta \\
\text{logit}\left(\phi_j\right) &= \beta^* + \nu_j, \qquad \nu_j \sim N\left(0, \sigma^2\right)
\end{aligned}$$

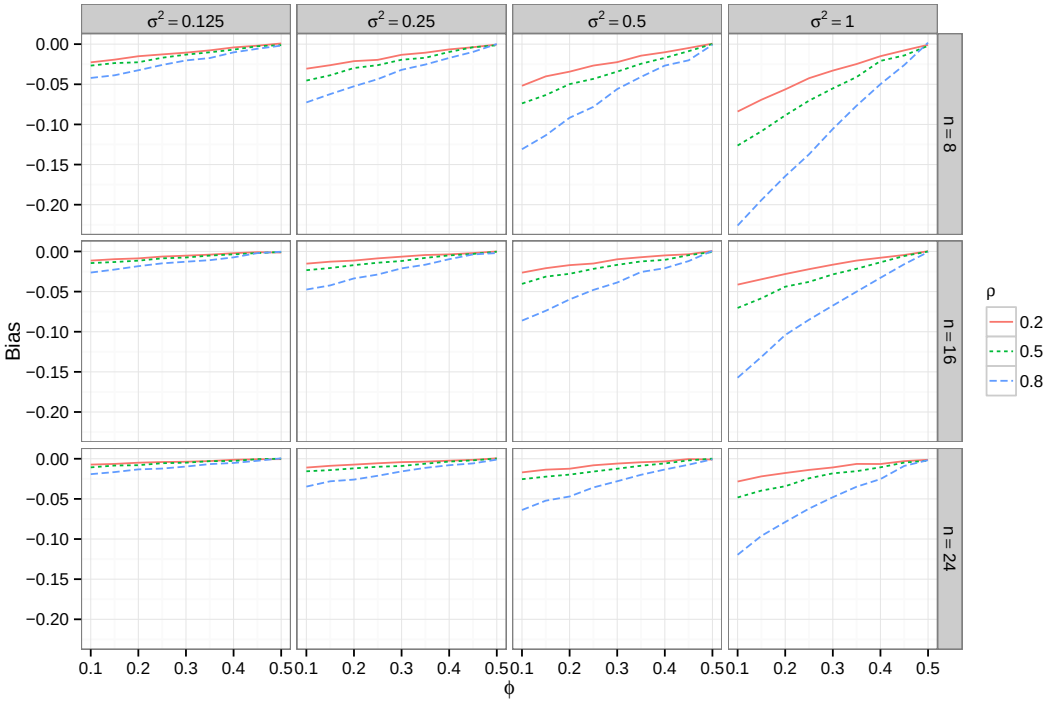
where $\nu_1, ..., \nu_n$ follow an AR(1) serial dependence model with auto-correlation ρ. The conditional mean β^* is defined implicitly by $\pi^r = \text{E}\left[\left(1 + \exp\left(-\beta^* - \nu_j\right)\right)^{-1}\right]$. For the event duration and interim time distributions, I use an alternating Poisson process model in which both F_D and F_E are exponential.

Table C.3 summarizes the design of the simulation. For each combination of factor levels, I simulate 40,000 series and calculate the following statistics. First, I calculate the marginal estimator $\hat{\beta}_I$ and model-based variance estimator V_R, both of which are based

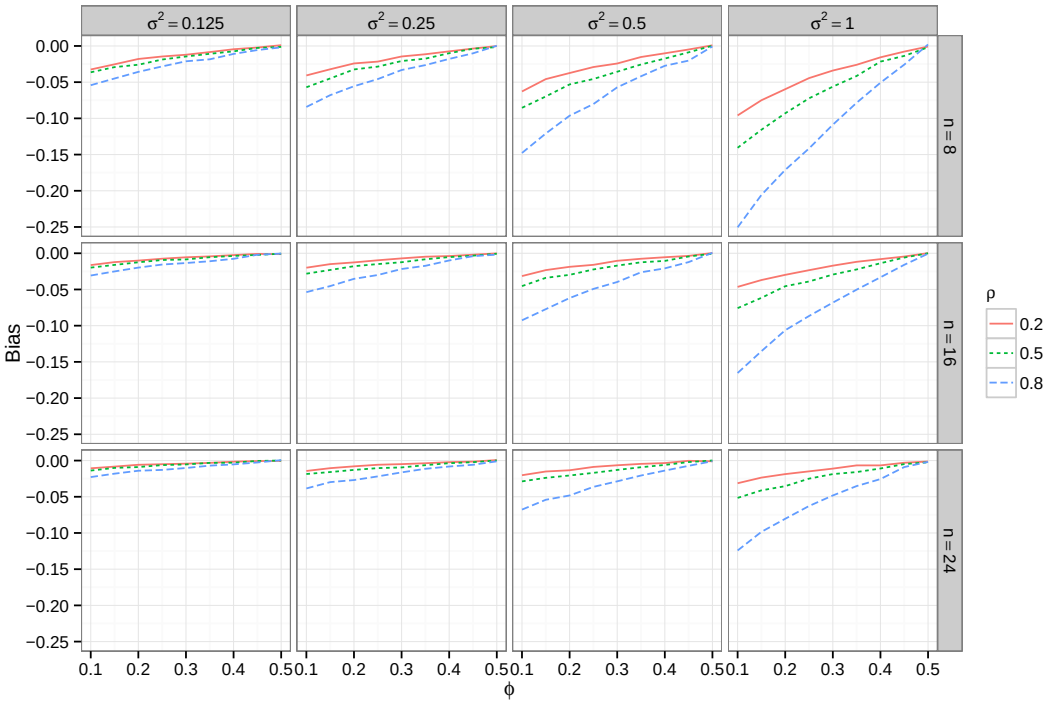
Table C.3. Simulation design for prevalence dependence model

Parameter	Definition	# Levels	Levels
π^C	Expected prevalence	5	0.1 (0.1) 0.5
ζ	Incidence	2	5, 20
σ^2	Latent variability	4	0.125, 0.250, 0.500, 1.000
ρ	Latent auto-correlation	3	0.2, 0.5, 0.8
n	Series length	5	8, 12, 16, 20, 24
-	Order of approximation	2	1, 2
-	Estimating equation	2	FML, RML

on the assumption that repeated measures are independent. In the stable phase model under consideration, $\hat{\beta}_I$ is simply the logit of the mean, and V_R is given in (6.8). Second, I calculate the theoretically optimal linear estimator $\hat{\beta}_{opt} = \text{logit}\left[\mathbf{1}'\boldsymbol{\Sigma}_r^{-1}\mathbf{Y}^r / \left(\mathbf{1}'\boldsymbol{\Sigma}_r^{-1}\mathbf{1}\right)\right]$, with $\boldsymbol{\Sigma}_r$ based on the empirical distribution of the 40,000 simulated series. Third, for a subset of 1,000 simulations, I calculate several different estimators for the variance of $\hat{\beta}_I$ as given in (6.21). For continuous recording or momentary time sampling data, the estimating equation for the nuisance parameters is based on an approximation to the true covariance structure of the data. In the simulations, I compare the performance of the first-order approximation (6.16) to the second-order approximation (6.17). For each approximation, I compute nuisance parameter estimates using both the full Gaussian pseudo-likelihood estimating equation (6.20), denoted as FML, and the restricted version (6.23), denoted as RML. I use the Wedderburn variance function for continuous recording data and the binomial variance function for momentary time sampling data.



(a) Continuous recording



(b) Momentary time sampling

Figure C.12. Average bias of the independence estimator $\hat{\beta}_I$ based on serial dependence model for (a) continuous recording and (b) momentary time sampling data.

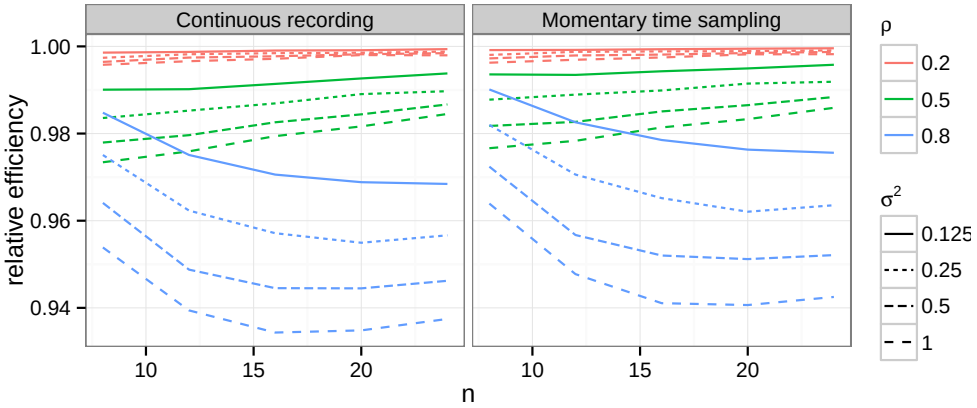


Figure C.13. Average relative efficiency of $\hat{\beta}_I$ versus $\hat{\beta}_{opt}$ based on continuous recording (left panel) or momentary time sampling data (right panel) in the stable-phase model, for varying values of latent variability σ^2 and latent auto-correlation ρ.

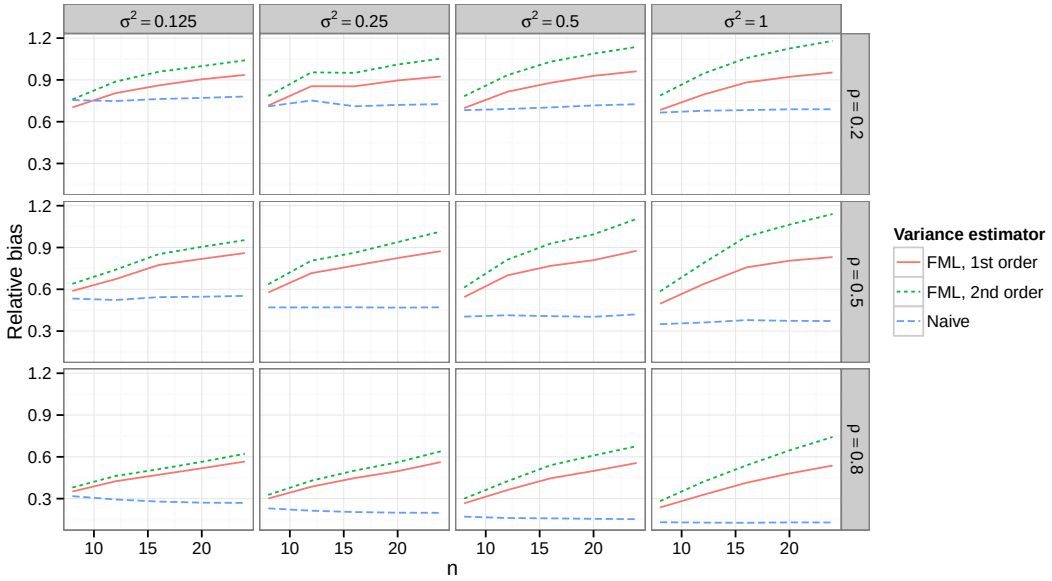
C.4.1. Bias and efficiency of log-prevalence odds estimators

Figure C.16 plots the average bias of the independence estimator $\hat{\beta}_I$ versus the expected prevalence level π^C, for varying levels of latent variability, auto-correlation, and sample size. Results for continuous recording and momentary time sampling are very similar, and so I focus on the former. The bias of the independence estimator depends strongly on the level of latent variability, with higher variability leading to larger bias that is approximately proportional to the logit of π^C. The bias depends to a lesser extent on the level of latent autocorrelation, with higher autocorrelation leading to larger bias. In combination, the effect of a highly variable and highly dependent latent process leads to moderate bias even for the largest sample size considered. The average bias of the optimal linear estimator $\hat{\beta}_{opt}$ is nearly identical to that of $\hat{\beta}_I$, and so it is not displayed.

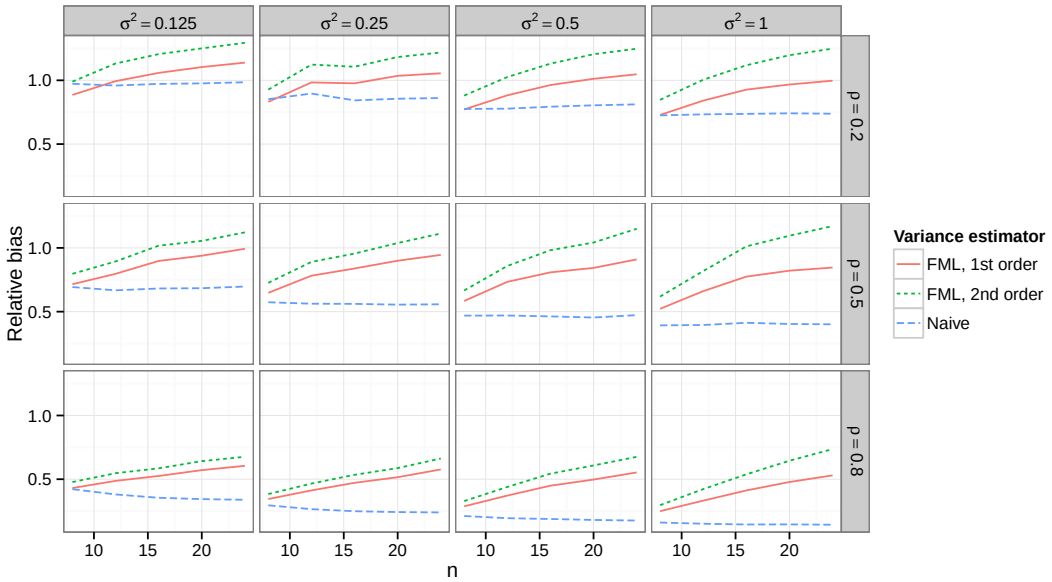
To understand the efficiency of the independence estimator, I compared its root mean-squared error to that of the theoretically optimal linear estimator. I calculate relative efficiency as $\sqrt{\mathrm{E}[(\hat{\beta}_{opt} - \beta)^2]/\mathrm{E}[(\hat{\beta}_I - \beta)^2]}$. Figure C.13 plots the average relative efficiency of the independence estimator versus sample size, for varying levels of latent variability and auto-correlation and for both continuous recording and momentary time sampling data. The degree of latent autocorrelation is the main factor determining relative efficiency. Except for highly autocorrelated processes, the independence estimator is nearly as efficient as the optimal linear estimator, with average relative efficiency of over 97% even for $\rho = 0.5$. The relative efficiency is also affected by latent variability, though this may be partly a consequence of the bias of both estimators at high levels of σ^2. Based on these results, it appears reasonable to use $\hat{\beta}_I$ as a point estimate rather than one where the weight matrix is based on estimated nuisance parameters. I therefore study the performance of variance estimators for $\hat{\beta}_I$ rather than for an estimator involving iterative evaluation of estimating equations, which substantially reduces computational intensity.

C.4.2. Bias and efficiency of variance estimators

Figure C.14 compares the average relative bias of the FML variance estimators based on 1st and 2nd order approximations to the covariance; the naive variance estimator, which does not account for possible serial dependence, is also plotted. The average is taken across levels of π^C and ζ. The relative bias of the RML variance estimators are not displayed because they are severely biased. For both measurement procedures, the naive estimator underestimates the true variance of $\hat{\beta}_I$ to an extent that depends on both σ^2 and ρ; it remains biased even as sample size increases. The 1st-order FML estimator also



(a) Continuous recording



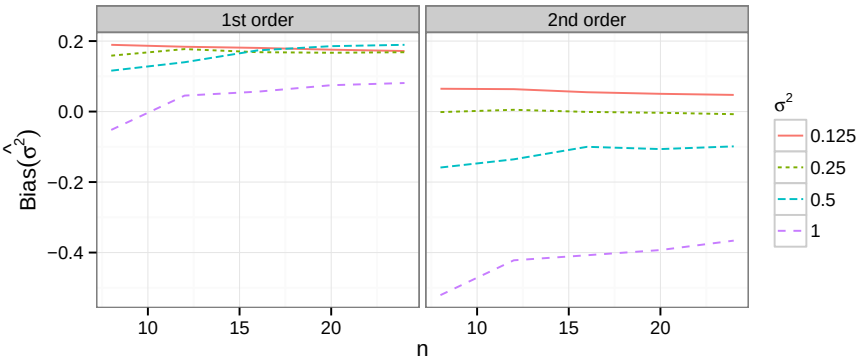
(b) Momentary time sampling

Figure C.14. Average relative bias of the 1st- and 2nd-order FML variance estimators and the naive variance estimator, based on serial dependence model for (a) continuous recording and (b) momentary time sampling data.

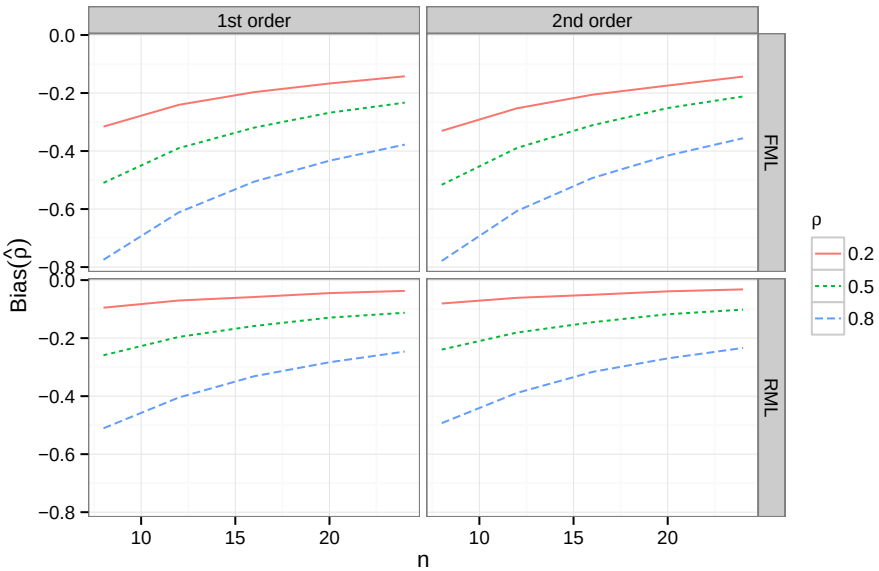
tends to underestimate the true variance to an extent that depends on σ^2 and ρ, but the bias decreases as sample size increases. The 2nd-order FML estimator performs similarly, but appears to overestimate the true variance for larger sample sizes. For series of length $n = 24$, both FML estimators perform reasonably except when the latent autocorrelation is very large. Across the parameter space, the 1st order FML estimator has lower rmse (not depicted), and so may be preferred on that basis. With either estimator, it appears that a fairly large sample size is required in order for the FML estimator to provide a reasonable estimate of variance.

C.4.3. Bias of nuisance parameter estimators

The performance of the FML variance estimators depends on how well estimates of the nuisance parameters are recovered. Figure C.15 displays the average bias of the nuisance parameter estimates of σ^2 and ρ versus the corresponding true parameter values, based on continuous recording data (results for momentary time sampling data are similar). In Figure C.15a, the 1st-order FML estimator of σ^2 has on average an upward bias that appears to be due to the approximation, because it persists as sample size increases. The 2nd-order FML estimator of has a negative bias for larger values of σ^2; the RML estimators of σ^2 have large, positive bias and are not displayed. Figure C.15b displays the bias of the FML and RML estimators of ρ. The 1st- and 2nd-order approximations perform very similarly. Both the FML and RML estimators have downward biases that are reduced as sample size increases. The bias of the RML estimators are smaller and more proportionate to the true parameter. Judging by the biases of the nuisance parameter



(a) FML estimators of latent variability σ^2



(b) 1st order FML and RML estimators of latent autocorrelation ρ

Figure C.15. Average bias of the nuisance parameter estimates, based on serial dependence model for continuous recording data.

estimators even at large sample sizes, more accurate approximations to the covariance matrix could be useful in improving the performance of the FML estimators.

Table C.4. Simulation design for incidence dependence model

Parameter	Definition	# Levels	Levels
π^E	Expected incidence	4	5, 10, 20, 40
ϕ	Prevalence	2	0.1, 0.3
σ^2	Latent variability	4	0.125, 0.250, 0.500, 1.000
ρ	Latent auto-correlation	3	0.2, 0.5, 0.8
n	Series length	4	8, 12, 16, 24
-	Estimating equation	2	FML, RML

C.5. Incidence dependence models

This section describes a set of simulations involving models in which the latent values of incidence are serially dependent from session to session, and where estimators are based on event counting data. The design of this simulation runs parallel to that reported in Section C.4.

I assume that the marginal distribution of the reported data is stable, so that $\beta = \ln\left[\mathrm{E}\left(Y_j^E\right)\right] = \ln\left(\pi^E\right)$. I assume that prevalence is constant from session to session and that dependence between sessions arises due to serial correlation in values of incidence, using the follow data generating model:

$$\begin{aligned}\left(Y_j^E|\nu_j\right) &\sim M_E\left(ARP\left[\phi_j, \zeta_j\right]\right) \\ \ln\left(\zeta_j\right) &= \beta - \frac{1}{2}\sigma^2 + \nu_j, \qquad \nu_j \sim N\left(0, \sigma^2\right) \\ \phi_j &= \phi\end{aligned}$$

where $\nu_1, ..., \nu_n$ follow an AR(1) serial dependence model with auto-correlation ρ. For the event duration and interim time distributions, I use an alternating Poisson process model in which both F_D and F_E are exponential.

Table C.4 summarizes the design of the simulation. For each combination of factor levels, I simulate 40,000 series and calculate the following statistics. First, I calculate the marginal estimator $\hat{\beta}_I$ and model-based variance estimator V_R, both under the assumption that repeated measures are independent; note that under the assumed stable phase model, $\hat{\beta}_I$ is simply the log of the mean. Second, I calculate the theoretically optimal linear estimator $\hat{\beta}_{opt} = \ln\left[\mathbf{1}'\boldsymbol{\Sigma}_E^{-1}\mathbf{Y}^E / \left(\mathbf{1}'\boldsymbol{\Sigma}_E^{-1}\mathbf{1}\right)\right]$, with $\boldsymbol{\Sigma}_E$ calculated using the empirical distribution of the 40,000 simulated series. Third, for a subset of 1,000 simulations, I calculate several different estimators for the variance of $\hat{\beta}_I$. I compute nuisance parameter estimates using both the full Gaussian pseudo-likelihood estimating equation (6.20), denoted as FML, and the restricted version (6.23), denoted as RML, then evaluate $V\left(\hat{\beta}_I\right)$ as given in (6.21).

C.5.1. Bias and efficiency of log-incidence estimators

Figure C.16 plots the average bias of the log-incidence estimators versus sample size n, for varying levels of latent variability and auto-correlation; the average is taken across levels of π^E and ϕ because the bias does not depend strongly on either. The bias of the independence estimator is nearly identical to that of the theoretically optimal linear estimator; only when the latent process is highly autocorrelated does the optimal estimator have reduced bias, and even then the improvement is slight. The bias of both estimators depends strongly on the level of latent variability, with higher variability leading to larger downward bias. The bias depends to a lesser extent on the level of latent autocorrelation, with higher autocorrelation leading to larger downward bias. In combination, the effect of a highly variable and highly dependent latent process produces large downward biases

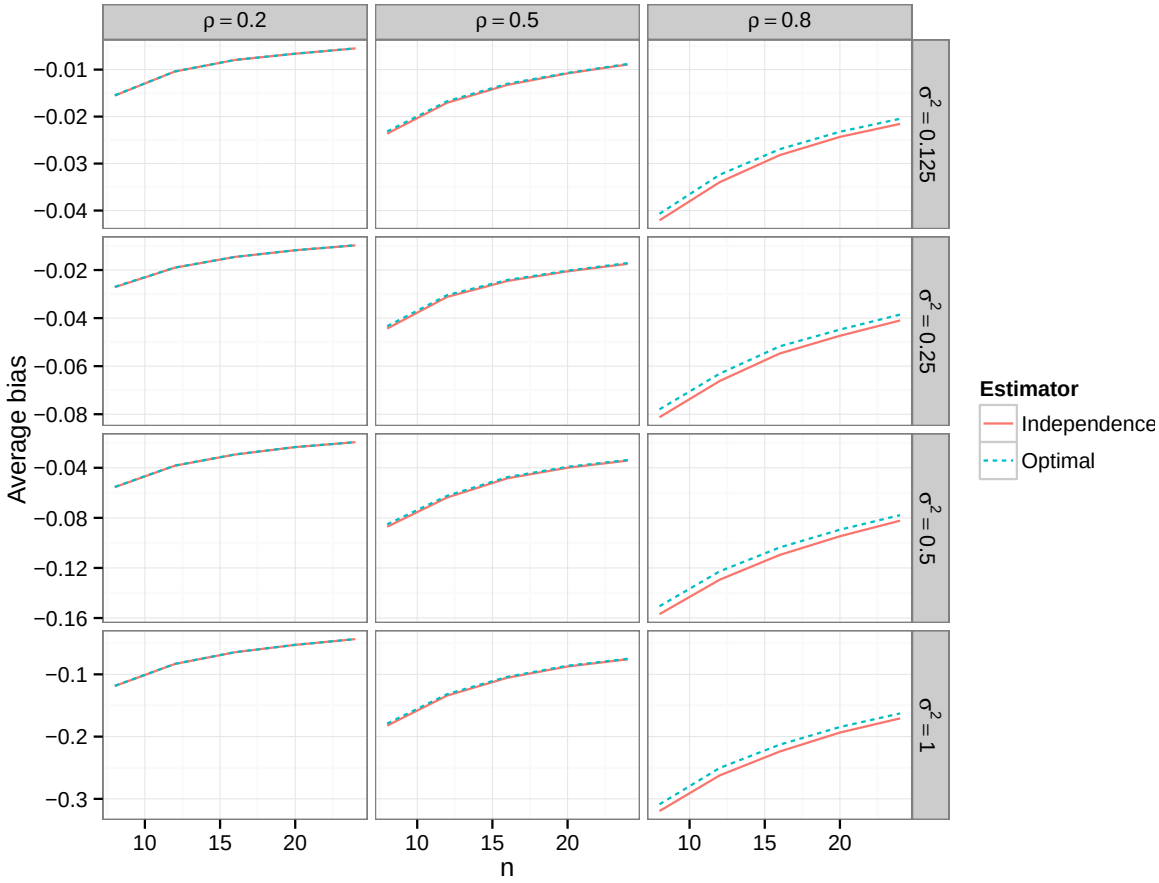


Figure C.16. Average bias of the independence estimator $\hat{\beta}_I$ and the optimal estimator $\hat{\beta}_{opt}$ based on serial dependence model for event-counting data.

in both estimators, even for the largest sample size considered. Considering that these biases are present even in the optimal linear estimator, other classes of estimators or second-order bias corrections will need to be considered if these biases are to be reduced.

To understand the efficiency of the independence estimator, I compared its root mean-squared error to that of the theoretically optimal linear estimator. I calculate relative efficiency as $\sqrt{\mathrm{E}[(\hat{\beta}_{opt}-\beta)^2]/\mathrm{E}[(\hat{\beta}_I-\beta)^2]}$. Figure C.17 plots the average relative efficiency of the independence estimator versus sample size, for varying levels of latent variability and auto-correlation. The degree of latent autocorrelation is the main factor determining

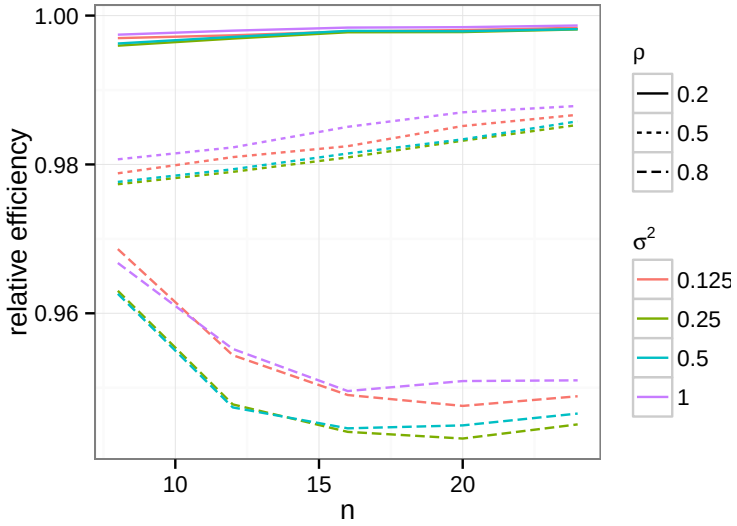


Figure C.17. Average relative efficiency of $\hat{\beta}_I$ versus $\hat{\beta}_{opt}$ based on event counting data in the stable-phase model, for varying values of latent variability σ^2 and latent auto-correlation ρ.

relative efficiency. Except for highly autocorrelated processes, the independence estimator is nearly as efficient as the optimal linear estimator, with average relative efficiency of over 97% even for $\rho = 0.5$. The relative efficiency is not strongly affected by latent variability, though this may be largely a consequence of the bias of both estimators at high levels of σ^2. Based on these results, it appears reasonable to use $\hat{\beta}_I$ as a point estimate rather than one where the weight matrix is based on estimated nuisance parameters. I therefore study the performance of variance estimators for $\hat{\beta}_I$ rather than for an estimator involving iterative evaluation of estimating equations, which substantially reduces computational intensity.

C.5.2. Bias and efficiency of variance estimators

Figure C.18 compares the average relative bias of the FML variance estimator $V\left(\hat{\beta}_I\right)$ to the naive variance estimator, the latter of which does not account for possible serial dependence; the average is taken across levels of π^E and ϕ. The relative bias of the RML variance estimator is not displayed because it is orders of magnitude larger than that of the FML estimator. As is to be expected, the naive estimator underestimates the true variance of $\hat{\beta}_I$ to an extent that depends on both σ^2 and ρ. The naive estimator remains biased even as sample size increases. The FML estimator also tends to under-estimate the true variance to an extent that depends on σ^2 and ρ, but the bias decreases as sample size increases. For series of length $n = 24$, the variance estimator is close to unbiased except when the latent autocorrelation is very large. In general, it appears that a fairly large sample size is required in order for the FML estimator to provide a reasonable estimate of variance.

C.5.3. Bias of nuisance parameter estimators

The performance of the FML variance estimator depends in turn on how well estimates of the nuisance parameters are recovered. Figure C.19 displays the average bias of the nuisance parameter estimates of σ^2 and ρ versus the corresponding true parameter values. It can be seen in Figure C.19a that the FML estimator of σ^2 has a downward bias that is approximately proportional to the true parameter; the proportionate bias decreases with sample size. This is consistent with the behavior of Gaussian maximum likelihood estimators in other contexts. The RML estimator of σ^2 has a large positive bias and is not displayed. Figure C.19b displays the bias of the FML and RML estimators of

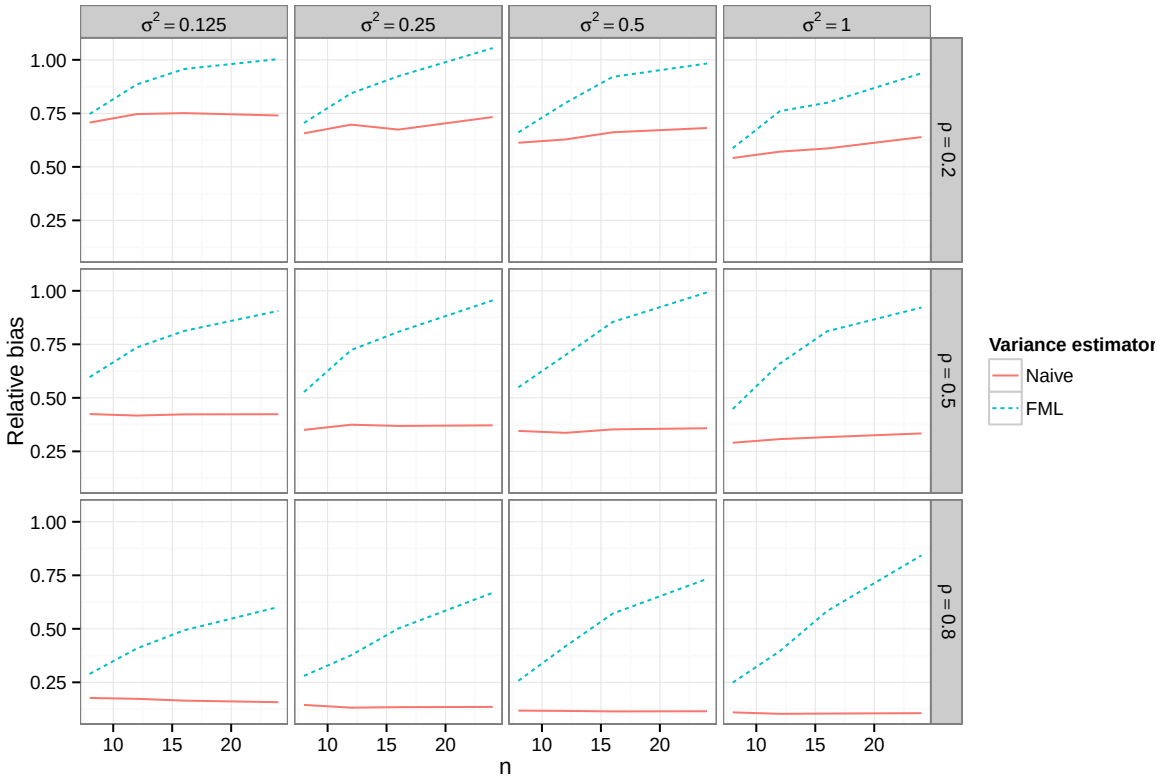
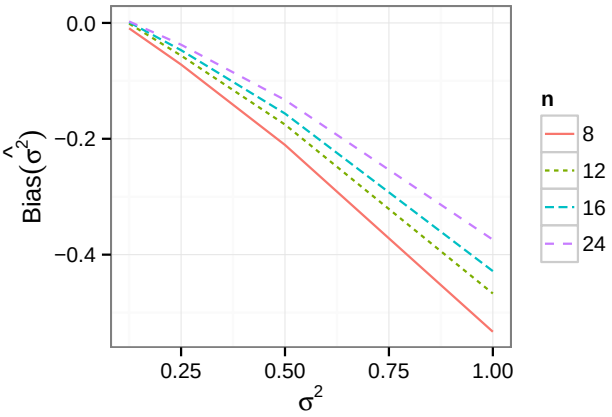
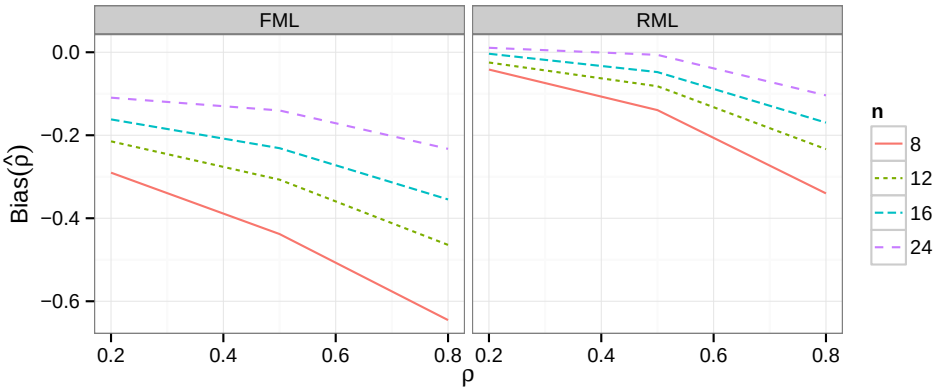


Figure C.18. Average relative bias of the FML variance estimator and the naive variance estimator, based on serial dependence model for event-counting data.

ρ. Both estimators have downward biases that are reduced as sample size increases. Curiously, the bias of the RML estimator is smaller and more proportionate to the true parameter. In combination, the biases of the FML estimators for σ^2 and ρ imply that the off-diagonals of the covariance matrix of $\mathbf{Y}^E$ tend to be under-estimated, leading in turn to underestimation of Var $\left(\hat{\beta}_I\right)$.



(a) FML estimator of latent variability σ^2



(b) FML and RML estimators of latent autocorrelation ρ

Figure C.19. Average bias of the nuisance parameter estimates, based on serial dependence model for event counting data.