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

James E. Pustejovsky

Chapter 1: Operational comparability and single-case research

Single-case designs comprise a set of research methods for evaluating the effects of interventions on individuals. The defining features of the designs are 1) controlled introduction (and possibly also removal) of an intervention on one or more individual cases, 2) repeated measurement of outcomes over time, and 3) use of each case as its own control. Empirical single-case research appears in many areas of psychology and education, but is particularly concentrated in Special Education, School Psychology, Clinical Psychology, Psychotherapy, Social Work, and Applied Behavior Analysis (Horner et al., 2005; Kazdin, 2011; Kennedy, 2004). By nature of the research designs and how study results are usually analyzed, single-case research emphasizes individual change. Single-case designs 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. Thus, while study designs may be highly relevant to individual participants, each study provides meager evidence for drawing generalized inferences.

Despite this idiographic orientation, there has long been interest in using meta-analytic synthesis techniques with single-case research, as a means to bolster the validity of single studies through replication, to study variation in treatment effectiveness across cases, and to generalize from a collection of studies (Allison & Gorman, 1993; Gingerich, 1984; Gorsuch, 1983). More recently, fields that use single-case research have begun to articulate standards of scientific evidence and have looked to quantitative synthesis as a means for establishing evidence-based practices (Chambless & Ollendick, 2001; Horner et al., 2005; Kratochwill & Stoiber, 2002; Odom et al., 2005). As a result, systematic reviews of single-case research now appear with increasing frequency (Maggin, O'Keeffe, & Johnson, 2011) and large research synthesis projects such as the What Works Clearinghouse (WWC) have recently broadened the scope of their evidence standards to include single-case research (Kratochwill et al., 2012).

Despite long-standing interest among single-case researchers and increased attention from the evidence-based practice movement, there remains 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 (Beretvas & Chung, 2008; Wolery, Busick, Reichow, & Barton, 2010). Nearly all are 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. Many of the criticisms of previous approaches to statistical analysis and effect size estimation stem from two features of single-case data: first, that single-case data series often display time trends; and second, that repeated measurements of the same case should be treated as serially dependent, rather than independent. Recent discussions of effect sizes for single-case research have emphasized the importance of accounting for both of these features (Horner, Swaminathan, Sugai, & Smolkowski, 2012; Wolery et al., 2010).

This dissertation addresses two challenges to defining and estimating effect sizes for single-case research. The first challenge is to find effect sizes that remain on a comparable metric across studies that use different research designs, such as single-case designs and simple randomized experiments; I call such effect sizes *design-comparable*. The second problem is to find effect sizes that can be applied across studies that use varied operational procedures for measuring the same construct; I call such effect sizes *measurement-comparable*. Abstractly, both of these problems are special cases of operational comparability—whether effect sizes are on a metric that is invariant across heterogeneous study operations. As I review in Section 1.1, questions of operational comparability arise in many different areas of meta-analysis. 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. Without it, a collection of effect sizes will exhibit heterogeneity due merely to procedural differences in how the study was carried out, making it more difficult to detect any substantive differences.

To address these problems of design-comparability and measurement-comparability, my broad strategy is to formulate structural models that capture essential features of multiple relevant operations (either design-related features or measurement-related features). I then use these structural models to precisely define target effect size parameters, study identification issues, and propose estimation strategies. Chapters 2 through 4 study design-comparability: Chapter 2 describes an abstract set of modeling criteria for constructing design-comparable effect sizes; Chapters 3 applies these general criteria to the family of standardized mean difference effect sizes, proposing a design-comparable effect size and estimation method; and Chapter 4 presents several applications of the proposed models and methods. Turning to measurement-comparability, Chapter 5 proposes a measurement-comparability model and defines effect size measures for use with the most common classes of outcomes in single-case research. Chapter 6 extends the proposed measurement-comparability model to incorporate more complex features, including time trends and serial dependence. Chapter 7 collects various further extensions, areas for further research, and concluding thoughts.

Chapter 2: A general framework for design-comparability models

Design-comparable effect sizes are needed in order to combine evidence from a collection of studies that used heterogeneous designs, such as single-case designs and between-groups designs. Such evidence exists in a number of different research areas, including reading fluency interventions, writing interventions, and phonological awareness training programs. Past syntheses on these topics have either reported separate meta-analyses for each type of design or have limited their scope to only one type of design.

A design-comparable effect size for single-case designs was first proposed by Hedges, Pustejovsky, and Shadish (2012, 2013, henceforth HPS), who used a particular hierarchical linear model to define a standardized mean difference effect size and explicitly demonstrate its equivalence to the standardized mean difference from a between-subjects randomized experiment. However, the HPS method is limited to a single model that makes strong assumptions regarding lack of time trends and homogeneity of treatment effects across cases. Though reliance on these strong assumptions limits the set of studies where the specific effect sizes described by HPS can be applied, the general approach has much broader application.

In this chapter, I explicate the general logic behind the HPS approach and demonstrate how design-comparable effect sizes can be defined under much more general conditions. Specifically, I outline a set of three criteria that a model must meet in order for a design-comparable effect size to be defined; I then describe how to use a model that meets those criteria to construct design-comparable effect sizes. The development in this chapter is abstract, rather than tied to any particular parametric model. Subsequent chapters apply the general logic to specific models and provide detailed applications.

In order for a design-comparable effect size to be defined, it must be sufficiently general that it can describe both a single-case design and a cross-sectional randomized experiment. A sufficiently general model meets the following three criteria. First, it must adequately describe the observed data from the SCD under analysis, including capturing the functional form of the outcome process. This is because treatment effects are identified by extrapolating baseline trends forward in time (Horner et al., 2012), and a model should provide a reasonable fit to the observed baseline data if it is to be the basis for extrapolation. The second criteria is that 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. This criterion ensures that the model is general enough to encompass a randomized experiment. The third criteria is that the model must be causally interpretable at the level of treatment assignment. I examine the third criteria in relation to three of the most common types of SCDs: the multiple baseline design, the treatment reversal design, and the alternating treatment design. For each design, I detail the set of potential outcomes encompassed by a causally interpretable model and examine how the most general models can be constrained by introducing structural assumptions.

A model meeting these three criteria allows one to construct a design-comparable effect size parameter by considering a hypothetical, cross-sectional experiment where 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. The effect size represents a contrast between the two potential outcome distributions identified in such an experiment. Thus, it depends may depend on an implementation time and a target follow-up time, both specified explicitly.

Chapter 3: Design-comparable standardized mean differences: Modeling and estimation

In this chapter, I apply the abstract modeling criteria outlined in Chapter 2 to define design-comparable effect sizes in the family of standardized mean differences. I propose a suite of multi-level models for the multiple baseline design and the treatment reversal design, thereby extending the HPS approach to incorporate time trends, heterogeneous treatment effects, and non-linear treatment response functions. I demonstrate how design-comparable effect sizes can be identified under these models, then describe an estimation method based on restricted maximum likelihood estimation with a further small-sample correction. Finally, I present several simulations examining the operating characteristics of the proposed estimators.

For multiple baseline designs, all of the models that I consider are based on a common specification for the repeated measurements on a given case, involving piece-wise linear time trends in the baseline and treatment phases and auto-regressive dependence in the errors. Given this within-case specification, I describe five models that differ in which of the within-case parameters are allowed to vary across cases. The models are selected to highlight those that would be interesting and useful in application to single-case research. For each model, I demonstrate how to derive a design-comparable standardized mean difference parameter.

Compared to models for multiple baseline designs, causally interpretable models for treatment reversal designs are more difficult to formulate, because they must allow for treatments to be removed and re-introduced. I consider several models with non-linear treatment response functions, in which the treatment effect does not reach full potency immediately and decays only gradually after the treatment is removed; with this within-case model, the equilibrium treatment effect is then either assumed to be constant or to vary across cases. For each of these models, I again demonstrate how to derive a design-comparable standardized mean difference parameter.

Having presented a variety of models for single case designs and demonstrated how to use those models to construct a target effect size parameter, I then propose an estimation method. Based on restricted maximum likelihood (REML) estimates of a model's component parameters, an initial effect size estimate is formed as the ratio of a linear combination of estimated fixed effects to the square root of a linear combination of estimated variance components. For some models, the exact form of these linear combinations depends on the specific times chosen for treatment introduction and follow-up. The initial effect size estimate is then corrected for small-sample bias by approximating its distribution using a *t*-distribution, in a fashion similar to Hedges' *g*-correction (Hedges, 1981); I will refer to the result as the c-REML estimator. The degrees of freedom in the *t*-approximation depend on the covariance matrix of variance component estimates, which I estimate via the inverse of the expected Fisher-information matrix. An estimator for the variance of the effect size is also based on a *t*-approximation.

I conducted several small simulation studies examining the operating characteristics of the c-REML estimator under varying designs and data-generating models. The first simulation used the same basic data-generating models studied in earlier work by HPS, so that the c-REML estimator could be compared directly to the HPS effect size estimator. I find that the c-REML estimator has only small biases, even at the smallest sample sizes considered. Furthermore, it performs comparably to the HPS method in terms of bias and mean-squared error, and so may be considered a viable alternative. The proposed variance estimator also has smaller bias than the corresponding variance estimator proposed by HPS.

Two further simulations examined the performance of the c-REML estimator under models with multiple between-case random effects. With a treatment reversal design and a data-generating model involving heterogeneous treatment effects, I find that the c-REML estimator has reasonably small biases in datasets with five independent cases and has only moderate biases in datasets containing only three independent cases; the corresponding variance estimator is conservative in that it tends to over-state the estimator's true variance. With a multiple baseline design and a data-generating model that allows time trends that vary across cases, the c-REML estimator has reasonably small biases in datasets with five independent cases; however, the corresponding variance estimator tends towards anti-conservatism, likely as a result of the larger number of fixed effect coefficients in the model. Future work should consider penalized likelihood methods (Chung, Rabe-Hesketh, Gelman, Liu, & Dorie, 2013) or small-sample corrections to fixed-effect standard errors (Kenward & Roger, 1997, 2009), which may yield better effect size estimates and variance estimates in models with multiple random effects.

Chapter 4: Design-comparable standardized mean differences: Applications

This chapter presents five detailed applications of the models and estimation methods described in Chapter 3, illustrating the process of model fitting and comparison. The applications are drawn from real single-case studies. Several of the examples were also analyzed by HPS, allowing me to highlight distinctions between the proposed c-RML estimator and the estimation methods proposed by HPS. The comparisons demonstrate some minor deficiencies in the HPS method for estimating nuisance parameters. More broadly, the chapter demonstrates the flexibility and extensibility of the c-REML estimator for a wide array of different models.

Chapter 5: Measurement-comparable effect sizes for free-operant behavior

A desirable characteristic of an effect size measure is that its magnitude should not depend strongly on operational details of how the outcome was measured. Without this property of measurement comparability, it becomes difficult to draw meaningful inferences from averages across and comparisons between effect sizes because true variation in magnitude is confounded by differences in measurement scales. Despite its importance, measurement comparability has received scant attention in discussions of effect sizes for single-case research. The measurement comparability of some commonly used effect sizes has been asserted based on heuristic arguments (Parker, Vannest, & Davis, 2011; Van den Noortgate & Onghena, 2003), but never examined with explicit statistical models.

This chapter develops effect size measures for single-case research that attend closely to the issue of measurement comparability. Setting aside issues of design-comparability, I focus on effect sizes for quantifying changes in the behavior of individual cases, rather than average changes in a population of cases. Also, rather than attempting to encompass any and all measurement operations used in single-case research, I focus only on the most common class of outcome measures: direct observation 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. When conducting direct observation in this context, several different procedures might be used to generate a quantitative summary measurement of the behavior; the most commonly used procedures are continuous recording, event counting, momentary time sampling, and interval recording. In practice, these procedures may be applied to measure very similar constructs, such as problem behavior. Measurement-comparable effect sizes are therefore needed in order to synthesize a set of studies that use such heterogeneous measurement procedures.

To define and study measurement-comparable effect sizes for free-operant behavior, I posit an equilibrium alternating renewal process (ARP) model for the behavior that is observed on a given measurement occasion, or what is sometimes called the “behavior stream.” Two characteristics of the behavior stream correspond directly to parameters of the ARP: prevalence, or the proportion of time that the behavior occurs, and incidence, or the rate at which new behavioral events occur. Given the assumptions of the ARP, several of the recording procedures produce measurements that correspond directly to either prevalence (continuous recording, momentary time sampling) or incidence (event counting). However, interval recording procedures are problematic because they produce measurements that are a complex function of both prevalence and incidence.

After describing the ARP model for the measurement process within a single session, I outline a simple model for the data collected on a single case over the course of multiple observation sessions, both before and after the introduction of a treatment. The model posits that subsequent measurements are independent and that the behavior follows a stable ARP within each phase; only the parameters of the ARP change from phase to phase.

Using both the within-session ARP model and the simple between-session model, I define several effect size parameters for measuring change in distinct aspects of the behavior stream, including the log-incidence ratio, the log-prevalence ratio, and the log-prevalence odds ratio. I then delineate the conditions under which these different metrics are equivalent, so that different effect sizes may be treated as either exactly or approximately measurement-comparable.

Having defined several effect size metrics for measuring change in directly observed behavior, I turn to questions of estimation. For measurements that correspond directly to behavioral parameters, some of the proposed effect sizes can be viewed as special cases of the log-response ratio, a well-known effect sizes used for meta-analysis in ecology and other disciplines (Hedges, Gurevitch, & Curtis, 1999) that is typically estimated by a basic moment estimator. Due to the small sample sizes available in many single-case studies, I instead propose estimators derived from a second-degree Taylor series approximation. Based on simulation studies reported in Appendix C.1, the bias-corrected estimators are preferred because they are nearly unbiased, even in quite small samples, and have mean-squared error that is comparable to the conventional moment estimators.

Next, I propose several approaches for estimating effect sizes from interval recording data, which measure directly neither the prevalence nor the incidence of the behavior. To deal with this construct invalidity, I introduce several distinct sets of further assumptions, each of which lead to bounds for a target effect size that can estimated from interval recording data. These approaches all differ in the assumptions on which they rely and in the information they yield, and thus will be appropriate in quite distinct empirical contexts.

In the final section of the chapter, I demonstrate the use of the proposed effect sizes for meta-analyzing a collection of single-case studies. The studies are drawn from a systematic review examining the effect of choice-making opportunities on the problem behavior of children with disabilities (Shogren, Faggella-Luby, Bae, & Wehmeyer, 2004). For each of the 27 cases in the synthesis, the outcome was a measure of problem behavior; however, a variety of different procedures were used to record the data. Measurement-comparable effect sizes provide a common metric for synthesizing the results across all of the cases, and have the advantage of being interpretable in terms of clear behavioral constructs. The results of the meta-analysis are quite sensitive to the assumptions employed to address the construct invalidity of interval recording data, due to the large number of cases measured using this method.

Chapter 6: Generalized linear models for free-operant behavior

The effect size model in Chapter 5 assumed that 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 prevalence and incidence of the behavior stream. I focus on the types of reported data that are direct measures of incidence or prevalence; applications to interval recording methods remain a topic for future work.

The central challenges in extending the model stem from a lack of tractable probability distributions that are also plausible for measures of directly observed behavior. Under the posited ARP, the first moments of such measurements 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 further moments of the event duration and interim time distributions, about which little will be known. Moreover, even if parametric distributions for event durations and interim times could be specified, the resulting probability distributions for session-level summary data would be intractable.

Absent plausible distributional models for the data-generating processes under study, I turn to quasi-likelihood methods. Quasi-likelihood estimating equations provide a natural and judicious approach to estimation, in that they require assumptions only regarding the mean and variance of the outcome, rather than its full parametric form. However, quasi-likelihood methods also carry the caveat that they provide asymptotic consistency, rather than exact, small-sample results. Given that applied single-case researchers and meta-analysts must deal with limited available data, I rely on small-sample simulation results to make assessments regarding the performance of the estimation techniques that I propose in this chapter.

I first consider generalized linear models that include time trends. For data from a given measurement procedure, I assume that the mean of the outcome relates to a linear predictor via a natural link function. The linear predictor contains a baseline time trend, an initial treatment effect, and a treatment-by-trend interaction term. In this model, a case-specific effect size is defined for a fixed, clinically meaningful duration of treatment, which can be expressed as a linear combination of the initial treatment effect and the treatment-by-trend interaction. To complete the model specification, I assume that the variance of the outcome can be expressed as the product of a set of dispersion parameters and a known variance-mean relationship. The dispersion parameters are allowed to vary between the baseline and treatment phases.

To estimate the target effect size, I use inter-linked quasi-score equations for the parameters in the linear predictor and the dispersion parameters (McCullagh & Nelder, 1989, Chapter 10). Because the quasi-score function for the linear predictor is an unbiased estimating equation, it yields an asymptotically consistent estimator under fairly general regularity conditions, regardless of whether the variance function is correctly specified. Exact variance functions corresponding to the ARP model are not typically available, even given specific parametric forms for the event duration and interim time distributions. Instead, I suggest variance functions for each of the measurement procedures under consideration, which capture the gross features of the mean-variance relationship but may still involve some degree of mis-specification. For example, I recommend the use of the Wedderburn variance function for continuous recording data because it approximates the exact variance when the event duration and interim time distributions are both exponential. I then consider alternative approaches to estimating the variance of the effect size. In a series of small simulations, I find that model-based estimators have substantially lower mean-squared error than empirical sandwich estimators, despite using a variance function that is not exact.

I then describe an extension to this generalized linear model that also incorporates 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 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; a between-session model for change in the behavior stream process—and thus an effect size—could conceivably be specified in terms of either. I follow the latter route, positing a generalized linear model for the marginal mean of the outcome, which implies a unique though possibly non-linear model for the conditional mean (cf. Heagerty & Zeger, 2000). Just as in the simpler models described earlier, case-specific effect sizes can then be expressed as linear combinations of parameters in the mean specification.

To estimate effect sizes under this model, I study two types of linear, unbiased estimating equations. The simpler approach ignores the serial dependence structure in the data, and is equivalent to quasi-likelihood estimation assuming independent measurements. I show that the approach of ignoring serial dependence is often nearly as efficient as an estimator based on a known serial dependence structure. An alternate approach is to estimate the serial dependence structure and using the results to estimate the mean structure more efficiently. In either case, an estimate of the dependence structure is needed in order to estimate the variance of the effect size.

I study Gaussian pseudo-likelihood estimating equations (Hall & Severini, 1998; Wang & Carey, 2004) for estimating the parameters of the dependence structure. I simulate data for each type of direct measurement procedure, based on a very simple model for the mean structure. For event counting data, modeled using a log-link function, I find that fairly long series (24 observations or more) are needed in order for the variance estimator to be approximately unbiased. Even longer series are needed for continuous recording and momentary time sampling data, modeled using a logit-link function, due to the use of approximations rather than exact expressions for the covariance of the data. In summary, the estimation approach that I have examined may require larger sample sizes for adequate performance than are typically available from single-case time series. Future work will explore several approaches to address this shortcoming.

Chapter 7: Future directions

In this chapter, I briefly discuss several further projects that I plan to pursue. 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. First, I sketch a method for estimating standardized mean difference effect sizes in from longitudinal data, including not just single case designs but other types of interrupted time series; the method is designed to be robust to mis-specification of the serial dependence structure for the repeated measurements. Second, I use the ARP model described in Chapter 5 to illustrate construct validity threats that can arise from the use of interval recording methods for direct observation. On a related third topic, I summarize in-progress work aimed at developing new methods of analyzing interval recording data that may remedy the construct validity shortcomings; the method makes use of finer-grained data from within an observation session, rather than just summarized, session-level data. Fourth, having argued for the importance of operational comparability and proposed new effect sizes for single-case research that are demonstrably design-comparable or measurement-comparable, I make some observations about the properties of several prominent effect sizes in the literature.