Journal of School Psychology 68 (2018) 99–112

Contents lists available at ScienceDirect

Journal of School Psychology

journal homepage: www.elsevier.com/locate/jschpsyc

Using response ratios for meta-analyzing single-case designs with behavioral outcomes☆

James E. Pustejovsky*

University of Texas at Austin, Austin, TX, USA

A R T I C L E I N F O

Action Editor: Stephen Kilgus

Keywords:
Single-case research
Meta-analysis
Effect size
Behavioral observation

A B S T R A C T

Methods for meta-analyzing single-case designs (SCDs) are needed to inform evidence-based practice in clinical and school settings and to draw broader and more defensible generalizations in areas where SCDs comprise a large part of the research base. The most widely used outcomes in single-case research are measures of behavior collected using systematic direct observation, which typically take the form of rates or proportions. For studies that use such measures, one simple and intuitive way to quantify effect sizes is in terms of proportionate change from baseline, using an effect size known as the log response ratio. This paper describes methods for estimating log response ratios and combining the estimates using meta-analysis. The methods are based on a simple model for comparing two phases, where the level of the outcome is stable within each phase and the repeated outcome measurements are independent. Although autocorrelation will lead to biased estimates of the sampling variance of the effect size, meta-analysis of response ratios can be conducted with robust variance estimation procedures that remain valid even when sampling variance estimates are biased. The methods are demonstrated using data from a recent meta-analysis on group contingency interventions for student problem behavior.

Studies that use single-case designs (SCDs) comprise a large and important part of the research base in certain areas of psychological and educational research. For instance, SCDs feature prominently in research on interventions for students with emotional or behavioral disorders (e.g., Lane, Kalberg, & Shepcaro, 2009), for children with autism (e.g., Wong et al., 2015), and for individuals with other low-incidence disabilities. SCDs are relatively feasible in these settings because they require fewer participants than between-groups research designs. Furthermore, SCDs involve within-case comparisons—using each case as its own control—and so can be applied even when cases exhibit highly heterogeneous or idiosyncratic problems.

A well-designed SCD makes it possible to draw inferences about the effects of an intervention for the participating individual(s). However, the growing focus on evidence-based practices in psychology and education has led to the need to address further, broader questions—not only about what works for individual research participants, but under what conditions and for what types of individuals an intervention is generally effective (Hitchcock, Kratochwill, & Chezan, 2015; Maggin, 2015). Such questions are difficult to answer based on data from individual SCDs because single studies rarely include broad variation in participant, setting, and intervention procedures, and of course most include only a few participants.

In light of the limitations of individual SCDs, there has long been interest in using meta-analysis methods to draw broader generalizations by synthesizing results across multiple SCDs (Gingerich, 1984; White, Rusch, Kazdin, & Hartmann, 1989). There have

☆ An earlier version of this paper was presented at the annual convention of the American Educational Research Association, April 28, 2017 in San Antonio, Texas. Supplementary materials are available at https://osf.io/c3fe9/.

* Department of Educational Psychology, University of Texas at Austin, 1912 Speedway, Stop D5800; Austin, TX 78712-1289, USA
E-mail address: pusto@austin.utexas.edu.

https://doi.org/10.1016/j.jsp.2018.02.003
Received 26 April 2017; Received in revised form 29 September 2017; Accepted 24 February 2018
0022-4405/

recently been many new developments in the methodology for analyzing and synthesizing data from SCDs (Manolov & Moeyaert, 2017; Shadish, 2014a), as well as increased production of systematic reviews and meta-analyses of SCDs (Maggin, O'Keeffe, & Johnson, 2011). Researchers have also designed frameworks for evaluating study quality, including influential design and evidence standards proposed by the What Works Clearinghouse (Kratochwill et al., 2013), Council for Exceptional Children (Council for Exceptional Children Working Group, 2014), and the Single-Case Reporting Guidelines in Behavioral Interventions (Tate et al., 2016).

A critical methodological decision in any meta-analysis is what effect size measure to use to quantify study results. In the context of SCDs, an effect size is a numerical index that quantifies the direction and magnitude of the functional relationship between an intervention and an outcome. A wide array of effect size indices have been proposed for summarizing SCD results, ranging from simple summary statistics such as the within-case standardized mean difference (Busk & Serlin, 1992; Gingerich, 1984), the percentage of non-overlapping data (PND; Scruggs, Mastropieri, & Casto, 1987), and the non-overlap of all pairs (NAP; Parker & Vannest, 2009), to more complex estimators based on linear regressions or hierarchical linear models (Maggin et al., 2011; Van den Noortgate & Onghena, 2008), as well as between-case standardized mean difference (BC-SMD) estimators that are designed to be comparable to effect sizes from between-groups designs (Shadish, Hedges, & Pustejovsky, 2014). However, there remains a lack of consensus about which effect size indices are most useful for meta-analyzing SCDs (Kratochwill et al., 2013).

To be useful in meta-analysis, an effect size should be in a metric that can be validly compared across studies (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hedges, 2008). In meta-analysis of between-case experimental designs, a key consideration in selecting an effect size metric is how the study outcomes are measured. For example, standardized mean differences are often used to summarize results for outcome constructs assessed using continuous, interval-scale measures such as psychological scales or academic achievement test scores, whereas odds ratios or relative risk ratios are typically used to summarize dichotomous outcomes, such as school dropout or mortality (Borenstein et al., 2009, Chapters 4–5). Some research synthesis projects even use multiple, distinct metrics to quantify effects for different outcome constructs (e.g., Tanner-Smith & Wilson, 2013). In contrast, existing effect size measures for SCDs are typically conceived as generic indices and are often applied with little consideration for how study outcomes are measured.

By analogy to effect sizes for between-case research, it is possible that useful effect size indices for SCDs can be identified by focusing not on single-case research in its entirety, but rather on studies that use a common class of outcome measures. There are at least two reasons for doing so. First, universally applicable effect size metrics are seldom needed because effect sizes are typically combined or compared within a given class of outcomes. Indeed, combining outcome constructs can risk the interpretability of the synthesis results (e.g., how should one interpret an average effect size that combines academic performance and disruptive behavior measures?). Second, all effect sizes are based on modeling assumptions, and outcome measurement properties are an important consideration in developing and validating such assumptions. Just as different modeling assumptions may be required for different classes of outcome measurements, different types of effect size measures may be needed as well.

The most widely used outcomes in single-case research are behavioral measures collected through systematic direct observation (Ayres & Gast, 2010). A variety of scoring procedures are used in conjunction with systematic direct observation, including continuous recording, frequency counting, and interval recording methods. The measurements resulting from these procedures are typically summarized in the form of counts, rates, or percentages. Researchers may also choose to record behavior for longer or shorter observation sessions, which will influence the variability of the resulting scores (i.e., longer observation sessions will produce less variable outcome measurements). Recent evidence indicates that behavioral observation data have features that are not well-described by regression models with normally distributed errors (Solomon, 2014; Solomon, Howard, & Stein, 2015), even though such models have been the predominant approach to statistical analysis of SCD data. As a result, methodologists have begun to emphasize the need for development of statistical analyses and effect size indices that are tailored to and more appropriate for the metrics commonly used with behavioral outcomes (Rindskopf & Ferron, 2014; Shadish, 2014b; Shadish, Hedges, Horner, & Odom, 2015).

One effect size index that may be particularly useful for describing the magnitude of functional relationships on behavioral measures is the log response ratio (LRR). The LRR is a general metric for comparing two mean levels; it is used in many areas of meta-analysis, including economics, medicine, and ecology (e.g., Hedges, Gurevitch, & Curtis, 1999). Pustejovsky (2015) introduced the LRR for meta-analysis of SCDs with behavioral outcome measures. In the context of SCDs, the LRR quantifies functional relationships in terms of the natural logarithm of the proportionate change between phases in the level of the outcome (a formal definition is given in the next section). The LRR is appropriate for outcomes measured on a ratio scale, such as frequency counts or percentage durations of a behavior.

The LRR has several advantageous features as an effect size measure for SCDs, including a direct relationship to percentage change, insensitivity to operational variation in behavioral measurement procedures, and—under certain conditions—comparability across different dimensional constructs. First, the LRR is directly connected to the metric of percentage change, a familiar and readily interpretable conceptualization of effect size that is consistent with how behavioral researchers and clinicians often quantify and discuss treatment impacts (Campbell & Herzinger, 2010; Marquis et al., 2000). Several past meta-analyses of single-case research have used percentage change indices as effect sizes, including syntheses of positive behavioral support interventions (Marquis et al., 2000), behavioral treatments for self-injurious behavior (Kahng, Iwata, & Lewin, 2002), and interventions for reducing problem behavior in individuals with autism (Heyvaert, Saenen, Campbell, Maes, & Onghena, 2014). However, these past applications lacked formal, statistical development for the effect size index—a limitation addressed by the LRR.

A second advantage is that the magnitude of the LRR is relatively insensitive to how the outcome variable was measured, such as use of different recording systems or different observation session lengths (Pustejovsky, 2015, 2018). For instance, a collection of SCDs might include some studies that used continuous recording for twenty-minute sessions and other studies that used 15-sec