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

James E. Pustejovsky

The University of Texas at Austin

Feburary 23, 2018

Forthcoming in *Journal of School Psychology*

Author note

James E. Pustejovsky, Ph.D. University of Texas at Austin, Austin, TX, USA.

A previous 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/.

Correspondence concerning this article should be addressed to James E. Pustejovsky, Department of Educational Psychology; University of Texas at Austin; 1912 Speedway, Stop D5800; Austin, TX 78712-1289. Phone: 512-471-0683. Email: pusto@austin.utexas.edu.

Abstract

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 auto-correlation 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.

Keywords: single-case research; meta-analysis; effect size; behavioral observation

Using Log Response Ratios for Meta-Analyzing Single-Case Designs with Behavioral Outcomes

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 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, Swaminathan 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 momentary time sampling for 10-min sessions. The magnitude of the LRR is unaffected by such procedural variation, making it possible to compare or combine effect sizes from studies that use different measurement procedures. This property is due to the fact that its magnitude depends only on the mean levels of the outcome in each phase. In contrast, other effect size indices such as the within-case standardized mean difference, PND, and NAP are defined in terms of the variability of the outcome measurements, making them sensitive to how the outcomes are measured (Pustejovsky, 2018).

Finally, LRR effect sizes based on different dimensional characteristics of a behavior can sometimes be directly compared (Pustejovsky, 2015). For example, a collection of SCDs might include some studies that use event counting to measure the frequency of a behavior and other studies that use momentary time sampling to measure the percentage duration of a behavior.

Researchers might be interested in comparing an intervention’s effects on behavioral frequency to its effects on percentage duration—or even in combining results across both behavioral dimensions. Pustejovsky (2015) described a theoretical model, called the alternating renewal process, that can be used to identify conditions under which LRR effect sizes for frequency outcomes are equivalent to LRR effect sizes for percentage duration, as well as other equivalence relationships. Although these conditions might not always hold precisely in practice, the framework remains useful as an approximate guide, as illustrated in the meta-analysis example described in a later section.

Along with these advantages, the LRR is also limited in several key respects. First, available methods for estimating the LRR are based on a model that assumes that the outcomes for a given case are stable within each phase of the design (i.e., lacking time trends). Second, methods for estimating the sampling variance of the LRR are based on an assumption that the outcome measurements are mutually independent, which runs counter to the growing consensus that statistical methods for SCDs should provide some means of accounting for serial dependence or auto-correlation (Horner, Swaminathan, Sugai, & Smolkowski, 2012; Wolery, Busick, Reichow, & Barton, 2010). A recent innovation in meta-analytic methodology called robust variance estimation (Hedges, Tipton, & Johnson, 2010) can be used to address this limitation when meta-analyzing LRR estimates, as explained in a later section. A third limitation is that applying the LRR to outcomes measured as proportions or percentages requires attention to how the outcomes are operationally defined, in order to ensure that the resulting effect sizes are on a common scale. Although application of the LRR does involve complexities beyond what is involved in calculating many other available effect size indices for SCDs, this degree of nuance is required precisely because the LRR is suited for quantifying effects on behavioral outcomes, which can be measured in a variety of ways.

Even though it has only recently been proposed for use in the context of single-case research, researchers have begun to apply the LRR in large-scale meta-analyses of SCDs (Common, Lane, Pustejovsky, Johnson, & Johl, 2017; Morano et al., 2017). However, available literature on the LRR is limited to a single, technically-focused article (Pustejovsky, 2015) on how the metric works in the context of a statistical model for systematic direct observation of behavior. There is therefore a need for further guidance about how to apply the LRR effect size in practice. The goal of the present paper is to fill this gap by providing a “user’s guide” for the LRR and demonstrating how this effect size index can be used to meta-analyze SCDs with behavioral outcome measures.

The remainder of the paper is organized as follows. The next section provides a formal definition of the LRR effect size, describes the basic calculations involved in estimating the LRR from data, and demonstrates the calculations with an example. The following section discusses further issues involved in using the LRR to conduct a synthesis of SCDs. The next section turns to methods for meta-analyzing a set of LRR effect size estimates, focusing in particular on methods for robust variance estimation. The meta-analysis methods are demonstrated using data from a recent systematic review of school-based group contingency interventions (Maggin, Pustejovsky, & Johnson, 2017). The final section discusses outstanding limitations and future research directions.

Log response ratios

The LRR effect size is defined based on a simple model for the data from a baseline phase and an intervention phase within a single-case design. Suppose that the baseline phase includes *m* sessions, with outcome data $Y_1^A, \dots Y_m^A$, and that the intervention phase includes *n* sessions, with outcome data $Y_1^B, \dots, Y_n^B$. Let us assume that the average level of the outcome is constant within each phase (i.e., lacking any systematic time trend). Let μ_A denote the mean level of the outcome during the baseline phase and μ_B denote the mean level of the outcome during the treatment phase, where both μ_A and μ_B are greater than zero. Let us further assume that outcome measurements are sampled independently. This is a strong and potentially unrealistic assumption. However, I will describe a method for meta-analyzing LRR effect sizes that remains valid even when the independence assumption is violated.

Under this simple model, the LRR effect size parameter is defined as

$$\psi = ln\left(\frac{\mu_B}{\mu_A}\right), \qquad (1)$$

where ln( ) denotes the natural logarithm function. From the algebraic properties of the natural logarithm, the LRR parameter can be written equivalently as $\psi = \ln(\mu_B) - \ln(\mu_A)$. If there is no change in the underlying level of the outcome—that is, if intervention has no effect whatsoever—then the LRR will be $\psi = 0$. If treatment leads to an increase in the level of the outcome, then the LRR will be positive; conversely, if treatment leads to a decrease in the level of the outcome, then the LRR will be negative.

One basic and advantageous property of the LRR is that it is scale-invariant, meaning that changing the units of the outcome measurements will not change the magnitude of the LRR. For instance, suppose that μ_A and μ_B represent average frequency counts of a behavior observed for 15-min sessions. Re-scaling the outcomes in terms of the rate of behavior per minute does not change the ratio of μ_B to μ_A. The LRR will therefore remain the same whether it is calculated from frequency counts or from standardized rates.

Another useful property of the LRR is that it can be transformed into the metric of percentage change, an intuitive way to interpret the magnitude of effect. The percentage change in the level of the outcome from baseline to the treatment phase is

$$\%\, change = 100\% \times \left(\frac{\mu_B - \mu_A}{\mu_A}\right) . \tag{2}$$

Equivalently, percentage change can be expressed in terms of the LRR parameter as

$$\%\, change = 100\% \times [exp(\psi) - 1], \tag{3}$$

where exp( ) is the exponential function. This relationship can be used to aid interpretation of meta-analysis results for LRR effect sizes.

Estimation

In practice, the true levels of the outcome in each phase are unknown and must be estimated from sample data. The simplest approach to doing so is to replace the unknown mean levels μ_A and μ_B with the corresponding sample mean outcomes from each phase. A complication arises from the possibility that the sample means might be equal to zero even when the true mean levels are positive. To account for this possibility, Pustejovsky (2015) proposed to use the following truncated sample means:

$$\tilde{y}_A = max\left\{\frac{1}{2Dm}, \frac{1}{m}\sum_{i=1}^{m} Y_i^A\right\}, \qquad \tilde{y}_B = max\left\{\frac{1}{2Dn}, \frac{1}{n}\sum_{i=1}^{n} Y_i^B\right\}, \tag{4}$$

where *m* and *n* are the number of observations in the baseline and treatment phases, respectively, and *D* is a constant that depends on the scale of the outcome variable. *D* is equal to 1 for outcomes in the metric of counts; equal to the observation session length (in minutes) for outcomes in the metric of rates per minute; equal to the number of intervals for outcomes measured as a proportion of intervals; and equal to the number of intervals divided by 100 for outcomes measured as a percentage of intervals (continuously duration recording is treated as equivalent to 1-sec interval recording). These values are chosen so that the truncated mean is invariant to changes of scale.

Using the truncated sample means, a basic estimator of the LRR can be calculated as

$$R_1 = ln(\tilde{y}_B) - ln(\tilde{y}_A)\,. \tag{5}$$

However, this basic estimator has a small-sample bias. Pustejovsky (2015) proposed a bias-corrected estimator for use when either phase includes only a small number of observations. The bias-corrected estimator is calculated as

$$R_2 = ln(\tilde{y}_B) + \frac{s_B^2}{2n\tilde{y}_B^2} - ln(\tilde{y}_A) - \frac{s_A^2}{2m\tilde{y}_A^2} \tag{6}$$

where s_A and s_B are the sample standard deviations of the outcome data from the baseline and treatment phases, respectively.

In addition to an estimate of effect size, conventional approaches to meta-analysis also require an estimate of the sampling variance of the effect size. Assuming that the outcomes in each phase are mutually independent, an estimate of the sampling variance of R_2 is given by

$$V^R = \frac{s_A^2}{m\tilde{y}_A^2} + \frac{s_B^2}{n\tilde{y}_B^2}\,. \tag{7}$$

Taking the square root of V^R gives an approximate standard error for R_2: $SE^R = \sqrt{V^R}$. It is important to note that the variance estimator and standard error will not be valid if the outcome measures are auto-correlated. In the presence of positive auto-correlation, they will tend to under-estimate the true sampling variability of the effect size index, and this limitation should be noted when reporting standard errors of LRR estimates for individual data series. However, for purposes of meta-analysis of LRR estimates, robust variance estimation techniques (Hedges et al., 2010) can be used to account for the possibility of inaccurate sampling variances, as detailed in a later section.

Example: McKissick et al. (2010)

I now demonstrate the calculation of LRR effect size estimates and corresponding standard errors using data from a single-case study evaluating the effects of a group contingency intervention on disruptive behavior and engagement in a classroom setting. McKissick and colleagues (2010) used a multiple baseline design across class periods to assess the effects of an interdependent group contingency in a second-grade, general education classroom. Notably, teachers participating in the study expressed a goal of 50% reduction in disruptive behavior. The researchers measured disruptive behaviors using frequency counting and measured student engagement using a partial interval recording procedure; for both outcomes, observations across the entire class of 26 students were taken during 20-min sessions. For purposes of illustration, I focus on the rates of disruptive behavior and calculate LRR effect size estimates for each tier (class period) of the multiple baseline design.

Raw data for analysis were extracted from Figure 2 of McKissick et al. (2010) using the XYit digitization software (Geomatrix, 2007); a graph of the raw data is included in the supplementary materials (https://osf.io/c3fe9/). Table 1 reports sample means, sample standard deviations, and the number of sessions for the baseline and treatment phase during each class period. These summary statistics can be used to calculate the plug-in estimator of the LRR (R_1), the bias-corrected LRR estimator (R_2), and the standard error (SE^R), which are reported in the final three columns of Table 1. As noted previously, the standard errors assume independent outcomes and will tend to be too small if the outcomes are positively auto-correlated. In this example, bias correction reduces the magnitude of the estimates by as much as 0.059 for period 3. The bias-corrected LRR estimates are quite similar across class periods, ranging from -0.610 to -0.807 (equivalent to percentage reductions of between 46% and 55%). The similarity of effects across the three class periods is consistent with the authors' visual analysis of the data. A benefit of using the LRR here is that it quantifies effect magnitude in the same terms as participants' stated goals of 50% reduction in target behavior.

Preparing LRR Estimates for Use in Meta-Analysis

When considering use of LRR effect sizes for synthesizing multiple SCD studies, researchers must address several further issues before carrying out effect size calculations and meta-analysis. This section describes three issues and methods for addressing each, including (1) how to determine whether the LRR is an appropriate effect size metric, (2) how to transform the effect sizes so that their signs (positive or negative) are consistent with the direction of therapeutic improvement for the behavior, and (3) how to calculate effect sizes for studies with more than two phases per case. The final part of this section provides an illustrative example.

Determining whether LRR is Appropriate

Researchers interested in using the LRR for meta-analyzing a collection of SCDs must first determine whether it is an appropriate metric. At least three considerations are relevant here. First, the definition of the LRR parameter requires that outcomes be measured on a ratio scale, such that a score of zero corresponds to absence of the outcome. Thus, researchers must consider whether the outcomes in the collection of SCDs have this property. It may be that some but not all outcome constructs were measured using ratio scales, in which case the researchers could use the LRR for the ratio-scale outcomes and other effect size metrics for other outcome constructs. For constructs measured predominantly on ratio scales, researchers might also need to exclude from effect size calculations a subset of studies that report non-ratio scale outcomes.

Second, the LRR conceptualizes effect size in terms of proportionate change. Researchers should thus consider whether describing intervention effects in terms of proportionate change is meaningful on a practical level. In many instances, proportionate change will likely be a meaningful and intuitive way to describe intervention effects (Campbell & Herzinger, 2010). However, proportionate change will not be meaningful for certain classes of interventions and outcomes. For instance, some meta-analyses of SCDs examine interventions for increasing behavior that is absent or nearly absent during baseline (e.g., antecedent social skills interventions for improving social initiations of students with autism; Ledford, King, Harbin, & Zimmerman, 2016). The LRR may be inappropriate in this context because nearly any increase in behavior would be extremely large in proportionate terms, and small differences in baseline level would lead to drastically different effect size magnitudes. Similarly, the LRR may not be useful for summarizing effects of interventions that consistently produce total extinction of a behavior because such changes are at the extreme of its scale (i.e., reductions of 100% correspond to LRRs of negative infinity). Researchers may need to consider other effect size metrics for such outcomes.

Third, currently available estimation methods for the LRR are based on the assumption that outcomes are stable within each phase (i.e., lacking time trends). Researchers will therefore need to determine whether this assumption is reasonable for the collection of included studies. Existing theory about target behaviors and interventions might indicate whether it is reasonable to expect time trends during baseline and treatment phases. For example, it might be unreasonable to assume stable baselines for academic outcomes such as curriculum-based reading fluency measures, where students improve over time due to instruction, repeated practice, and natural maturation. Similarly, it would not be appropriate to assume stable treatment phases for interventions that aim primarily to affect the rate of change in an outcome, rather than the overall level of the outcome. In addition to theoretical considerations, researchers should also use visual inspection of study outcomes to assess whether systematic time trends are prevalent in the collection of SCDs to be synthesized. Still, considering that many other effect size measures for SCDs are premised on the assumption of stable phases yet are still applied in research syntheses, there are likely to be many instances where the stable phase assumption is reasonable.

Valence Transformation

A collection of studies to be included in a meta-analysis might include some that examine behavior where increase is therapeutically desirable (a positive-valence outcome; e.g., initiations of peer interaction) and others that examine behavior where decrease is therapeutically desirable (a negative-valence outcome; e.g., episodes of physical aggression). If both types of studies are to be included in a meta-analysis, then the effect sizes must first be transformed so that the sign of the estimate is consistent with the direction of therapeutic improvement across all of the outcomes. For some other effect size indices (e.g., standardized mean differences and NAP), this transformation is simply a matter of reversing the sign of the effect size index. However, the appropriate transformation process for LRR effect sizes depends on whether the metric of the outcome variable is a natural rate or a proportion.

For studies that measure behavior on a natural rate or frequency metric, the transformation process is simply a matter of reversing the sign of effect size indices (i.e., multiplying by -1) to be consistent with the direction of therapeutic improvement. For example, suppose that we want positive values of the effect size index to correspond to therapeutic improvement. We would then need to identify all cases that assessed negative-valence behavior and multiply the effect size estimates for these cases by -1. Alternately, if most studies in the meta-analysis examined negative-valence behavior, we might prefer to use negative values of the

LRR to represent therapeutic improvements. In this case, we would identify all cases that assessed positive-valence behavior and multiply the effect size estimates for these cases by -1. The sampling variance and standard error of the effect size are not affected by sign changes and so can be used without further modification.

For studies that measure behavior on a proportion or percentage metric (e.g., percentage of time on task), the transformation process is more involved. Instead of simply changing the sign of the effect size estimate, the outcome variable must be redefined so that the direction of therapeutic improvement is consistent. For example, suppose that most studies examine behavior where decrease is desirable, so that we want negative values of the LRR to represent therapeutic improvement. Now suppose that one of the studies measures percentage of time on-task—a behavior where increase is desirable. Before calculating the LRR, we must first transform the data from this latter study by subtracting the original scores from 100%, yielding percentage of time off-task. The LRR and its variance can then be calculated based on the transformed data. Equivalently, we could calculate the sample mean and variance based on the original data, subtract the mean from 100%, and then calculate the LRR and its variance using the transformed mean. The example at the end of this section demonstrates how to carry out these calculations.

Alternately, suppose that the majority of studies in a synthesis focus on behavior where increase is desirable, so that positive values of the LRR should correspond to therapeutic improvement. If one study measures the proportion of intervals with problem behavior, we must first transform this outcome by subtracting the original scores from 1.00 (or equivalently, subtracting the sample means for each phase from 1.00), yielding proportion of intervals *without* problem behavior. We can then calculate the LRR effect size estimates and variances based on the transformed outcome data.

For count or rate outcomes, transformation changes the sign—but not the magnitude—of the effect size. In contrast, the transformation for proportion or percentage outcomes changes both the sign and the magnitude of the effect size. As a result, there are two distinct ways that the LRR can be applied to proportion outcomes, depending on whether therapeutic improvement corresponds to negative or positive values of the LRR, and researchers will have to choose which approach to take. To distinguish between them, I shall refer to the effect size as LRR-d (for decreasing) when negative values correspond to therapeutic improvement and LRR-i (for increasing) when positive values correspond to therapeutic improvement. If the majority of studies in a set of SCDs focus on negative-valence outcomes, then LRR-d would likely be the preferred approach; similarly, if most studies focus on positive-valence outcomes, then LRR-i would be the natural choice. I revisit the question of choosing between these two approaches in later sections.

Handling Multiple Pairs of Phases

Thus far, I have described methods for estimating the LRR comparing a single baseline phase to a single treatment phase. In practice, some types of SCDs involve several replications of the baseline-treatment contrast for each case (e.g., ABAB designs), and researchers will need to determine how to apply the LRR to represent effect sizes in such cases. I briefly note two approaches that have been used in previous meta-analyses of SCDs and then outline a third, preferred approach.

One approach that has been used in previous reviews is to select a single pair of phases to represent the functional relationship of interest. This might be between the initial baseline phase and the initial treatment phase (Heath, Ganz, Parker, Burke, & Ninci, 2015), or between the initial baseline phase and the final treatment phase (Heyvaert et al., 2014). Although procedurally simple, this approach fails to make full use of the data and ignores contrasts between some phases that are taken into consideration as part of visual analysis.

Another approach is to pool data across multiple phases and calculate a single LRR estimate comparing all baseline phases to all treatment phases (cf. White et al., 1989). For example, in an ABAB design, we would combine the data from the initial baseline (A1) and the return to baseline (A2) phases, and similarly combine the data from the initial treatment (B1) and reintroduction of treatment (B2) phases. This approach assumes that the level of the outcome is constant across all phases under the same treatment condition. It would therefore only be appropriate for studies where the outcome is expected to immediately return to baseline levels upon removal of the intervention.

Although both of the above approaches have been used in practice, neither is consistent with the logic of visual analysis, the predominant method of drawing conclusions from SCDs (Kratochwill, Levin, Horner, & Swoboda, 2014). A third, preferred approach is to calculate LRR estimates for each pair of adjacent phases, then combine those estimates into a single summary effect size for the case. For instance, in an ABAB design we would calculate LRR estimates for the A1-B1 comparison and for the A2-B2 comparison, then average the estimates together. Let R_2^1 and R_2^2 denote the estimates for the first and second pair of phases, with corresponding sampling variances V^{R1} and V^{R2}. The composite effect size estimate is calculated as $R_2 = (R_2^1 + R_2^2)/2$, with sampling variance estimate $V^R = (V^{R1} + V^{R2})/4$. I prefer this approach because it uses all available data and thus produces more precise estimates of effect size than approaches involving only a single pair of phases. Further, it is more consistent with the logic of visual analysis because it relies exclusively on comparisons between adjacent phases and avoids the assumption that the outcome returns to the initial baseline level after treatment is removed. I demonstrate this method in the following example.

Example: Schmidt (2007)

Schmidt (2007) used an ABAB design to evaluate the effects of Class-wide Function-Based Intervention Teams, a group-contingency intervention, on the on-task and disruptive behavior of three focal students in a first-grade class. The behaviors of each focal student were observed daily for 10 min sessions, using duration recording for on-task behavior and frequency counting for disruptive behavior. Note that on-task behavior is a positive-valence outcome on a percentage metric and disruptive behavior is a negative-valence outcome on a natural rate metric. The valence transformation methods described above therefore need to be applied.

Graphs of the raw data for each outcome are included in the supplementary materials (https://osf.io/c3fe9/). Table 2 reports summary statistics for each outcome, student, and phase of the ABAB design. These summary statistics can be used to calculate LRR effect size estimates and accompanying variance estimates for each pair of phases in the design.

In Table 3, Columns (1) through (6) report the LRR-d form of the effect size estimates, which are encoded so that negative values correspond to therapeutic improvements in behavior. Columns 1 and 2 report the effect size estimates comparing the initial baseline (A1) and initial treatment (B1) phases; Columns 3 and 4 report estimates for the return to baseline (A2) and re-introduction of treatment (B2) phases. For disruptive behavior, the LRR-d estimates were calculated directly from the summary statistics. For on-task behavior, the LRR-d estimates were calculated after transforming the outcome to percentage duration of *off-task* behavior (i.e., by subtracting the means of each phase from 100) so that the outcome has negative valence. Column 5 reports the combined LRR-d estimate for each case, calculated by taking the average of the

LRR-d estimates from each pair of phases; Column 6 reports the variance of the combined LRR-d estimate, calculated by taking the sum of the variances from each pair of phases, divided by 4.

The final two columns of Table 3 report the LRR-i form of the effect size estimates, which are encoded so that positive values correspond to therapeutic improvements in behavior. (I report only the combined effect size estimates, which are averages of the estimates for the A1B1 comparison and the A2B2 comparison.) Note that the LRR-i estimates for disruptive behavior were transformed by multiplying the LRR-d estimates by -1, while the LRR-i estimates for on-task behavior were calculated directly from the summary statistics in Table 2.

It is noteworthy that the magnitude of the LRR-d estimates for disruptive behavior are fairly similar to the magnitude of the LRR-d estimates for on-task behavior, whereas the magnitude of the LRR-i estimates is more discrepant across the two outcomes. This pattern suggests that the LRR-d form might be more appropriate for synthesizing results because it is more consistent across outcomes. Of course, this is only a single study, and in practice one will need to examine the consistency of results across the full set of studies to be synthesized.

Meta-Analysis with Robust Variance Estimation

Meta-analysis is a set of statistical techniques for synthesizing results across studies in order to draw generalizations about overall patterns of findings (Borenstein et al., 2009). Meta-analysis can be used to address questions about the overall average magnitude of effects, the degree of consistency or inconsistency (heterogeneity) of results across studies, and characteristics of participants or studies that moderate the magnitude of effect sizes.

In synthesis of between-groups research designs, basic meta-analysis methods involve one effect size estimate per study and effect sizes from different studies are typically assumed to be independent. In contrast, LRR effect sizes from SCDs describe results at the level of the individual case rather than at the level of the study. Studies that include multiple cases thus contribute *multiple* effect size estimates. This leads to a hierarchical structure, in which effect size estimates for individual cases are nested within studies. Working with other effect size metrics, Van den Noortgate and Onghena (2008) proposed a three-level, hierarchical meta-analysis model for synthesizing effect size estimates from SCDs. I follow this approach by applying a hierarchical model for synthesizing LRR effect size estimates, while introducing robust variance estimation techniques that account for the possibility of inaccurate sampling variances.

Suppose that we have identified a collection of K studies for inclusion in the meta-analysis, where study k includes a total of n_k cases. Let R_{jk} denote the LRR effect size estimate and V_{jk}^R denote the corresponding sampling variance, both for case j from study k (for simplicity, I drop the subscript distinguishing R_1 from R_2). The multi-level meta-analysis model describes the LRR estimate for a given case in terms of an overall average effect size γ, a study-level error term v_k, a case-level error term u_{jk}, and a sampling error e_{jk}:

$$R_{jk} = \gamma + v_k + u_{jk} + e_{jk}. \tag{8}$$

The sampling error e_{jk} corresponds to the difference between the effect size estimate R_{jk} and the *true* effect size parameter for that case; it is assumed to have mean zero and known variance V_{jk}^R. Note that this assumption will be violated if V_{jk}^R is not an accurate estimate of the true sampling variance, as would be the case if the outcome data were auto-correlated. This potential problem is the main reason to focus on robust variance estimation techniques, which are valid even if the sampling variances of the effect size estimates are inaccurate.

The case-level error term u_{jk} corresponds to the difference between the true effect for case j and the average true effect for all cases in study k. It assumed to be normally distributed with mean zero and unknown variance ω^2. The variance parameter ω^2 describes the degree of heterogeneity in the effects across the population of cases within a given study. Larger values of ω^2 indicate that the effects of intervention are less consistent (more variable) across cases within a study. Finally, the study-level error term v_k corresponds to the difference between the average true effect for study k and the overall average effect; it is the assumed to be normally distributed with mean zero and variance τ^2. The variance parameter τ^2 describes the degree of heterogeneity in the effects across studies; larger values of τ^2 indicate that intervention effects are less consistent across studies. Such heterogeneity might arise from variation in intervention procedures, implementation fidelity, or study populations.

Estimation

The main parameters of interest in the multi-level meta-analysis model are the overall average effect size (γ) and the variance components ω^2 and τ^2, which quantify the degree to which effect sizes are heterogeneous across cases and across studies, respectively. Estimates of the variance components can be obtained through maximum likelihood or restricted maximum likelihood methods, which are iterative numerical estimation procedures. These methods are implemented in many widely used statistical analysis software packages, including SAS PROC MIXED (Littell, Milliken, Stroup, Wolfinger, & Schabenberber, 2006), the metafor package in R (Viechtbauer, 2010), and the gllamm command for Stata (Rabe-Hesketh, Skrondal, & Pickles, 2004). It is important to note that the estimates of the variance components are obtained based on the assumption that the sampling variances are accurate. Violation of this assumption might affect the accuracy of the variance component estimates. Still, even if the sampling variances are not fully accurate, I would nonetheless recommend estimating and reporting the case-level and study-level variance components because they remain informative as rough indicators of heterogeneity.

Assume that estimates of both variance components, denoted $\widehat{\omega}^2$ and $\hat{\tau}^2$, have been obtained using software. An estimate of the overall average effect size γ can then be constructed as a weighted average of the LRR estimates, with weights chosen so that the overall average effect estimate is as precise as possible. It is helpful to break this calculation into two steps. In the first step, study-specific average effect sizes are estimated using a weighted average of the case-level LRR estimates:

$$\bar{R}_k = \frac{1}{G_k}\sum_{j=1}^{n_k}\frac{R_{jk}}{\widehat{\omega}^2 + V_{jk}^R}, \quad \text{where} \quad G_k = \sum_{j=1}^{n_k}\frac{1}{\widehat{\omega}^2 + V_{jk}^R} \tag{9}$$

for $k = 1,\ldots,K$. The overall average effect size estimate, denoted $\hat{\gamma}$ is then calculated as a weighted average of the study-level averages:

$$\hat{\gamma} = \frac{1}{H}\sum_{k=1}^{K} h_k \bar{R}_k \quad \text{where} \quad h_k = \frac{1}{\hat{\tau}^2 + \frac{1}{G_k}} \quad \text{and} \quad H = \sum_{k=1}^{K} h_k. \tag{10}$$

When the sampling variances and estimated variance components are accurate, this weighted average is optimal because it is as precise as possible. Even if the sampling variances and estimated variance components are inaccurate, though, the weighted average remains a valid estimate of the overall average effect size.

Robust Variance Estimation

In conventional meta-analysis, the standard error of the overall average effect size estimate would be estimated as $1/\sqrt{H}$. Unlike the point estimate of the average effect size, the validity of its standard error is contingent on the accuracy of the LRR sampling variances and the estimated variance components, which might not be valid if the outcome data are auto-correlated. Robust variance estimation techniques (Hedges et al., 2010) can be used to calculate a standard error for $\hat{\gamma}$ that does not rely on the accuracy of $\widehat{\omega}^2$, $\hat{\tau}^2$, or the V_{jk}^Rs. Basic robust variance techniques require a large number of studies to achieve unbiased variance estimates and confidence intervals with correct coverage levels (Hedges et al., 2010). However, finite-sample corrections for robust variance estimation are available (Tipton, 2014) and provide more accurate variance estimates and coverage levels when the number of studies is small or moderate.

A robust variance estimator is given by the weighted sample variance of the study-level average effect size estimates:

$$V^{\gamma} = \frac{1}{H^2} \sum_{k=1}^{K} \frac{h_k^2 (\bar{R}_k - \hat{\gamma})^2}{1 - \frac{h_k}{H}} \,. \tag{11}$$

The robust standard error for $\hat{\gamma}$ is calculated as the square root of the variance estimator: $SE^{\gamma} = \sqrt{V^{\gamma}}$. This standard error is robust in the sense that it does not rely on the accuracy of the sampling variances or variance component estimates. Roughly speaking, it works by treating the study-level average effect estimates $\bar{R}_1, \dots, \bar{R}_K$ as a random sample from a distribution with unknown variance and estimating their variance empirically (using the sample variance of the study-level average effect estimates) rather than by relying on modeling assumptions.

A $(1 - \alpha) \times 100\%$ confidence interval for the overall average effect size can be calculated as follows. Let $t(\alpha/2\,, f)$ denote the $\alpha/2$ critical value from a *t* distribution with f degrees of freedom. The end-points of the confidence interval are then calculated as

$$\gamma_L = \hat{\gamma} - SE^{\gamma} \times t\left(\frac{\alpha}{2}, f\right), \quad \gamma_U = \hat{\gamma} + SE^{\gamma} \times t\left(\frac{\alpha}{2}, f\right), \tag{12}$$

where f is a small sample degrees-of-freedom approximation that is computed automatically in robust variance estimation software (Pustejovsky, 2017). If the included studies all have a similar number of cases (e.g., all studies have 3 or 4 cases), then the degrees of freedom will be approximately $f \approx K - 1$. This confidence interval is robust to use of inaccurate sampling variances and variance components because it uses the robust standard error and small-sample degrees of freedom.

Converting to Percentage Change

As an aid to substantive interpretation, it is helpful to translate the estimated overall average LRR effect size and its confidence interval into the metric of percentage change. The percentage change equivalent to the overall average LRR effect size can be calculated as $100\% \times [\exp(\hat{\gamma}) - 1]$. A level-$\alpha$ confidence interval is given by

$$100\% \times [exp(\gamma_L) - 1], 100\% \times [exp(\gamma_U) - 1]. \tag{13}$$

I demonstrate these calculations in the subsequent example.

Meta-Regression

The basic multi-level meta-analysis model consists of an overall average effect size and variance components, which characterize the extent to which the effects vary across cases and across studies. In order to examine whether participant or study characteristics moderate the magnitude of effects, the basic meta-analysis model can be extended by including predictor variables that encode these characteristics. The resulting model is known as a meta-regression or mixed-effects model.

Suppose we wish to examine whether the magnitude of effects varies depending on the age of the participant. Let $(\text{Age})_{jk}$ denote the age in years of case *j* within study *k*, centered at the average age of the full sample of cases. The combined form of the meta-regression would be

$$R_{jk} = \gamma_0 + \gamma_1 (Age)_{jk} + v_k + u_{jk} + e_{jk}. \tag{14}$$

Here, γ_0 corresponds to the overall average effect size when $(\text{Age})_{jk}$ is equal to zero and γ_1 corresponds to the *difference* in average effect sizes between cases that differ in age by one year. Positive values of γ_1 would indicate that larger intervention effects are associated with older participants. In this model, the study-level errors (v_k) and case-level errors (u_{jk}) now represent *residual* variation in true effect sizes, after accounting for variation due to age.

Meta-regression models are quite flexible in that they can include one or multiple predictor variables, which vary across cases or vary only across studies; indicator variables for categorical moderators can also readily be included. The software packages mentioned previously provide functionality for estimating meta-regression models, including with robust variance estimation. For a more detailed discussion of meta-regression models, readers may refer to Borenstein and colleagues (2009, Chapters 20–21).

Example: Meta-Analysis of Group Contingency Interventions

Maggin and colleagues (2017) conducted a systematic review and synthesis of single-case studies on group-contingency interventions. Using systematic search and review criteria, they identified 40 studies that met What Works Clearinghouse design standards for SCDs with or without reservations. The authors' original meta-analysis of these studies was based on the BC-SMD effect size index (Shadish et al., 2014), a metric designed to facilitate direct comparison with effect sizes from between-groups experimental studies.

I re-analyzed the studies identified by Maggin and colleagues (2017) using the LRR effect size. This analysis complements the original analysis in several ways. First, the LRR can be applied to data from all studies that met inclusion criteria, whereas the BC-SMD could be estimated for only 27 (68%) of included studies due to technical limitations of the index. Second, the LRR is a case-level effect size index and can therefore be used to investigate research questions that pertain to individual-level variation, such as: To what extent does the magnitude of intervention effects vary across cases from the same study? To what extent do intervention effects vary based on individual participant characteristics such as age or behavioral function?

Such questions cannot be addressed with the BC-SMD because it is a study-level effect size that pertains to *average* effects across cases, potentially concealing heterogeneity across cases in a given study.

For purposes of illustration, I focused on the 33 studies and 111 cases that assessed effects of group contingency interventions on problem behavior. All calculations were conducted in R, using the metafor package for meta-analysis and meta-regression (Viechtbauer, 2010) and the clubSandwich package for robust variance estimation (Pustejovsky, 2017). Complete raw data and computer code for replicating all calculations are available in the supplementary materials (https://osf.io/c3fe9/).

Included studies assessed problem behavior using a variety of different recording procedures, including event counting, momentary time sampling, and partial interval recording, all of which yielded ratio-scale outcome measures. Because some outcomes were quantified as proportions or percentages, the effect size estimates based on the LRR-d form of the index differ in magnitude from those based on LRR-i. Given that problem behavior is a negative-valence outcome, I focused on the LRR-d form of the index, although I also computed LRR-i for comparison purposes.

Figure 1 displays density plots of both LRR-d and LRR-i, with separate distributions plotted for outcomes measured on a count/rate metric or a proportion/percentage metric. In the left-hand panel, it can be seen that there are only minor differences in the distributions of LRR-d. In contrast, the LRR-i estimates for proportion/percentage metrics are much smaller and less symmetrically distributed than the LRR-i estimates for count/rate metrics. The substantial difference between the two distributions suggests that the LRR-i form of the index may be less appropriate for synthesizing effects across both types of outcome metrics. Following on the theoretical framework described in Pustejovsky (2015), the LRR-d index would be expected to be comparable across outcome metrics if interventions primarily affect the frequency or spacing of behaviors, but not the duration of individual behavioral episodes. This is a plausible theory of how group contingency interventions affect behavior and so lends further support for the choice to focus on LRR-d.

Model 1 in Table 4 reports the overall average LRR-d effect size estimate, robust standard error, degrees of freedom, confidence interval, and variance component estimates. Across all 111 cases from 33 studies, the overall effect of group contingencies was estimated as -1.18, 95% CI: [-1.35, -1.01], which corresponds to a decrease in problem behavior of 69%, 95% CI: [64%, 74%]. The between-study variance in average effects was estimated as $\hat{\tau}^2 = 0.180.$ To characterize the magnitude of this variance component estimate, consider that if average effects are normally distributed, then approximately two thirds of effects should fall within one SD of the average effect, or between $\hat{\gamma} - \hat{\tau} = -1.61$ (% change = -80%) and $\hat{\gamma} + \hat{\tau} = -0.76$ (% change = -53%). Thus, there is a substantial degree of heterogeneity in effects across studies. In comparison, the within-study variance in individual-specific treatment effects was estimated as $\widehat{\omega}^2 = 0.045$, substantially smaller than the between-study variance.

In addition to estimates of the overall average effect size and variance components, it is of interest to identify characteristics of the participants, interventions, or studies that explain variation in the magnitude of intervention effects. The original analysis by Maggin and colleagues (2017) examined a large number of potential moderators. For purposes of illustration, I consider just two: study setting (general versus special education class) and unit of analysis (group-level or individual). Of the 33 included studies, 24 were conducted in general education settings and 9 were conducted in special education classes. Further, 24 studies examined effects at the group level (i.e., assessing behavior at the aggregate classroom level) and 9 examined effects for individual focal students.

To examine the extent to which these factors account for variation in magnitude of effects, I fit a meta-regression model that included separate intercepts for each combination of setting and unit of analysis. Results are reported in Model 2 of Table 4. Average effect sizes within each of the four sub-groups are statistically distinguishable from zero at the 5% significance level. The results indicate an interaction between moderators. For studies in general education settings, effects for studies with individuals as the unit of analysis were larger than for studies with groups as the unit of analysis (estimated difference = -0.70, SE = 0.26, *p* = .036), whereas for studies in special education settings, effects for studies with individual as the unit of analysis were smaller than for studies with groups as the unit of analysis, although the difference is not statistically distinguishable from zero (estimated difference = 0.33, SE = 0.29, *p* = .29). The residual between-study variance in Model 2 was estimated as $\hat{\tau}^2 = 0.106$, indicating that the combination of study setting and unit of analysis explain 42% of the between-study variation in average effects. An important caveat to these findings is that the moderators were not identified *a priori* (i.e., as part of pre-registered analysis plan) and so must be considered purely exploratory.

Discussion

In this paper, I have demonstrated the use of a recently proposed effect size index, the log response ratio, for meta-analysis of SCDs with behavioral outcome measures. Compared to meta-analysis based on other effect size indices, the proposed methods are distinctive in several respects.

First, development of the LRR was motivated by a realistic model for systematic direct observation procedures (Pustejovsky, 2015), and the index is thus designed to work well with behavioral outcomes. Other effect size indices, such as non-overlap measures or within-case standardized mean differences, do not specifically account for the features of how behavioral outcomes are measured. In a synthesis of SCDs examining effects of choice-making on directly observed behavioral outcomes, Pustejovsky (2015) reported an example in which meta-analysis produced uninterpretable results when based on the within-case standardized mean difference (WC-SMD) index. More broadly, Pustejovsky (2018) demonstrated using simulations that several non-overlap measures, as well as the WC-SMD, are influenced by incidental details of how behavioral outcomes are assessed, such as use of longer or shorter observation sessions. Such procedural sensitivity is problematic for interpreting these indices as measures of effect magnitude.

Compared to other effect indices, the calculations involved in estimating the LRR entail some additional complexities, including the use of truncated sample means and the need to attend to the valence of outcomes measured as percentages or proportions. These complexities are a direct consequence of how the LRR accounts for the features of behavioral outcome data. Furthermore, the estimates can still readily be computed by hand (or with a spreadsheet) from basic summary statistics, as evidenced by the examples provided in previous sections. On balance, this additional cost seems worth the return of more interpretable meta-analysis results.

Another distinctive feature of the LRR is its close connection to using percentage change between phases as an effect measure. Others have argued that percentage change is a conceptually appealing and intuitive index for quantifying the magnitude of functional relationships (Campbell & Herzinger, 2010; Marquis et al., 2000). Indeed, many researchers use percentage change as an “informal” effect measure in primary studies (e.g., Call, Simmons, Mevers, & Alvarez, 2015) and even in some systematic reviews (Heyvaert et al., 2014; Kahng et al., 2002; Marquis et al., 2000). As I have demonstrated, it is possible to translate LRR estimates and meta-analytic averages directly into percentage change terms. Similar translation approaches are used in other areas of meta-analysis, such as when bivariate correlations are meta-analyzed on the Fisher-*z* scale but translated into Pearson-*r* coefficients for purposes of interpretation or when randomized trials with binary outcomes are meta-analyzed using log-odds ratios but interpreted in terms of percentage changes relative to specified levels of baseline risk.

A final distinctive aspect of the methods I have described is the use of robust variance estimation (Hedges et al., 2010) to account for potential auto-correlation in the data. Although the presence and consequences of auto-correlation in SCD time series has long been debated (Huitema & McKean, 1998; Matyas & Greenwood, 1996), scholars have recently emphasized that statistical methods for SCDs should not merely assume it away (Horner et al., 2012; Wolery et al., 2010). Robust variance estimation methods effectively side-step the auto-correlation issue. Instead of trying to estimate and account for the degree of auto-correlation in individual data series, they provide a means to synthesize effect size estimates and conduct moderator analysis that remains valid regardless of the presence or absence of auto-correlation. This approach is particularly appealing given the challenges of detecting and estimating even simple forms of serial dependence in short time series (Huitema & McKean, 2007).

Although I have presented robust variance estimation methods in the context of meta-analyzing LRR effect sizes, these techniques are not limited to this single effect size. Use of robust variance estimation has also been recommended for meta-analyzing BC-SMD indices (Zelinsky & Shadish, 2016). They could also be applied for meta-analyzing other case-level effect size indices, such as non-overlap of all pairs, for which sampling variance formulas are valid only in the absence of auto-correlation. Use of robust variance estimation in these contexts warrants further investigation.

Limitations

The methods described in this paper have several limitations, which point towards areas in need of further methodological research. First, appropriately assessing and accounting for time trends is a critical part of visual assessment of single-case data (Kratochwill et al., 2014) and is likewise important for valid statistical analysis (Horner & Kratochwill, 2012). However, available methods for estimating the LRR assume that the level of the outcome is constant within each phase—an assumption that would be violated if the outcomes follows a systematic time trend during the baseline phase or if the full effect of intervention is not immediate. In further work, I am investigating how to address this limitation of the LRR by developing a generalized non-linear regression model, in which certain parameters correspond to LRR effect sizes (Swan & Pustejovsky, 2017). This non-linear model accounts for time trends during treatment phases (as well as return-to-baseline phases in reversal designs) that arise when an intervention has gradual effects on an outcome. Using similar modeling techniques, it may be possible to extend the LRR to handle baseline time trends as well. Until such extensions become available, researchers conducting syntheses of SCDs will need to use theory and visual analysis to determine whether it is reasonable to apply LRR estimation methods that assume no time trends.

Second, the variance estimator for the LRR is premised on the assumption that the outcome data are independent. I have proposed robust variance estimation methods for use in meta-analysis of LRR estimates, which work even when the sampling variance of the effect size is not valid. However, this limitation of the variance estimator remains a problem if researchers wish to use an LRR estimate to draw inferences about an individual case. I have emphasized that the independence assumption should be acknowledged whenever analysts report standard errors for individual LRR estimates. Furthermore, this paper has purposely *not* described methods for conducting inferential statistical tests or constructing confidence intervals for individual LRR values because such methods carry the same limitations as the standard error, and so may have limited utility.

Future research should investigate methods that work with auto-correlated outcome data. This might be possible either by inverting randomization tests (cf. Michiels, Heyvaert, Meulders, & Onghena, 2017) or by developing estimators to directly quantify auto-correlation in the outcome data. Such methods could provide evidence about the degree to which auto-correlation is a concern for behavioral outcome data, as well as inferential tests or confidence intervals for LRR effect sizes when considered individually, rather than as part of a synthesis.

A third limitation is that the LRR is a within-case effect size and—like many other effect size indices for SCDs—is not directly comparable to any effect size for between-group experimental designs. Currently, the BC-SMD (Pustejovsky, Hedges, & Shadish, 2014; Shadish et al., 2014) is the only available effect size metric that is on a common scale across both types of research designs. The BC-SMD has been applied and interpreted as a general-purpose effect size measure for SCDs (i.e., used regardless of the class of outcome measure), yet there remain outstanding questions regarding the extent to which its underlying assumptions are robust when applied to behavioral outcome data. Recognizing this, the developers of the BC-SMD index have noted the need for other effect size metrics that are appropriate for the types of outcome measurements generated by behavioral observation systems (Shadish et al., 2015, p. 82). This need could be met by developing a “between-case” extension of the LRR or, more broadly, by developing methods for joint synthesis of SCDs and between-case designs with behavioral outcome measures. Still, the within-case LRR remains useful for syntheses of SCD results.

Fourth, when applied to outcomes measured as proportions (or percentages), the magnitude of LRR effect sizes depends on whether the LRR-d or LRR-i form is used, and researchers must decide which is a more appropriate measure of effect magnitude. I have suggested several factors that can inform this decision, including theoretical considerations, drawing on the framework from Pustejovsky (2015); the predominant valence of outcomes in the studies to be summarized; and the degree of alignment in the distribution of effect sizes between frequency count outcomes and proportion outcomes. Visual analysis could also be informative, by examining which form of the index better corresponds to visual determinations of effect magnitude. On a broader level, there is a need to understand the degree of correspondence between the LRR and visual analysis. Some degree of discrepancy might be expected because visual inspection is typically conceived as an inferential technique (Kratochwill et al., 2014), which involves assessing the *degree of evidence* for a functional relationship, similar to a hypothesis test. In contrast, the LRR estimates the magnitude of a functional relationship *separately* from its degree of certainty.

Finally, the utility of LRR effect sizes is limited by definition to contexts where percentage change is a meaningful and interpretable way to quantify effect magnitude, and so it will not work well for all research areas where SCDs are used. At a basic level, percentage change is only meaningful for outcomes measured on ratio scales, where a score of zero corresponds to the total absence of the outcome. Thus, it will not be appropriate for outcomes such as rating scale measures of student engagement. At a more substantive level, percentage change is unlikely to be a meaningful way to quantify effects of interventions on behaviors that are totally or nearly absent during baseline, such as the number of words read correctly for a student who cannot read, because practically *any* improvement in behavior will appear very large in percentage terms. Similarly, percentage change might not be meaningful for describing effects of interventions that consistently produce total extinction of a behavior (i.e., 100% reductions) because all of the effect sizes will be at or near ceiling levels. In such instances, quantities that capture other features of the intervention's effects, such as duration of treatment needed to extinguish the behavior, might be more relevant and meaningful measures of effect size.

This final limitation of the LRR highlights the need to consider the context of application—including especially the types of outcome measures reported in a set of studies to be synthesized—when selecting an effect size index for meta-analysis of SCDs. Rather than searching for generic metrics to be applied across any set of SCDs, the field should instead consider developing metrics that work well in circumscribed areas of application. Following this logic, I have proposed the log response ratio as an effect size metric for meta-analyzing SCDs with behavioral outcomes. Although limited to a single outcome domain, the prevalence and prominence of direct observation measures within single-case research suggests that the log response ratio might nonetheless find broad application.

Acknowledgements

This work was supported by Grant R305D160002 from the Institute of Educational Sciences, U.S. Department of Education. The opinions expressed are those of the author and do not represent the views of the Institute or the U.S. Department of Education. The author is grateful to Daniel Maggin, David Rindskopf, Tsuyoshi Yamada, and Kathleen Zimmerman for feedback on draft versions of this paper.

References

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.

Borenstein, M., Hedges, L. V, Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester, UK: John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470743386

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.

Call, N. A., Simmons, C. A., Mevers, J. E. L., & Alvarez, J. P. (2015). Clinical outcomes of behavioral treatments for pica in children with developmental disabilities. *Journal of Autism and Developmental Disorders*, *45*(7), 2105–2114. https://doi.org/10.1007/s10803-015-2375-z

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.

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

Council for Exceptional Children Working Group. (2014). Council for Exceptional Children: Standards for evidence-based practices in special education. *TEACHING Exceptional Children*, *46*(6), 206–212. https://doi.org/10.1177/0040059914531389

Geomatrix. (2007). XYit. Retrieved from http://geomatix.net/xyit

Gingerich, W. J. (1984). Meta-analysis of applied time-series data. *The Journal of Applied Behavioral Science*, *20*(1), 71–79. https://doi.org/10.1177/002188638402000113

Heath, A. K., Ganz, J. B., Parker, R. I., Burke, M., & Ninci, J. (2015). A meta-analytic review of functional communication training across mode of communication, age, and disability. *Review Journal of Autism and Developmental Disorders*, *2*(2), 155–166. https://doi.org/10.1007/s40489-014-0044-3

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*(3), 167–171. https://doi.org/10.1111/j.1750-8606.2008.00060.x

Hedges, L. V, Gurevitch, J., & Curtis, P. S. (1999). The meta-analysis of response ratios in experimental ecology. *Ecology*, *80*(4), 1150–1156. https://doi.org/10.1890/0012-9658(1999)080[1150:TMAORR]2.0.CO;2

Hedges, L. V, Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods*, *1*(1), 39–65. https://doi.org/10.1002/jrsm.5

Heyvaert, M., Saenen, L., Campbell, J. M., Maes, B., & Onghena, P. (2014). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: An updated quantitative synthesis of single-subject research. *Research in Developmental Disabilities*, *35*(10), 2463–2476. https://doi.org/10.1016/j.ridd.2014.06.017

Hitchcock, J. H., Kratochwill, T. R., & Chezan, L. C. (2015). What Works Clearinghouse standards and generalization of single-case design evidence. *Journal of Behavioral Education*, *24*(4), 459–469. https://doi.org/10.1007/s10864-015-9224-1

Horner, R. H., & Kratochwill, T. R. (2012). Synthesizing single-case research to identify evidence-based practices: Some brief reflections. *Journal of Behavioral Education*, *21*(3), 266–272. https://doi.org/10.1007/s10864-012-9152-2

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. https://doi.org/10.1353/etc.2012.0011

Huitema, B. E., & McKean, J. W. (1998). Irrelevant autocorrelation in least-squares intervention models. *Psychological Methods*, *3*(1), 104–116. https://doi.org/10.1037//1082-989X.3.1.104

Huitema, B. E., & McKean, J. W. (2007). Identifying Autocorrelation Generated by Various Error Processes in Interrupted Time-Series Regression Designs: A Comparison of AR1 and Portmanteau Tests. *Educational and Psychological Measurement*, *67*(3), 447–459. https://doi.org/10.1177/0013164406294774

Kahng, S., Iwata, B. a, & Lewin, A. B. (2002). Behavioral treatment of self-injury, 1964 to 2000. *American Journal of Mental Retardation : AJMR*, *107*(3), 212–221. https://doi.org/10.1352/0895-8017(2002)107<0212:BTOSIT>2.0.CO;2

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education*, *34*(1), 26–38. https://doi.org/10.1177/0741932512452794

Kratochwill, T. R., Levin, J. R., Horner, R. H., & Swoboda, C. M. (2014). Visual analysis of single-case intervention research: Conceptual and methodological issues. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-Case Intervention Research: Methodological and Statistical Advances* (pp. 91–125). Washington, DC: American Psychological Association.

Lane, K. L., Kalberg, J. R., & Shepcaro, J. C. (2009). An examination of the evidence base for function-based interventions for students with emotional and/or behavioral disorders attending middle and high schools. *Exceptional Children*, *75*(3), 321–340.

Ledford, J. R., King, S., Harbin, E. R., & Zimmerman, K. N. (2016). Antecedent social skills interventions for individuals with ASD: What works, for whom, and under what conditions? *Focus on Autism and Other Developmental Disabilities*, forthcoming. https://doi.org/10.1177/1088357616634024

Littell, R. C., Milliken, G. A., Stroup, W. W., Wolfinger, R. D., & Schabenberber, O. (2006). *SAS system for linear mixed models*. Cary, NC: SAS Institute.

Maggin, D. M. (2015). Considering generality in the systematic review and meta-analysis of single-case research: A response to Hitchcock et al. *Journal of Behavioral Education*, *24*(4), 470–482. https://doi.org/10.1007/s10864-015-9239-7

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. https://doi.org/10.1080/09362835.2011.565725

Maggin, D. M., Pustejovsky, J. E., & Johnson, A. H. (2017). A meta-analysis of school-based group contingency interventions for students with challenging behavior: An update. *Remedial and Special Education*, *38*(6), 353–370. https://doi.org/10.1177/0741932517716900

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), 301–321. https://doi.org/10.1016/j.jsp.2011.03.004

Manolov, R., & Moeyaert, M. (2017). Recommendations for choosing single-case data analytical techniques. *Behavior Therapy*, *48*(1), 97–114. https://doi.org/10.1016/j.beth.2016.04.008

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.

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.

McKissick, C., Hawkins, R. O., Lentz, F. E., Hailley, J., & McGuire, S. (2010). Randomizing multiple contingency components to decrease disruptive behaviors and increase student engagement in an urban second-grade classroom. *Psychology in the Schools*, *47*(9), 944–959. https://doi.org/10.1002/pits.20516

Michiels, B., Heyvaert, M., Meulders, A., & Onghena, P. (2017). Confidence intervals for single-case effect size measures based on randomization test inversion. *Behavior Research Methods*, *49*(1), 363–381. https://doi.org/10.3758/s13428-016-0714-4

Morano, S., Ruiz, S., Hwang, J., Wertalik, J. L., Moeller, J., Karal, M. A., & Mulloy, A. (2017). Meta-analysis of single-case treatment effects on self-injurious behavior for individuals with autism and intellectual disabilities. *Autism & Developmental Language Impairments*, *2*, 1–26. https://doi.org/10.1177/2396941516688399

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single-case research: Nonoverlap of all pairs. *Behavior Therapy*, *40*(4), 357–67. https://doi.org/10.1016/j.beth.2008.10.006

Pustejovsky, J. E. (2015). Measurement-comparable effect sizes for single-case studies of freeoperant behavior. *Psychological Methods, 20*(3), 342–359. https://doi.org/10.1037/met0000019

Pustejovsky, J. E. (2017). clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections. Retrieved from https://cran.r-project.org/package=clubSandwich

Pustejovsky, J. E. (2018). Procedural sensitivities of effect sizes for single-case designs with behavioral outcome. *Psychological Methods*, forthcoming. Retrieved from https://osf.io/p3nuz/

Pustejovsky, J. E., Hedges, L. V, & Shadish, W. R. (2014). Design-comparable effect sizes in multiple baseline designs: A general modeling framework. *Journal of Educational and Behavioral Statistics, 39*(5), 368–393. https://doi.org/10.3102/1076998614547577

Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). *GLLAMM Manual* (Division of Biostatistics Working Paper Series No. 160). Berkeley, CA. Retrieved from http://biostats.bepress.com/ucbbiostat/paper160

Rindskopf, D. M., & Ferron, J. M. (2014). Using multilevel models to analyze single-case design data. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-case intervention research: Methodological and statistical advances.* (pp. 221–246). Washington, DC: American Psychological Association. https://doi.org/10.1037/14376-008

Schmidt, A. C. (2007). *The effects of a group contingency on group and individual behavior in an urban first-grade classroom*. University of Kansas. Retrieved from http://gradworks.umi.com/14/43/1443719.html

Scruggs, T. E., Mastropieri, M. A., & Casto, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*(2), 24–43. https://doi.org/10.1177/074193258700800206

Shadish, W. R. (2014a). Analysis and meta-analysis of single-case designs: An introduction. *Journal of School Psychology*, *52*(2), 109–122. https://doi.org/10.1016/j.jsp.2013.11.009

Shadish, W. R. (2014b). Statistical analyses of single-case designs: The shape of things to come. *Current Directions in Psychological Science*, *23*(2), 139–146. https://doi.org/10.1177/0963721414524773

Shadish, W. R., Hedges, L. V, Horner, R. H., & Odom, S. L. (2015). *The role of between-case effect size in conducting, interpreting, and summarizing single-case research*. Washington, DC. Retrieved from http://ies.ed.gov/ncser/pubs/2015002/

Shadish, W. R., Hedges, L. V, & Pustejovsky, J. E. (2014). Analysis and meta-analysis of single-case designs with a standardized mean difference statistic: A primer and applications. *Journal of School Psychology*, *52*(2), 123–147. https://doi.org/10.1016/j.jsp.2013.11.005

Solomon, B. G. (2014). Violations of assumptions in school-based single-case data: Implications for the selection and interpretation of effect sizes. *Behavior Modification*, *38*(4), 477–496. https://doi.org/10.1177/0145445513510931

Solomon, B. G., Howard, T. K., & Stein, B. L. (2015). Critical Assumptions and Distribution Features Pertaining to Contemporary Single-Case Effect Sizes. *Journal of Behavioral Education*, *24*(4), 438–458. https://doi.org/10.1007/s10864-015-9221-4

Swan, D. M., & Pustejovsky, J. E. (2017). A gradual effects model for single-case designs. Retrieved from https://osf.io/f3mr2/

Tanner-Smith, E. E., & Wilson, S. J. (2013). A meta-analysis of the effects of dropout prevention programs on school absenteeism. *Prevention Science*, *14*(5), 468–478. https://doi.org/10.1007/s11121-012-0330-1

Tate, R. L., Perdices, M., Rosenkoetter, U., Shadish, W. R., Vohra, S., Barlow, D. H., … Wilson,

B. (2016). The Single-Case Reporting Guideline In BEhavioural Interventions (SCRIBE) 2016 Statement. *Evidence-Based Communication Assessment and Intervention*, *10*(1), 44–58. https://doi.org/10.1080/17489539.2016.1190525

Tipton, E. (2014). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375–393. https://doi.org/10.1037/met0000011

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. https://doi.org/10.1080/17489530802505362

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

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.

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. https://doi.org/10.1177/0022466908328009

Wong, C., Odom, S. L., Hume, K. A., Cox, A. W., Fettig, A., Kucharczyk, S., … Schultz, T. R. (2015). Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. *Journal of Autism and Developmental Disorders*, *45*(7), 1951–1966. https://doi.org/10.1007/s10803-014-2351-z

Zelinsky, N. A. M., & Shadish, W. R. (2016). A demonstration of how to do a meta-analysis that combines single-case designs with between-groups experiments: The effects of choice making on challenging behaviors performed by people with disabilities. *Developmental Neurorehabilitation*, forthcoming. https://doi.org/10.3109/17518423.2015.1100690

Table 1. Summary statistics and LRR effect size estimates for frequency of disruptive behavior data from McKissick et al. (2010).

	Baseline phase			Treatment phase					
Case	$\tilde{y}_A$	s_A	M	$\tilde{y}_B$	s_B	n	R_1	R_2	SE^R
Period 1	13.983	1.626	3	6.146	3.025	7	-0.822	-0.807	0.198
Period 2	17.652	5.577	5	9.211	7.766	7	-0.650	-0.610	0.349
Period 3	13.441	2.330	9	5.997	4.183	4	-0.807	-0.748	0.354

Table 2. Summary statistics by phase for disruptive behavior and on-task behavior data from Schmidt (2007).

Case	Phase A1			Phase B1			Phase A2			Phase B2		
	$\tilde{y}_{A1}$	s_{A1}	m_1	$\tilde{y}_{B1}$	s_{B1}	n_1	$\tilde{y}_{A2}$	s_{A2}	m_2	$\tilde{y}_{B2}$	s_{B2}	n_2
Disruptive behaviors (frequency count)												
Albert	18.63	7.16	9	3.61	3.03	8	20.29	9.48	3	3.90	2.28	5
Faith	23.38	13.09	8	7.37	3.05	14	21.52	8.93	3	5.21	2.29	5
Lilly	29.31	11.43	9	5.07	3.01	13	16.67	9.46	3	6.23	6.51	6
On-task behavior (% duration)												
Albert	67.69	26.01	9	94.22	4.70	8	71.63	23.70	3	92.67	12.79	5
Faith	75.44	13.38	8	76.36	27.31	14	41.74	48.55	3	95.71	2.22	5
Lilly	59.80	28.25	9	92.17	12.83	13	88.18	17.32	3	93.49	6.54	6

Table 3. LRR-d and LRR-i effect size estimates and variances for disruptive behavior and on-task behavior data from Schmidt (2007).

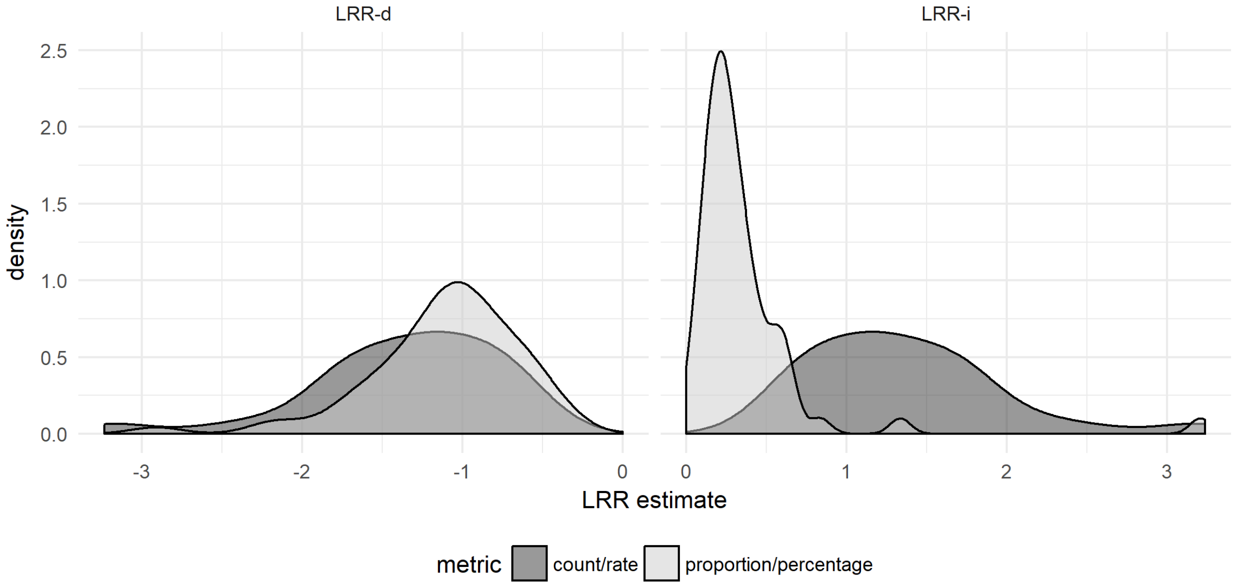
	LRR-d						LRR-i	
	A1B1		A2B2		Combined		Combined	
Case	R_2^1	V^{R1}	R_2^2	V^{R2}	R_2	V^R	R_2	V^R
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Disruptive behaviors (frequency count)								
Albert	-1.605	0.104	-1.651	0.141	-1.628	0.061	1.628	0.061
Faith	-1.168	0.051	-1.428	0.096	-1.298	0.037	1.298	0.037
Lilly	-1.749	0.044	-0.947	0.289	-1.348	0.083	1.348	0.083
On-task behavior (% duration)								
Albert	-1.716	0.155	-1.165	0.842	-1.440	0.249	0.282	0.014
Faith	-0.009	0.132	-2.698	0.285	-1.353	0.104	0.310	0.116
Lilly	-1.560	0.261	-0.870	0.884	-1.215	0.286	0.237	0.010

Table 4. Meta-analysis results based on LRR-d effect sizes for problem behavior outcomes from Maggin et al. (2017).

	Studies	Cases	Est.	SE	d.f.	95% CI	$\hat{\tau}^2$	$\widehat{\omega}^2$
Model 1							0.180	0.045
Overall average	33	110	-1.18	0.08	31.1	[-1.35, -1.01]		
Model 2							0.105	0.046
General Ed., group	19	46	-0.95	0.07	17.1	[-1.11, -0.80]		
General Ed., individual	5	22	-1.65	0.25	3.9	[-2.36, -0.95]		
Special Ed., group	5	15	-1.53	0.15	3.6	[-1.98, -1.09]		
Special Ed., individual	4	28	-1.21	0.24	2.8	[-2.01, -0.41]		

Notes: Est. = estimate. SE = standard error. d.f. = small-sample degrees of freedom. CI = confidence interval. $\hat{\tau}^2$ = estimated between-study variance. $\widehat{\omega}^2$ = estimated within-study variance.

Figure 1. Distribution of LRR-d and LRR-i effect size estimates by outcome metric.


LRR-d
LRR-i
density
LRR estimate
metric
count/rate
proportion/percentage