Single-Case Synthesis Tools II: Comparing Quantitative Outcome Measures

Kathleen N. Zimmerman[1], James E. Pustejovsky[2], Jennifer R. Ledford[1], Erin E. Barton[1], Katherine E. Severini[1], & Blair P. Lloyd[1]

[1]Vanderbilt University [2]University of Texas at Austin

January 2, 2018. Forthcoming in *Research in Developmental Disabilities*. This paper is not the version of record and may not exactly replicate the final, published version of the article. The final article will be available, upon publication, at https://doi.org/10.1016/j.ridd.2018.02.001

Author acknowledgements: The research reported in this article was supported in part by Grant R305D160002 from the Institute of Educational Sciences, U.S. Department of Education. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education.

Abstract

Varying methods for evaluating the outcomes of single case research designs (SCD) are currently used in reviews and meta-analyses of interventions. Quantitative effect size measures are often presented alongside visual analysis conclusions. Six measures across two classes—overlap measures (percentage non-overlapping data, improvement rate difference, and Tau) and parametric within-case effect sizes (standardized mean difference and log response ratio [increasing and decreasing])—were compared to determine if choice of synthesis method within and across classes impacts conclusions regarding effectiveness. The effectiveness of sensory-based interventions (SBI), a commonly used class of treatments for young children, was evaluated. Separately from evaluations of rigor and quality, authors evaluated behavior change between baseline and SBI conditions. SBI were unlikely to result in positive behavior change across all measures except IRD. However, subgroup analyses resulted in variable conclusions, indicating that the choice of measures for SCD meta-analyses can impact conclusions. Suggestions for using the log response ratio in SCD meta-analyses and considerations for understanding variability in SCD meta-analysis conclusions are discussed.

Keywords: single case research design; meta-analysis; overlap measures; within-case effect sizes

Single-Case Synthesis Tools II: Comparing Quantitative Outcome Measures

Federal law requires special education practitioners to implement evidence-based practices (EBPs) identified through rigorous, experimental research including single case research designs (SCD; Individuals with Disabilities Education Act [IDEA], 2004). Identifying EBPs may increase the likelihood that children with disabilities receive effective treatments, thus improving long-term academic and social outcomes. Classifying interventions based on their evidence of effectiveness may facilitate the selection of appropriate, effective treatments. This can help families avoid the unnecessary stress, financial strains, or resource burden associated with ineffective, popular treatments.

Antecedent Sensory-Based Interventions

Antecedent sensory-based interventions (SBI; e.g., weighted vests, weighted blankets, therapy balls, therapy cushions, multi-sensory environments) are a popular treatment for young children with autism spectrum disorders (ASD) to improve engagement and problem behaviors (May-Benson & Koomar, 2010). Based on the theory of sensory integration (Ayres, 1979), antecedent SBI are interventions implemented before children interact with their environment, which are intended to aide in processing sensory input in the environment (Blanche, Chang, Guiterrez, & Gunter, 2016). However, conclusions regarding their effectiveness vary across existing syntheses (Barton, Reichow, Schnitz, Smith, & Sherlock, 2015; Case-Smith, Weaver, & Fristad, 2015; Lang et al., 2012; Leong, Carter, & Stephenson, 2015; May-Benson & Koomar, 2010; Watling & Hauer, 2015; Yunnus et al., 2015). Investigating sources of variability in intervention reviews—particularly for frequently used interventions with mixed results such as antecedent SBI—may improve and advance the methods used to identify EBPs.

Synthesizing Evidence

Visual analysis is the traditional method for assessing study outcomes in SCD (Ledford, Wolery, & Gast, 2014). Systematic procedures for conducting visual analysis exist (cf. Ledford, Lane, & Severini, 2017), but visual analysis does not produce a single summary measure to quantify the magnitude of behavior change. The lack of a numerical index of effectiveness makes it difficult for researchers to quantitatively summarize outcomes across multiple SCD studies. In contrast, quantitative measures of effectiveness allow SCD researchers to use statistical procedures such as meta-analysis to more clearly synthesize evidence from SCD studies.

Meta-analysis is a set of statistical procedures for summarizing and studying patterns of evidence from multiple studies on a single topic or intervention (Borenstein, Hedges, Higgins, & Rothstein, 2009). When meta-analysis is used, the average magnitude and distribution of intervention effects are estimated by combining results—typically in the form of effect size estimates—from each study (Borenstein et al., 2009). Meta-analytic techniques for synthesizing group design research are now well established and used extensively throughout the social and behavioral sciences (Beretvas, 2010; Borenstein et al., 2009; Kavale, 2007). Findings from meta-analyses of group design studies are considered one of the highest standards for determining the overall effectiveness of an intervention (Cook et al., 2015; Shadish, Rindskopf, & Hedges, 2007). However, meta-analytic techniques for synthesizing evidence from SCD have not yet achieved the same degree of consensus. Existing systematic reviews and syntheses of SCD have most frequently used measures representing the degree of overlap in data level between adjacent conditions (i.e., overlap measures; Heyvaert et al., 2015; Maggin et al., 2011), but several other types of effect size measures and synthesis techniques are available, including some that have only recently been introduced.

To be useful for purposes of meta-analysis, an effect size index should “describe the magnitude of an effect on a common scale” for a body of research on a common topic or intervention (Shadish, Hedges, Horner, & Odom, 2015, p. 27). When conducting systematic reviews of studies using SCD, standardized effect sizes facilitate reviewers’ ability to (a) combine results across SCD using a common scale of magnitude and (b) combine results across multiple SCD that investigate a common intervention (Shadish et al., 2015). However, it is not yet clear which effect size indices are well suited to describe SCD data.

Overlap Measures

Overlap measures (Chen et al., 2016; Parker, Vannest, & Davis, 2011; Vannest & Ninci, 2015) are numerical indices used to quantify the extent to which data overlap in level between adjacent conditions. Percentage of non-overlapping data (PND) was one of the first quantitative tools used to synthesize SCD (Scruggs, Mastropieri, & Casto, 1987; Scruggs & Mastropieri, 1998). Since the introduction of PND, several additional overlap measures have been proposed and applied, including the improvement rate difference (IRD; Parker, Vannest, & Brown, 2009), non-overlap of all pairs (NAP; Parker & Vannest, 2009), Kendall’s Tau for non-overlap between groups ($\text{Tau}_{\text{novlap}}$ or Tau; Parker, Vannest, Davis, & Sauber, 2011), and Tau-U (Parker et al., 2011b). Each measure quantifies the amount of overlap between adjacent conditions differently (Parker, Vannest, & Davis, 2011), although some evidence indicates various overlap measures are highly inter-correlated (Chen et al., 2016).

Several studies have been conducted comparing overlap measures with visual analysis (Barton et al., 2016; Chen et al., 2016; Ma, 2006; Parker & Hagan-Burke, 2007; Parker & Vannest, 2009; Parker et al., 2011a; Rakap, Snyder, & Pasia, 2014; Wolery et al., 2010). One study found visual analysis and overlap measures lead to similar conclusions (Parker et al., 2011a). However, comparison studies have not led to convergent conclusions regarding which overlap measures most closely align with visual analysis. When visual analysis suggests an effect is present, overlap measures have produced variable conclusions (Ma, 2006; Rakap et al., 2014; Wolery et al., 2010). This may not be a critical flaw, given that visual analysis is designed to determine *whether a functional relation exists*, whereas overlap measures are designed to describe overlap of data. Conclusions may also be variable because decisions about functional relations are based on six different data characteristics (trend, level, variability, overlap, immediacy of change, consistency), whereas overlap measures solely examine one (overlap). However, some reviews have reported overlap measures in the absence of functional relation conclusions (e.g., Dart, Collins, Klingbeil, & McKinley, 2014), suggesting authors interpret overlap measures either as an indicator of the presence of a functional relation or as an estimate of magnitude of effect.

Multiple general limitations of overlap measures have been noted (Pustejovsky, 2016b; Pustejovsky, 2016c; Ledford et al., 2016; Rakap et al, 2014; Wolery et al., 2010). Overlap measures fail to measure the magnitude of an effect; rather they measure the degree of overlap in the level of data points between adjacent conditions (see Authors, in press for a detailed discussion including illustrative examples). Overlap measures do not address replication logic (Wolery et al., 2010), a critical feature of SCD that determines if results are interpretable (Ledford et al., 2014; Wolery et al., 2010). Limited sensitivity of overlap measures may make them unsuitable for quantifying intervention effects (Barton et al., 2016; Chen et al., 2016). The magnitude of overlap measures is also influenced by procedural factors such as design type (Chen et al., 2016), the number of data points in a condition (Pustejovsky, 2016c; Tarlow, 2016), the length of observation sessions, and the type of recording system (Pustejovsky, 2016b).

Consequently, overlap measures do not provide a fair basis for comparing multiple SCD that vary on one or more of these procedural dimensions (Pustejovsky, 2016b). Furthermore, data patterns such as outliers, extinction bursts, and delayed change, even when predicted, also influence the magnitude of overlap measures (Barton et al., 2016; Ledford et al., 2016; Rakap et al., 2014). Despite a growing body of evidence suggesting caution when using overlap measures as quantitative synthesis tools for SCD (cf. Pustejovsky, 2016b), they remain widely used in SCD research (Ledford et al., 2016; Maggin, O'Keefe, & Johnson, 2011). No studies to date have compared the utility of overlap measures to within-case parametric effect size estimates.

Within-Case Parametric Effect Sizes

In contrast to overlap measures, which are defined without reference to distributional assumptions, parametric effect size indices are defined in terms of parametric models for the data. Two types of parametric effect sizes have been proposed for SCD: within-case and between-case. Within-case parametric effect sizes quantify the magnitude of intervention effects for each case (or tier) within an SCD study, whereas between-case effect sizes quantify the magnitude of *average* intervention effects across multiple cases in a study. Because they are case-specific, within-case SCD effect sizes can only be applied to data from SCD. In contrast, between-case effect sizes are constructed so as to be comparable to effect size estimates from group design studies, at least under certain theoretical assumptions about the data-generating process (Shadish et al., 2015). However, even the simplest forms of between-case effect sizes cannot always be calculated from SCD studies because they require data from at least three participants across at least two conditions. Furthermore, because they quantify average effects—rather than examining results for each replication—between-case effect sizes address a somewhat different goal than traditional methods of visual analysis (i.e., average magnitude may not have a direct relationship with functional relation determination) and non-overlap measures.

Multiple within-case effect size indices have been proposed for use with SCD. Gingerich (1984) and Busk and Serlin (1992) introduced a within-case standardized mean difference (SMD) index, defined as the difference in average outcomes between intervention and baseline conditions, scaled by the pooled standard deviation of the outcomes within each condition. Although its mathematical form is similar to the SMD used in group designs, the scaling factor of the within-case SMD captures only within-case variability and so is not comparable to SMD estimates from group designs (Van den Noortgate & Onghena, 2008; Shadish, Hedges, & Pustejovsky, 2014). More recently, Pustejovsky (2014) proposed a family of within-case, parametric effect sizes that quantify intervention effects in terms of proportional change (i.e., percentage change from baseline to intervention). These log response ratio effect size measures may be particularly well suited for studies that use behavioral outcomes assessed through direct observation (Pustejovsky, 2014; Pustejovsky & Ferron, 2017). Although this class of effect size measures has not yet been widely used, conceptualizing effect magnitude in proportional terms has precedents in previous systematic reviews of SCD (e.g., Campbell, 2003; Kahng, Iwata, & Lewin, 2002; Marquis et al., 2000) and, more informally, as a way to convey results of SCD to clinicians (Campbell & Herzinger, 2010).

Purpose and Research Questions

The purpose of this review is to evaluate the extent to which the measure used to estimate the magnitude of treatment effects (overlap measures and within-case effect sizes) impacts conclusions about the effectiveness of antecedent SBI. Three overlap measures (PND, IRD, and Tau), and three parametric, within-case effect sizes (standardized mean difference [SMD], log response ratio-decreasing [LRRd], and log response ratio-increasing [LRRi]) were used to evaluate the effectiveness of antecedent SBI for children in early childhood settings. The overlap measures were selected based on current recommendations for use and correlations between groups of measures (Chen et al., 2016; Rakap et al., 2014; Wolery et al., 2010). The parametric effect sizes were selected based on our evaluation of the within-case effect sizes suitable for use with the types of outcome measures employed in this body of literature (e.g., designs evaluating outcomes for a single case [participant] using interval-scale measurements [e.g., interval-based measures]; see Pustejovsky & Ferron, 2017 for a detailed discussion). Between-case parametric effect sizes were not included because (a) most designs evaluated in this review failed to meet criteria to use between-case effect sizes (only 3 out of 8 articles met criteria; Hodgetts, 2011; Krombach, 2016; Leew, 2010) and (b) between-case effect sizes are not directly comparable to overlap measures, the most commonly used metric in the current SCD literature.

The following research questions guided the review: (1) To what extent do meta-analytic summaries using overlap measures and parametric effect sizes indicate that antecedent SBI result in positive behavior change for young children? (2) To what extent do meta-analytic summaries align with visual analysis conclusions regarding the effectiveness of antecedent SBI? and (3) Do conclusions regarding antecedent SBI differ across and within classes of effect size measures?

Method

Included Studies

Search procedures, article eligibility, and article selection can be found in the companion document to this outcome review (Synthesis Tools Part I). Eleven articles (69 designs) were included in a two-part hierarchical review of antecedent SBI. Designs refer to single SCD (e.g., A-B-A-B, multiple baseline designs). For alternating treatment designs that included three ongoing conditions, three comparisons were present (condition A to B, condition A to C, and condition B to C). First, the quality and rigor of included designs were evaluated (see Synthesis Tools Part I). Then, outcomes were evaluated for studies with at least three potential demonstrations of effect using procedures described below. Eight articles including 21 cases across 17 total outcomes met inclusion criteria (see Table 1).

Data Extraction

Three graduate students in special education (including the first author) used an electronic software program, Version 2.0 of Plot Digitizer (2015), to extract values from each graph to calculate overlap measures and effect sizes. Some obvious extraction errors were adjusted. This includes: (a) Negative values were adjusted to 0 and values over 100 adjusted to 100 for continuous outcomes with a 0-100 scale. (b) Non-integer values were adjusted to the closest whole number for categorical outcomes coded as whole numbers (i.e., alertness states, Tunson & Candler, 2010). (c) Values expressed as rate (count per minute) were rounded to the nearest possible value given the session length and total count. For example, in a 5-min session during which problem behaviors were counted, an extracted 1.19 behaviors per min corresponds to 5.95 behaviors, an impossible value. However, an adjusted value of 1.2 behaviors per min corresponds with a total of 6 behaviors. Values for count and percentage outcomes were not changed after data extraction.

Reliability data were collected for at least 33% of designs during the data extraction process for the purposes of calculating inter-observer agreement (IOA) using point-by-point agreement: (number of agreements/total number of data points) multiplied by 100 (Ayres & Ledford, 2014). The agreement window was 2% (for outcomes on a percentage or rate scale), 1 integer (for counts), and exact agreement (for categorical outcomes). Average overall agreement between data extraction coders was 97% (range across designs: 78-100%). A data collector erroneously skipped two data points when extracting data from a single comparison in a design that resulted in the 78% agreement calculation. Disagreements were resolved via consensus.

Visual Analysis Procedures

Visual analysis of the level, trend, variability, immediacy, overlap, and consistency of data was conducted for each design based on the guidelines outlined in the *Single Case Analysis and Review Framework* (SCARF; Authors, 2016). These guidelines for visual analysis were selected because they provided a comprehensive evaluation of SCD quality in a previous review of the same body of literature evaluated in this review. The first author served as the primary analyst for visual analysis; a special education graduate student trained in visual analysis served as a reliability coder for at least 33% of designs and an SCD expert (third author) served as a disagreement mediator (see Synthesis Tools Part I). Coders reached 100% exact agreement on the presence/absence of a functional relation and had one disagreement on the 4-point rating for one outcome in one design, which was resolved via consensus.

Calculation Procedures using Extracted Data

Overlap measures and parametric effect sizes were calculated for each A-B comparison across all designs. For multiple baseline designs across participants (MBPs), indices were calculated separately for each tier. For A-B-A-B designs with two A-B comparisons, indices were calculated for each pair (e.g., A1, B1 and A2, B2) and averaged, yielding a single effect size estimate per case. Simple averages were calculated for IRD, SMD with a Hedges' *g* small sample correction, LRRd, and LRRi. For the remaining overlap measures, weighting schemes were used to calculate averages, with *m* representing the number of points in condition A and *n* representing the number of points in condition B. Each index was weighted by *n* when calculating PND and by the total number of pairs across conditions (*m* x *n*) for Tau.

In studies using ATDs, indices were calculated for each comparison with the ongoing baseline condition denoted as A and the intervention condition denoted as B. Only comparisons of ongoing baseline versus the treatment were included in meta-analysis calculations; comparisons of weighted as compared to unweighted vests were not included, as neither condition reflected ‘baseline’ condition in which the intervention was absent.

For each design, overlap measures and parametric effect size estimates were calculated using the R statistical computing environment (R Core Team, 2016). Complete data and computer syntax for replicating all of the calculations is available on the Open Science Framework at (website redacted for peer review).

The second author served as the primary analyst for all overlap measure and effect size calculations; the first author served as the reliability analyst. The first author used a beta version of an online SCD effect size calculator tool to calculate reliability data for the overlap measures and parametric effect sizes (Pustejovsky, 2016d). IOA was calculated using point-by-point agreement to four decimal places (number of agreements/total number of data points) multiplied by 100 (Ayres & Ledford, 2014). Average agreement was 100% for PND, IRD, Tau, SMD, LRRd, and LRRi estimates. A coding error yielded an initial overall agreement of 0% for Tau, SMD, LRRd/LRRi standard error estimates. After correcting the error, agreement for Tau, SMD, LRRd, and LRRi standard error estimates was 100%.

Overlap Metrics

PND. PND quantifies overlap in data from two conditions (e.g., baseline and treatment; Scruggs et al., 1987; Scruggs & Mastropieri, 1998). For outcomes in which an increase is desirable, PND is defined as the percentage of data points in the second condition (treatment) that exceed the highest data point from the first condition; for outcomes in which a decrease is desirable, it is the percentage of data points in the second condition that are below the lowest data point from the first condition (Scruggs et al., 1987). PND is bound to values between 0-100% and is calculated for each A-B comparison in a single design. Scruggs and Mastropieri (1998) suggested PND values less than 50% denote an ineffective intervention, values between 50-70% denote unclear effects, values between 70-90% denote an intervention is effective, and values above 90% denote an intervention is very effective. No standard errors are available.

IRD. IRD (Parker et al., 2009) is a non-overlap measure defined as the number of “improved data points” divided by the total number of data points in a condition in each A-B comparison (Parker et al., 2009, p. 139). For outcomes in which an increase is desirable, an improved data point is one in the intervention condition that meets or exceeds any data point in the baseline condition. For outcomes in which a decrease is desirable, an improved data point is one in the intervention condition that meets or is less than any data point in the baseline condition. IRD can be expressed as a percentage or as a positive or negative number, but the range of possible values depends on the number of data points (Pustejovsky, 2016b). A score of 1.0 IRD indicates all data in the intervention condition are higher than scores in the baseline condition (Parker et al., 2009). As tentative benchmarks for IRD, Parker et al. (2009, p. 147) proposed that values below .50 correspond to “questionable” effects, values between .50 and .70 correspond to “medium” effects, and values above .70 correspond to “large” effects. No standard errors are available.

Tau. Kendall’s tau for non-overlap between groups ($\text{Tau}_{\text{novlap}}$ or Tau; Parker et al., 2011) was calculated rather than Tau-U because the calculation for Tau-U is unclear and has been noted to change across papers written by the creators of the measure (Pustejovsky, 2016c; Tarlow, 2016). Tau is a non-overlap measure defined in terms of comparisons between pairs of data points, including one data point from a baseline condition and one data point from a treatment condition. Tau is an analysis of pairwise comparisons in which each pair of data points is coded as "(a) positive or improving over time, (b) negative or decreasing, or (c) tied" (Parker et al., 2011, p. 288). The difference between the number of positive and negative comparisons is divided by the total number of pairs. Scores further from zero indicate fewer overlapping data points. Standard errors for Tau were calculated using a formula that is valid when the outcomes are mutually independent (Pustejovsky, 2016a); in the presence of positive auto-correlation, the formula will tend to under-state the true standard error. Tau is related to another overlap metric, non-overlap of all pairs (NAP; Parker & Vannest, 2009), by a simple linear transformation. Results for Tau therefore apply directly to NAP as well. Furthermore, due to the direct correspondence between the metrics, benchmark values for NAP can be used to derive benchmarks for Tau. Based on guidelines for NAP proposed by Parker and Vannest (2009), Tau values of less than .30 might be characterized as "weak," values between .30 and .84 as "medium," and values larger than .84 as "large."

Within-Case Effect Sizes

Within-Case SMD. A within-case SMD with a Hedges' *g* small sample correction was calculated. For a single A-B comparison, the within-case SMD is calculated as the difference between the average outcomes in the B phase versus the A phase, scaled by the standard deviation of the A phase; a small-sample correction is then applied to reduce bias from short A phases (Pustejovsky & Ferron, 2017). Because it involves standardizing by a measure of variability in the baseline condition, it cannot be calculated if the outcome series is constant in the baseline condition. Consequently, several data series with outcomes of zero throughout the baseline condition were excluded when calculating within-case SMDs. Standard errors were calculated based on the assumption that the outcomes are mutually independent (Pustejovsky & Ferron, 2017). It is important to note that this will tend to under-state the true standard error if there is positive auto-correlation. Harrington and Velicer (2015) proposed values between 0 and 1 as “small,” 1 and 2.5 as “medium,” and greater than 2.5 as “large.”

Within-Case LRRd and LRRi. LRR (Pustejovsky, 2014) effect size indices provide quantitative measures of the proportional change in behavior between conditions (Pustejovsky & Ferron, 2017). For outcomes measured as counts or rates, the two forms of the index, LRRd and LRRi, differ only in sign. However, for outcomes measured as proportions or percentages—which comprised the majority of outcomes in identified studies—LRRd and LRRi can differ not only in sign but also in magnitude, making it necessary to distinguish between them (Pustejovsky, 2017). For a single A-B comparison and an outcome where an increase is desirable, LRRi is calculated by taking the natural logarithm of the ratio of the average outcome in the B phase to the average outcome in the A phase. The natural logarithm transformation is used so that the range of possible values is less restricted (Pustejovsky, 2017). LRR indices are appropriate for dependent variables measured using a ratio scale, where a score of zero indicates the absence of a measured outcome (e.g., percentage correct responding; Pustejovsky & Ferron, 2017). The indices cannot be calculated if the mean in one or both conditions is equal to zero. Because of this requirement, analysis of LRRd and LRRi excluded several data series.

After LRR values are synthesized, the average effect size estimates can be translated into a percentage change for purposes of interpretation (Pustejovsky & Ferron, 2017). LRR decreasing (LRRd) was developed to evaluate interventions designed to decrease behavior occurrence whereas LRR increasing (LRRi) was developed to evaluate interventions designed to increase behavior occurrence over time; both can be used if interventions evaluate increasing and decreasing behavior (Pustejovsky & Ferron, 2017). As with Tau and the within-case SMD, standard errors for LRRd and LRRi were calculated based on the assumption that the outcome data were mutually independent; limitations in the presence of auto-correlation are similar. There is not yet a known scale to quantify values.

Meta-analysis

Overall average effect sizes were calculated for the indices with known standard errors, including one overlap measure (Tau) and the three effect sizes (SMD, LRRd, and LRRi). Average effects were estimated based on a multi-level meta-analysis model including between-study and within-study variation in true effect sizes. The overall meta-analytic model was:

$$T_{ij} = \mu + \eta_j + \epsilon_{ij} + \nu_{ij}$$

where T_{ij} is the effect size estimate for case i in study j; ν_{ij} is the sampling variance of T_{ij} (i.e., the squared standard error), which is treated as known; μ is the overall average effect size; η_j is a random effect for study j; and $\epsilon_{ij}$is a random effect for case i in study j. Study-specific random effects were assumed to be normally distributed with mean zero and standard deviation τ; case-specific random effects were assumed to be normally distributed with mean zero and standard deviation ω. The model was estimated using restricted maximum likelihood. Standard errors and confidence intervals for overall average effect sizes were calculated using robust variance estimation (clustering by study, with small sample adjustments; Hedges, Tipton, & Johnson, 2010; Tipton, 2015; Tipton & Pustejovsky, 2015) to account for the possibility of inaccurate sampling variances, as could occur if the outcomes were auto-correlated. Calculations were conducted in R using the *metafor* (Viechtbauer, 2010) and *clubSandwich* (Pustejovsky, 2016e) packages. Subgroup analyses of effects by intervention and outcome type were conducted using the aforementioned model with separate intercepts for each category. For PND and IRD, which do not have known sampling variances, overall average effect sizes were calculated using simple, un-weighted averages. Robust variance estimation methods (clustering by study, with small-sample corrections) were again used to account for the unknown sampling variances of the estimates, as well as the possible dependence among effect sizes from a given study.

Results

Visual Analysis

Among the included designs, visual analysis using the SCARF (Authors, 2016) yielded zero designs with positive outcomes. In short, visual analysis indicated functional relations were not present between antecedent SBI and targeted behaviors.

Meta-Analysis

The sample distributions for each overlap measure and parametric effect size estimate can be found in Table 2. Outcomes were calculated for all possible comparisons that met the criteria for each measure (e.g., ratio scale for LRRd and LRRi). Individual estimates for each included design can be found in the supplementary materials.

Random effects analysis. Table 3 reports the results from the random effects meta-analysis, including estimates of between-study and within-study heterogeneity. An omnibus effect size was calculated using each measure, along with effect sizes by dependent and independent variables (see Table 3). Overall average effectiveness was not statistically significant for Tau, SMD, LRRd, or LRRi, with confidence intervals crossing zero for each measure (see Table 3). The estimated average magnitude of effect was small using Tau (0.281, 95% CI [-0.025, 0.587]) and SMD (0.340, 95% CI [-0.044, 0.725]). The magnitude of effects for LRRd and LRRi were interpreted as percentage change (Pustejovsky & Ferron, 2017). The average estimated effect using LRRd (-0.269) was equivalent to a -23% change (reduction) in behavior (95% CI [-50%, 19%]). The average estimated effect using LRRi (0.027) was equivalent to a 3% change (increase) in behavior (95% CI [-11%, 19%]).

When designs were grouped by whether dependent variables were desirable (intended to increase) or undesirable (intended to decrease), overall effect sizes for each group were not distinguishable from zero for Tau, SMD, LRRd, or LRRi. Intervention effectiveness also was evaluated with designs grouped by independent variable type: weighted vests (Cox et al., 2009; Hodgetts et al., 2011; Leew et al., 2010; Reichow et al., 2010), sensory seating (Krombach, 2016; Olson, 2015; Umeda & Deitz, 2011), and multi-sensory environment (Tunson & Candler, 2010). Tau and SMD produced statistically significant results for weighted vest interventions (Tau = 0.281, 95% CI [0.094, 0.468]; SMD 0.314, 95% CI [0.063, 0.565]). LRRd and LRRi found outcomes for weighted vests interventions were not distinguishable from zero. Outcomes were not significantly different from zero for sensory seating interventions across any measures. The statistical significance of results for multi-sensory environment interventions could not be assessed because only one study provided relevant data. LRRd and LRRi could not be calculated for multi-sensory environments because the dependent variable used in the study was categorical rather than ratio scale, and thus did not meet statistical assumptions.

Estimated between- and within-study variance components indicated a large amount of heterogeneity across cases, suggesting that interventions may have harmful effects for some participants and positive effects for others (Tau, SMD, LRRd; Table 3). Compared to the other effect size measures, there was less between-study heterogeneity in LRRi. Furthermore, SMD and LRRi had a smaller estimated degree of within-study heterogeneity than Tau or LRRd.

Simple un-weighted analysis. Overall average effect sizes were calculated using un-weighted averages and robust variance estimation methods for measures without known sampling variances (PND and IRD; see Table 4). PND indicated antecedent SBI resulted in small positive effects that were not statistically significant (average PND = 0.223, 95% CI [-0.005, 0.451]), whereas IRD indicated antecedent SBI produced moderate, positive and statistically significant effects (average IRD = 0.430, p<0.05, 95% CI [0.261, 0.598]). When designs were grouped by dependent or independent variable type, the overall effectiveness of each group was not distinguishable from zero using PND. However, subgroup analyses using IRD did yield positive results that were distinguishable from zero when designs were grouped by dependent and independent variable type following the same method as in the random effects analysis. Antecedent SBI produced moderate, positive changes in desirable (average IRD = .438, 95% CI [0.154, 0.721]) and undesirable (average IRD = 0.416, 95% CI [0.221, 0.610]) behaviors. IRD also yielded moderate positive effects for weighted vest interventions (average IRD = 0.408, 95% CI [0.383, 0.434]); outcomes were indistinguishable from zero for sensory seating interventions. Estimates could not be calculated for multi-sensory environments due to an insufficient sample size for robust variance estimation.

Comparisons Across Effect Size Measures

Overlap measures. Conclusions differed among overlap measures (Tables 3 and 4). PND (average PND = 0.223, 95% CI [-0.0005, 0.451]) and Tau (average Tau = 0.281, 95% CI [-0.025, 0.587]) resulted in non-significant overall effects for antecedent SBI improving child behaviors, whereas IRD (average IRD = 0.430, 95% CI [0.261, 0.598]) resulted in moderate positive outcomes. Subgroup analyses also yielded variable conclusions across measures. Using IRD, antecedent SBI resulted in moderate, positive outcomes for desirable (average IRD = 0.438, 95% CI [0.154, 0.721]) and undesirable (average IRD = 0.416, 95% CI [0.154, 0.721]) behaviors. PND and Tau did not result in significant outcomes across either behavior type. Tau (average Tau = 0.281, 95% CI [0.383, 0.434]) and IRD (average IRD = 0.408, 95% CI [0.383, 0.434]) yielded positive outcomes for weighted vest interventions whereas PND values were not distinguishable from zero. Sensory seating interventions did not result in significant positive outcomes using any overlap measure.

Parametric within-case effect sizes. The sample distributions for each parametric within-case effect size can be found in Table 2. Due to differing requirements across measures, outcomes were not calculated for all comparisons using each parametric effect size; thus, comparisons across measures do not include the same A-B comparisons. Values by comparison can be found in the supplemental materials. All three parametric effect sizes pointed towards the conclusion that antecedent SBI do not result in overall outcomes distinguishable from zero (Table 3). Evaluation of the confidence intervals (CI), however, suggests variability in possible outcomes for single studies. The CI for LRRd included many values indicating improvements (reduction) in behavior, thus the possibility of an average 50% improvement in behavior is not ruled out. LRRi has a tight CI around zero that rules out large effects (positive or negative).

Conclusions from subgroup analyses also varied across measures. All measures yielded outcomes not distinguishable from zero for all dependent variable subgroups and sensory seating interventions. SMD resulted in conclusions of significant, small positive outcomes for weighted vest interventions (average SMD = 0.314, 95% CI [0.063, 0.565]). With each of the parametric effect sizes, heterogeneity statistics suggested variability in outcomes were present across all subgroup analyses, although the LRRi appeared to be more homogeneous than other measures.

Discussion

The current review provided the first comparison of overlap measures and parametric within-case effect sizes as indices to quantify intervention effectiveness in the context of SCD meta-analyses. These results have important implications for the use of antecedent SBI *and* SCD outcome synthesis. Meta-analytic statistics calculated with five of the six measures (PND, Tau, SMD, LRRd, LRRi) and visual analyses concluded that antecedent SBI are not an effective intervention. These results are consistent with previous reviews in special education journals (Barton et al., 2015; Lang et al., 2012; Leong et al., 2015).

Implications for Researchers

Conclusions regarding the consistency and helpfulness of quantitative measures are important given the widespread use of varying measures and the recent emphasis of use of these measures by federal funding agencies (Shadish et al., 2015). Differences across measures can lead to differences in conclusions regarding intervention effectiveness. Meta-analyses conducted with PND, Tau, and within-case parametric effect sizes resulted in omnibus effect sizes that were not significantly different than zero. However, PND, LRRd, and LRRi were the only measures that detected non-significant results for the omnibus effect size *and* for subgroup analyses, suggesting these measures may be more conservative estimates of effectiveness that may be less likely to increase Type I error rates.

The applicability of the measures also varied across tools. PND could be calculated for more A-B comparisons than LRRd or LRRi, and may have been particularly applicable to this body of literature because visual analysis concluded most designs had no effect (Rakap, Snyder, & Pasia, 2014; Wolery et al., 2010). However, previous studies have found overlap measures may fail to detect a range of effects present in a body of literature (Rakap, Snyder, & Pasia, 2014; Ledford, Wolery, & Gast, 2014; Wolery et al., 2010; Parker et al., 2011a). Furthermore, methodological research indicates that the magnitude of PND—and particularly it's magnitude when the true effect is null—is affected by the number of observations in the baseline condition

(Allison & Gorman, 1994; Pustejovsky, 2016b). Similarly, the value of IRD depends on the number of observations in the baseline and treatment conditions (Pustejovsky, 2016b). These properties make interpretation of PND and IRD results difficult, particularly considering that baseline condition lengths ranged from 4 to 20 observations across included A-B comparisons.

Application of the LRRd and LRRi effect sizes was limited to dependent variables measured using a ratio scale and where the average level of the outcome was greater than zero during the baseline condition (Pustejovsky & Ferron, 2017). Although this resulted in the exclusion of some A-B comparisons, most of the exclusions were for comparisons that were not relevant to the ultimate synthesis (i.e., comparisons between no treatment and unweighted vests).

Recommendations for Synthesis Method Selection

Selecting an appropriate quantitative measure to use should occur *in conjunction* with outcome evaluation using visual analysis (Authors, in press). Meta-analyses conducted with PND and IRD are difficult to interpret since the null values of each index are not zero. The values produced by these overlap measures are also influenced by procedural variations in designs (see Pustejovsky, 2016b; Authors, in press), thus decreasing the confidence one may have that outcome values are a reflection of behavioral changes rather than procedural features (e.g., number of data points in a condition). Given this limitation, coupled with numerous well-established limitations of overlap measures in extant literature (Chen et al., 2010; Wolery et al., 2010), we do not recommend the use of PND or IRD in meta-analytic summaries of SCD interventions as an estimate of *magnitude of effect*.

A third non-overlap measure, Tau, does not have the same shortcomings in terms of sample-size dependence. However, simulation evidence suggests that its magnitude can be influenced by other procedural features such as the choice of observational recording system and length of observation session for behavior outcomes. Methods for estimating the standard error of Tau are also available (Pustejovsky & Ferron, 2017), but these rely on the assumption that the outcomes are not auto-correlated. We therefore recommend considering use of Tau, or the directly equivalent NAP measure, instead of other overlap metrics. In particular, Tau may be useful for SCD that use outcome measurements that do not satisfy the assumptions necessary for parametric within-case effect size indices (i.e., ratio scale measurements).

Parametric within-case effect size selection should be considered in conjunction with guidelines presented by Pustejovsky and Ferron (2017) such as design type, dependent variable, and observational measurement system. If dependent variables are not measured using a ratio scale (e.g., percentage change), LRRd and LRRi cannot be calculated. If outcomes are measured using a ratio scale, LRRd and/or LRRi may be the most useful magnitude estimators. Results of LRRd and LRRi can be compared to determine which measure produces more homogeneous results (LRRi in this review) and outcomes can be reported using that measure. For comparisons included in this review, LRRi was applicable to more designs than SMD or LRRd, and was least likely to potentially over-estimate positive effects. However, due to the limited applications of LRRi, it should be used cautiously until further studies examine LRRi in the context of other types of interventions, dependent variables, and SCD designs (e.g., multiple baseline across contexts, when trends are present in data).

Limitations and Future Research

As noted in Synthesis Tools Part I, the sample size of the present review is relatively small and results must therefore be considered in the context of some limitations. First, the A-B comparisons included in each meta-analysis varied across tools due to the applicative limitations of each parametric effect size (Pustejovsky & Ferron, 2017). Variability in outcomes between methods may be due in part to the inclusion of different A-B comparisons in addition to differences in the measures themselves. Additionally, decisions regarding parametric within-case effect size selection and meta-analysis calculations were conducted by a statistician with expertise in SCD meta-analysis. Although guidelines for conducting SCD meta-analyses exist (cf. Pustejovsky & Ferron, 2017), the feasibility of completing an SCD meta-analysis without the assistance of an expert in the procedures cannot be determined.

Interpretability of outcomes may also be limited when discussing overlap measures and within-case parametric effect sizes because results of SCD are typically visually analyzed and discussed by design rather than each adjacent comparison. Both classes (overlap measures and parametric effect sizes) were calculated using simple A-B comparisons. Thus, they should be interpreted as estimates of magnitude of behavior change between baseline and intervention conditions. Because data patterns and characteristics were *not* considered with respect to determining whether experimental control was demonstrated, neither overlap measures nor parametric effect sizes should be interpreted as indicative of absence or presence of a functional relation. For example, an A-B-A-B design with mean values of 10, 20, 20, and 40 (per condition, respectively) might indicate large effects with both types of measures due to differences in level between A and B conditions. However, given the likelihood of maturation and lack of behavior change between intervention and the second baseline condition, visual analysis would suggest a functional relation is not present. Additionally, interpretability may be impacted by reporting outcomes by each simple A-B comparison rather than design. Calculating and reporting aggregated overlap measures and within-case parametric effect sizes by design in future reviews may assist consumers' interpretation of results. Future research is needed to investigate how methods of aggregating A-B comparisons to design-level statistics (e.g., weighted or unweighted averages) may impact conclusions regarding outcome magnitude or intervention effectiveness. Future research is also needed to investigate how to aggregate results across studies with and without experimental control.

Finally, antecedent SBI included few positive outcomes as determined via visual analysis. Comparisons of synthesis tools may produce different results for bodies of literature with predominantly positive effects or bodies of literature with greater heterogeneity of effects across individuals. The homogeneous outcomes in this body of literature may limit the generalizability of conclusions regarding the agreement between synthesis methods. The lack of positive outcomes also could have prevented us from detecting potential under-estimation of effects. While results of this review may be limited to bodies of literature that have few positive outcomes, our conclusions may be better understood as additional reviews compare types of synthesis methods across intervention literatures with diverse outcomes. Specifically, future research should compare synthesis methods using bodies of literature with mostly positive outcomes (e.g., differential reinforcement) and mixed outcomes (e.g., social stories) to develop broadly relevant guidelines for selecting an appropriate synthesis method.

Conclusions from this review suggest evaluations of the effectiveness of interventions examined in the context of SCD can be influenced by the synthesis method selected. Recommendations for methods to evaluate design quality (Synthesis Tools Part I) and outcomes (Synthesis Tools Part II), in relation to the methods evaluated in this review, may also assist researchers in selecting an appropriate method to synthesize SCD when evaluating interventions to identify evidence-based practices.

References

References marked with an asterisk (*) indicate studies included in the review.

Allison, D. B., & Gorman, B. S. (1994). “Make things as simple as possible, but no simpler.” A rejoinder to Scruggs and Mastropieri. *Behaviour Research and Therapy, 32,* 885-890.

Authors (2016, April). Single case analysis and review framework (SCARF). Retrieved from (website redacted for blind peer review).

Authors (in press). Synthesis and meta-analysis of single case research. In J. R. Ledford & D. L. Gast (Eds.), *Single Case Research Methodology,* New York, New York: Routledge.

Ayres, A. J. (1979). *Sensory integration and the child*. Los Angeles, CA: Western Psychological Services.

Ayres, K., & Ledford, J. R. (2014). Dependent measures and measurement systems. In D. L. Gast & J. R. Ledford (Eds.), *Single case research methodology: Applications in special education and behavioral sciences* (pp. 124-153). New York, NY: Routledge.

Barton, E. E., Meadan, H., Fettig, A., & Pokorski, B. (2016, February). Evaluating and comparing visual analysis procedures to non-overlap indices using the parent implemented functional assessment based intervention research. Poster presented at the Conference on Research Innovations in Early Intervention, San Diego, CA.

Barton, E. E., Reichow, B., Schnitz, A., Smith, I. C., & Sherlock, D. (2015). A systematic review of sensory-based treatments for children with disabilities. *Research in Developmental Disabilities, 37,* 64-80.

Ben-Sasson, A., Hen, L., Fluss, R., Cermak, S. A., Engel-Yeger, B., & Gal, E. (2009). A meta-analysis of sensory modulation symptoms in individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders, 39,* 1–11.

Beretvas, S. N. (2010). Meta-Analysis. In G. R. Hancock & R. O. Mueller (Eds.), *The Reviewer's Guide to Quantitative Methods in the Social Sciences*, (pp. 255-263). New York, New York: Routledge.

Blanche, E. I., Chang, M. C., Gutierrez, J., & Gunter, J. S. (2016). Effectiveness of a sensory-enriched early intervention group program for children with developmental disabilities. *American Journal of Occupational Therapy*, *70*, 1-8.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. West Sussex, England: John Wiley & Sons.

Burns, M. K., Zaslofsky, A. F., Kanive, R., & Parker, D. C. (2012). Meta-analysis of incremental rehearsal using phi coefficients to compare single-case and group designs. *Journal of Behavioral Education, 21*, 185-202.

Busk, P., & Serlin, R. (1992). Meta-analysis for single-case research. In T. Kratochwill & J. Levin (Eds.), *Single-Case Research Design and Analysis: New Directions for Psychology and Education* (pp. 187–212). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Campbell, J. M. (2003). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: A quantitative synthesis of single-subject research. *Research in Developmental Disabilities, 24*, 120–138.

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.

Case-Smith, J., Weaver, L. & Fristad, M. (2015). A systematic review of sensory processing interventions for children with autism spectrum disorders. *Autism, 19*, 133-148.

Chen, M., Hyppa-Martin, J. K., Reichle, J. E., & Symons, F. J. (2016). Comparing single case design overlap-based effect metrics from studies examining speech generating device interventions. *American Journal on Intellectual and Developmental Disabilities*, *121*, 169-193.

Cohen. J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.) Hillsdale, NJ Erlbaum Associates.

*Cox, A., Gast, D., Luscre, D., & Ayres, K. (2009). The effects of weighted vests on appropriate in-seat behaviors of elementary-age students with autism and severe to profound intellectual disabilities. *Focus on Autism and Other Developmental Disabilities, 24*, 17-26.

Dart, E. H., Collins, T. A., Klingbeil, D. A., & McKinley, L. E. (2014). Peer management interventions: A meta-analytic review of single-case research. *School Psychology Review, 43,* 367-384.

Gingerich, W. J. (1984). Meta-analysis of applied time-series data. *The Journal of Applied Behavioral Science, 20*, 71-79.

Harrington, M. & Velicer, W. F. (2015). Comparing visual and statistical analysis in single-case studies using published studies. *Multivariate Behavioral Research, 50*, 162-183.

Hedges, L. V, Pustejovsky, J. E., & Shadish, W. R. (2012). A standardized mean difference effect size for single case designs. *Research Synthesis Methods, 3,* 224-239.

Hedges, L. V, Tipton, E., & Johnson, M. C. (2010). Robust variance estimation in meta-regression with dependent effect size estimates. *Research Synthesis Methods, 1*, 39-65.

Heyvaert, M. Wendt, O., Van den Noortgate, M., & Onghena, P. (2015). Randomization and data-analysis items in quality standards for single-case experimental studies. *The Journal of Special Education, 49*, 146-156.

*Hodgetts, S., Magill-Evans, J., & Misiaszek, J. E. (2011). Weighted vests, stereotyped behaviors and arousal in children with autism. *Journal of Autism and Developmental Disorders, 41*, 805-14.

Individuals with Disabilities Education Act Pub. L. No. 108-446, 1400 Stat. (2004).

Kahng, S., Iwata, B. A, & Lewin, A. B. (2002). Behavioral treatment of self-injury, 1964 to 2000. *American Journal of Mental Retardation : AJMR*, *107*, 212-221.

*Krombach, P. A. (2016). *The effects of stability ball seating on children with autism spectrum disorder.* Available from ProQuest Dissertations & Theses Global. (1781592012).

Lang, R., O’Reilly, M., Healy, O., Rispoli, M., Lydon, H., Streusand, W., Davis, T., Kang, S., …Sigafoos, J. (2012). Sensory integration therapy for autism spectrum disorders: A systematic review. *Research in Autism Spectrum Disorders, 6*, 1004-1018.

Ledford, J. R., Lane, J. D., & Severini, K. E. (2017). Systematic use of visual analysis for assessing outcomes. *Manuscript in preparation.*

Ledford, J. R., Lane, J. D., Zimmerman, K. N., & Shepley, C. (2016, February). *Bigger, better, & more complex: To what extent do newer overlap-based metrics adequately describe single case data?* Poster presented at the Conference on Research Innovations in Early Intervention. San Diego, CA.

Ledford, J. R., Wolery, M., & Gast, D. L. (2014). Controversial and critical issues in single case research. In D. L. Gast & J. R. Ledford (Eds.), *Single Case Research Methodology,* (pp. 377-396). New York, New York: Routledge.

*Leew, S. V., Stein, N. G., & Gibbard, W. B. (2010). Weighted vests' effect on social attention for toddlers with autism spectrum disorders. *The Canadian Journal of Occupational Therapy, 77*, 113-24.

Leong, H. M., Carter, M., & Stephenson, J. (2015). Systematic review of sensory integration therapy for individuals with disabilities: Single case design studies. *Research in Developmental Disabilities, 47*, 334-351.

Ma, H. H. (2006). An alternative method for quantitative synthesis of single-subject researches: Percentage of data points exceeding the median. *Behavior Modification, 30*, 598-617.

Maggin, D. M., Briesch, A. M., Chafouleas, S. M., Ferguson, T. D., & Clark, C. (2014). A comparison of rubrics for identifying empirically supported practices with single-case research. *Journal of Behavioral Education, 23*, 287-311.

Maggin, D. M., O'Keefe, B. V., & Johnson, A. H. (2011). A quantitative synthesis of methodology in meta-analysis of single-subject research for students with disabilities: 1985-2009. *Exceptionality, 19*, 109-135.

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.

May-Benson, T. A., & Koomar, J. A. (2010). Systematic review of the research evidence examining the effectiveness of interventions using a sensory integrative approach for children. *American Journal of Occupational Therapy, 64,* 403-414.

O'Keefe, B. V., Slocum, T. A., Burlingame, C., Snyder, K., & Bundock, K. (2012). Comparing results of systematic reviews: Parallel reviews of research on repeated reading. *Education and Treatment of Children, 35*, 333-366.

*Olson, N. A. (2015). *Investigating stability balls in the classroom: Effects on student behavior* *and academic productivity.* Available from ProQuest Dissertations & Theses Global. (1686537353).

Parker, R. I., Hagan-Burke, S., & Vannest, S. (2007). Percentage of all nonoverlapping data (PAND): An alternative to PND. *Journal of Special Education, 40*, 194-204.

Parker, R. I., & Vannest, K. J. (2009). An improved effect size for single case research: Nonoverlap of all pairs (NAP). *Behavior Therapy, 40*, 357-367.

Parker, R. I., Vannest, K. J., & Brown, L. (2009). The improvement rate differences for single-case research. *Exceptional Children, 75*, 133-150.

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011a). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification, 35*, 303-322.

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011b). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, *42*, 284-299.

Plot Digitizer (2015). Retrieved from http://plotdigitizer.sourceforge.net.

Pustejovsky, J. E. (2016a). Standard errors and confidence intervals for NAP. Retrieved from: http://jepusto.github.io/NAP-SEs-and-Cis.

Pustejovsky, J. E. (2016b). Procedural sensitivities of effect sizes for single-case designs with behavioral outcome measures. Retrieved from: http://jepusto.github.io/working_papers/.

Pustejovsky, J. E. (2016c). Tau-U. Retrieved from http://jepusto.github.io/Tau-U.

Pustejovsky, J. E. (2016d). SCD-effect-sizes: A web application for calculating effect size indices for single-case designs (Version 0.1) [Web application]. Retrieved from: https://jepusto.shinyapps.io/SCD-effect-sizes

Pustejovsky, J. E. (2016e). clubSandwich: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections. R package version 0.2.1. https://cran.r-project.org/package=clubSandwich

Pustejovsky, J. E. (2014). Measurement-comparable effect sizes for single-case studies of free-operant behavior. *Psychological Methods, 20*, 342-359.

Pustejovsky, J. E. (2017). Using response ratios for meta-analyzing single-case designs with behavioral outcomes. http://doi.org/10.17605/OSF.IO/RX5WF

Pustejovsky, J. E., & Ferron, J. M. (2017). Research synthesis and meta-analysis of single-case designs. In J. M. Kauffman, D. P. Hallahan, & P. C. Pullen (Eds.), *Handbook of Special Education*, 2nd Edition. New York, NY: Routledge.

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*, 368-393.

Pustejovsky, J., & Swan, D. (2015). Four methods for analyzing partial interval recording data, with application to single-case research. *Multivariate Behavioral Research, 50*, 365-380.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Rakap, S., Snyder, P., & Pasia, C. (2014). Comparison of nonoverlap methods for identifying treatment effect in single-subject experimental research. *Behavioral Disorders, 39*, 128-145.

*Reichow, B., Barton, E. E., Sewell, J. N., Good, L., & Wolery, M. (2010). Effects of weighted vests on the engagement of children with developmental delays and autism. *Focus on Autism and Other Developmental Disabilities, 25*, 3-11.

Scruggs, T. E., Mastropieri, M. A., & Castro, G. (1987). The quantitative synthesis of single-subject research: Methodology and validation. *Remedial and Special Education, 8*, 24-33.

Shadish, W. R., Hedges, L. V., Horner, R. H., & Odom, S. L. (2015). The role of between-case effect sizes in conducting, interpreting, and summarizing single-case research. (NCER 2015-002) Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

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*, 123-147.

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation,113*, 95-109.

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, *2*, 188-196.

Tarlow, K. R. (2016). An improved rank correlation effect size statistic for single-case designs: Baseline corrected tau. *Behavior Modification*. *Advanced online publication.*

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta- regression. *Psychological Methods*, *20*, 375–393.

Tipton, E., & Pustejovsky, J. E. (2015). Small-sample adjustments for tests of moderators and model fit using robust variance estimation in meta-regression. *Journal of Educational and Behavioral Statistics*, *40*, 604-634.

*Tunson, J., & Candler, C. (2010). Behavioral states of children with severe disabilities in the multisensory environment. *Physical & Occupational Therapy in Pediatrics, 30*, 101-110.

*Umeda, C., & Deitz, J. (2011). Effects of therapy cushions on classroom behaviors of children with autism spectrum disorder. *American Journal of Occupational Therapy*, *65*, 152-9.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 2,* 142-151.

Vannest, K. J., & Ninci, J. (2015). Evaluating intervention effects in single-case designs. *Journal of Counseling & Development, 93,* 403-411.

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

Watling, R. & Hauer, S. (2015). Effectiveness of Ayres Sensory Integration® and sensory-based interventions for people with autism spectrum disorder: A systematic review. *American Journal of Occupational Therapy, 69,* 1-8.

Wendt, O., & Miller, B. (2012). Quality appraisal of single-subject experimental designs: An overview and comparison of different appraisal tools. *Education and Treatment of Children, 35*, 235-268.

Wolery, M. (2013). A commentary: Single-case design technical document of the what works clearinghouse. *Remedial and Special Education, 43,* 39-43.

Wolery, M., Busick, M., Reichow, B., & Barton, E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *Journal of Special Education, 44*, 18-28.

Yunus, F. W., Liu, K. P. Y., Bissett, M., & Penkala, S. (2015). Sensory-based intervention for children with behavioral problems: A systematic review. *Journal of Autism and Developmental Disorders, 45,* 3565-3579.

Table 1

Summary of Included Studies

	Year	Design	Intervention	Cases	Outcomes
Cox	2009	ATD	WV	2	1
Hodgetts	2011	ABAB	WV	3	1
Krombach	2016	MBP	SS	4	2
Leew	2010	MBP	WV	4	2
Olson	2015	ABAB	SS	1	5
Reichow	2010	ATD	WV	3	3
Tunson	2010	ABAB	MSE	2	1
Umeda	2011	ABAB	SS	2	2

Note. Articles are indicated by first author. ATD=alternating treatments design. MBP=multiple baseline across participants design. WV=weighted vest. SS=sensory seating. MSE=multi-sensory environment. Cases=total number of cases included in meta-analytic summaries. Outcomes=total number of outcome variables included in meta-analytic summaries.

Table 2

Sample Distribution Summary Statistics

ES	n	Min	Q1	Q2	Q3	Max	Mean
PND	63	0.00	0.00	0.04	0.33	1.00	0.22
IRD	63	0.00	0.19	0.33	0.60	1.00	0.41
Tau	63	-1.00	-0.25	0.16	0.55	1.00	0.15
SMD	55	-4.02	-0.15	0.30	0.89	6.78	0.31
LRRd	49	-4.23	-0.68	-0.07	0.16	1.82	-0.30
LRRi	60	-0.72	-0.06	0.01	0.13	1.00	0.05

Note. ES=effect size. n=number of included cases. Min=minimum. Q=quartile. Max=maximum. PND=percentage of non-overlapping data. IRD=improvement rate difference. SMD=standardized mean difference. LRRd=log response ratio decreasing. LRRi=log response ratio increasing.

Table 3

Random Effects Meta-Analysis of Indices with Known Sampling Variances

Category	n	Est.	SE	DF	95% CI	Between SD	Within SD
Tau							
Overall	**41**	**0.281**	**0.128**	**6.7**	**[-0.025, 0.587]**	**0.279**	**0.301**
DV Type						0.311	0.289
Desirable	26	0.191	0.153	6.5	[-0.176, 0.559]		
Undesirable	15	0.434	0.196	4.5	[-0.085, 0.953]		
IV Type						0.339	0.303
Weighted Vest	22	0.281	0.058	2.9	[0.094, 0.468]		
Sensory Seating	17	0.294	0.370	1.9	[-1.377, 1.964]		
MSE	2	0.162	-	-	-		
SMD							
Overall	**38**	**0.340**	**0.161**	**6.7**	**[-0.044, 0.725]**	**0.377**	**0.000**
DV Type						0.407	0.000
Desirable	25	0.259	0.211	6.2	[-0.252, 0.770]		
Undesirable	13	0.508	0.244	3.4	[-0.220, 1.236]		
IV Type						0.445	0.000
Weighted Vest	19	0.314	0.077	2.9	[0.063, 0.565]		
Sensory Seating	17	0.410	0.467	2.0	[-1.616, 2.436]		
MSE	2	0.198	-	-	-		
LRRd							
Overall	**36**	**-0.269**	**0.181**	**6.0**	**[-0.711, 0.173]**	**0.417**	**0.241**
DV Type						0.440	0.224
Desirable	23	-0.212	0.217	5.7	[-0.750, 0.327]		
Undesirable	13	-0.429	0.389	4.4	[-1.472, 0.613]		
IV Type						0.439	0.244
Weighted Vest	19	-0.175	0.063	3.0	[-0.375, 0.024]		
Sensory Seating	17	-0.385	0.453	2.0	[-2.343, 1.573]		
LRRi							
Overall	**38**	**0.027**	**0.060**	**5.6**	**[-0.122, 0.177]**	**0.132**	**0.076**
DV Type						0.128	0.076
Desirable	23	0.052	0.061	5.5	[-0.100, 0.205]		
Undesirable	15	0.000	0.070	5.0	[-0.179, 0.180]		
IV Type						0.142	0.078
Weighted Vest	21	-0.023	0.065	2.8	[-2.40, 0.194]		
Sensory Seating	17	0.079	0.118	2.0	[-0.431, 0.589]		

Note. n=total number of cases. SE=standard error. DF=degrees of freedom. CI=confidence interval. Tau=Kendall's tau for non-overlap between groups. SMD=between-case standardized mean difference with a Hedges' *g* small sample correction. LRRd=log response ratio decreasing. LRRi=log response ratio increasing. DV=dependent variable. IV=independent variable. MSE=multi-sensory environment. - = computation unavailable because only one article is included.

Table 4

Simple Un-Weighted Averages for Indices without Known Sampling Variances

Category	n	Estimate	SE	DF	95% CI
PND					
Overall	**41**	**0.223**	**0.090**	**5.3**	**[-0.005, 0.451]**
DV Type					
Desirable	26	0.257	0.129	4.6	[-0.082, 0.595]
Undesirable	15	0.164	0.088	2.5	[-0.148, 0.476]
IV Type					
Weighted Vest	22	0.143	0.074	2.1	[-0.159, 0.446]
Sensory Seating	17	0.352	0.167	1.8	[-0.433, 1.136]
MSE	2	-0.000	-	-	-
IRD					
Overall	**41**	**0.430**	**0.067**	**5.3**	**[0.261, 0.598]**
DV Type					
Desirable	26	0.438	0.108	4.6	[0.154, 0.721]
Undesirable	15	0.416	0.055	2.5	[0.221, 0.610]
IV Type					
Weighted Vest	22	0.408	0.006	2.1	[0.383, 0.434]
Sensory Seating	17	0.486	0.168	1.8	[-0.304, 1.276]
MSE	2	0.187	-	-	-

Note. n=total number of cases. SE=standard error. DF=degrees of freedom. CI=confidence interval PND=percentage of non-overlapping data. IRD=improvement rate difference. DV=dependent variable. IV=independent variable. MSE=multi-sensory environment. - = computation unavailable because only one article is included.