Research Synthesis and Meta-Analysis of Single-Case Designs

James E. Pustejovsky
University of Texas at Austin

And

John M. Ferron
University of South Florida

This is not the version of record. Please cite this chapter as:

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

Research Synthesis and Meta-Analysis of Single-Case Designs

In some areas of special education, much of the research base consists of studies that use single-case designs (SCDs). Single-case research methods play a prominent role in clinical and applied intervention research—especially research on low-incidence disabilities—because SCDs can be conducted with relatively few participants, and in settings where other types of experimental designs are difficult or infeasible. For example, Wong and colleagues (2015) conducted a comprehensive review of focused intervention practices for children with autism. Of 456 studies identified in the review, 89% used SCDs and only 11% used between-groups research designs. Similarly, a systematic review of positive behavioral interventions for children with challenging behavior identified 62 single-case studies but only 1 between-subjects design (Conroy, Dunlap, Clarke, & Alter, 2005). Use of SCDs appears to have become somewhat more common over the past three decades, as indicated by frequency of publication in prominent Special Education journals (Hammond & Gast, 2010) and growth in citations of "multiple baseline," the most common type of SCD (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2013).

Given the size and breadth of the research base that makes use of the designs, as well as their prevalence in some research areas, there is a need for systematic reviews and syntheses of evidence from SCDs. Research synthesis is the process of integrating results from multiple empirical studies for purposes of drawing generalizations (Cooper, Hedges, & Valentine, 2009). Often, this integration of study results is conducted using the statistical techniques of meta-analysis. By systematically comparing and combining evidence across individual studies, careful syntheses can be used to draw more generalized inferences than would be warranted either from individual studies or from narrative literature reviews—particularly in areas where individual studies are small, heterogeneous, and potentially come to ambiguous or discrepant conclusions.

In the context of special education, syntheses and meta-analyses of SCDs can be important research endeavors for at least three reasons. First, syntheses can serve as a basis for establishing evidence-based practices and directing clinicians, educators, and caregivers towards more effective interventions. Across many fields where between-groups research (e.g., randomized clinical trials) predominates, research syntheses are now recognized as a crucial tool for guiding policy and decision-making. A number of organizations exist to facilitate the production and use of syntheses for these aims, including the Institute of Education Science's What Works Clearinghouse (WWC; http://ies.ed.gov/ncee/wwc/) in education, the Cochrane Collaboration (http://www.cochrane.org/) in health care, and the Campbell Collaboration (http://www.campbellcollaboration.org/) in social sciences, among others. Many such large-scale synthesis efforts initially ignored evidence from SCDs (Shadish & Rindskopf, 2007), with consequent limitations in the evidence base for areas such as early intervention and early childhood special education (Odom, Collet-Klingenberg, Rogers, & Hatton, 2010). However, over the past decade, several prominent research organizations have developed guidelines for establishing evidence-based practices on the basis of evidence from SCDs, including Divisions 12 and 16 of the American Psychological Association (Chambless & Hollon, 1998; Chambless & Ollendick, 2001; Kratochwill & Stoiber, 2002), the Council for Exceptional Children's Division for Research (Council for Exceptional Children Working Group, 2014; Horner et al., 2005), and the WWC (Kratochwill et al., 2013). These guidelines have been applied to identify effective intervention practices for children with autism (e.g., Wong et al., 2015), as well as to demonstrate a lack of supporting evidence for certain practices (e.g., Losinski, Sanders, & Wiseman, 2016).

In addition to identifying evidence-based practices, syntheses of SCDs can be used to study variation in a treatment's efficacy across sub-populations, settings, conditions, and procedures (Gingerich, 1984; D. M. White, Rusch, Kazdin, & Hartmann, 1989). For example, in a recent meta-analysis of SCDs on behavioral self-management interventions, Briesch and Briesch (2016) found that intervention effects were moderated by both student disability type and by use of specific intervention components. It would be difficult to establish such relationships without using meta-analysis because individual studies seldom contain sufficient variability across moderating characteristics, much less a sufficient number of cases to be able to distinguish systematic associations from chance.

A final, less widely recognized advantage is that research syntheses can contribute to improvement and innovation in the methodology of a discipline. In clinical medicine, such investigations have identified puzzling and troubling patterns in empirical findings, such as that effect sizes from initial trials tend to be inflated compared to later trials (e.g., Ioannidis, 2008). More relevant to special education, scholars have used research synthesis methods to evaluate the methodological characteristics of published SCDs against existing standards and best-practice recommendations (Shadish & Sullivan, 2011; Smith, 2012) and to examine the extent to which data from SCDs meet the assumptions required for conventional statistical analyses (Solomon, 2014). In projects such as these, using research synthesis provides a unique vantage point from which to assess the workings of an entire field.

In light of these potential advantages, there has long been interest in applying research synthesis methods to SCDs (Allison, Faith, & Franklin, 1995; Gingerich, 1984; Scruggs, Mastropieri, Cook, & Escobar, 1986). The production of systematic reviews and syntheses of single-case research has also increased considerably over the past decade (Maggin, O'Keeffe, & Johnson, 2011), likely spurred by the increasing emphasis on evidence-based practice. Simultaneously, methodological developments over the past decade have greatly expanded the range of statistical tools available for meta-analyzing SCDs (Shadish, 2014)—particularly in the areas of effect size metrics and meta-analysis methods. Although there is now a wide array of statistical techniques available, many of these tools are still under development and investigation, and there remains a lack of consensus about which tools are best suited for use with single-case data (Lane & Carter, 2012). As a result, a researcher interested in conducting a synthesis of single-case research is currently faced with what might seem to be a dizzying array of options and differences of methodological opinion.

The goal of this chapter is to survey some of these developments and provide guidance about how to conduct a synthesis of SCDs. We begin by providing an overview of the process of conducting a research synthesis and discussing the tasks and procedures involved in the initial stages of a synthesis project. Because these initial stages closely resemble the processes used for synthesizing between-groups research, we provide only a brief overview. In subsequent sections, we discuss available effect size metrics for quantifying the magnitude of functional relationships between interventions and outcomes, then discuss methods for synthesizing results across multiple cases and SCD studies. These later sections are more detailed (though still selective) because the topics involves technical methods that are more specialized to single-case research. Unfortunately, there remain a number of outstanding methodological questions about these stages of the process, which prevents us from offering definitive recommendations about how to proceed on certain fronts. Given the lack of methodological consensus, we conclude by highlighting areas in need of further research and suggesting some ways to proceed until the outstanding methodological issues are better resolved.

The Research Synthesis Process

The work involved in designing and conducting a research synthesis is in many respects analogous to the process of conducting a primary research study such as a survey, except that it involves sampling and collecting measurements from already-conducted studies (typically, as described in published or unpublished research reports) rather than from individual people (Cooper, 1982). The process begins by formulating research questions and making decisions about how to operationalize the constructs involved. Similar to how one would sample and screen participants for eligibility in a primary study, research synthesis involves systematically searching for and screening *studies* for inclusion. Then, rather than surveying or measuring participant characteristics, research synthesis involves coding the characteristics of identified studies and extracting data. Typically (though not always), extracted data are used to calculate effect sizes, which are indices that quantify the magnitude of the functional relationships observed in the study. In primary research, statistical analysis (e.g., analysis of variance, regression modeling, hypothesis testing) is used to draw and support inferential conclusions on the basis of sample data; in a research synthesis, inferential conclusions are often drawn using the techniques of meta-analysis, a specialized set of statistical tools for combining and analyzing information (in the form of effect sizes or raw data) from multiple studies. At every stage of the process, researchers aim to apply scientific methods by using systematic, clearly operationalized, and replicable procedures, with the goal of obtaining a comprehensive and unbiased view of the available evidence.

This section provides an overview of the initial steps involved in conducting a research synthesis. The issues and methods involved in these stages are mostly generic, and apply whether the synthesis focuses on single-case or between-groups research. We therefore keep the discussion brief, while highlighting areas of particular concern in syntheses of SCDs. Researchers planning to conduct a synthesis will find it helpful to also consult more detailed guides to the process. Cooper (2010) and Lipsey and Wilson (2001) provide book-length treatments focused on synthesis of social science research. The *Handbook of Research Synthesis and Meta-Analysis* (Cooper et al., 2009) is a comprehensive and authoritative source for learning about the methodological issues involved in research synthesis.

Formulating and operationalizing research questions

Synthesis projects begin by formulating a set of research questions that can be addressed by examining the results of studies that have already been conducted. Questions addressed through research synthesis vary in scope (Cooper, 2009), but syntheses of SCDs are often either intervention-focused or problem-focused. Intervention-focused syntheses pose questions about a particular class of interventions or practices and may examine effects across a range of different outcomes. For instance, Fowler, Konrad, and Walker (2007) examined the effects of self-determination interventions on academic outcomes for students with cognitive disabilities. In contrast, problem-focused syntheses focus on a broader range of interventions or practices for addressing a specified problem. For example, Heyvaert and colleagues (2014) examined the efficacy of behavioral interventions—ranging from antecedent exercise to mindfulness-based interventions to differential reinforcement of alternative behavior—for reducing problem behavior in individuals with autism. Regardless of whether the research questions center around an intervention or around a problem, most syntheses of SCDs also limit scope to a certain population of individuals (e.g., children with autism, students with emotional and behavioral disorders).

Once formulated, the research questions to be addressed in a synthesis will invoke constructs pertaining to the population of individuals, types of interventions, and outcome variables of interest. To address the research questions, one must formulate for each dimension a set of criteria that define whether a given study falls within the scope of relevant evidence. For example, Heyvaert and colleagues (2014) specified detailed criteria for determining whether a study's participants included individuals with autism and for determining whether the study assessed problem behavior as an outcome. Providing detailed operational definitions enables the research team to maintain objectivity when assessing studies for inclusion, as well as allowing the study's audience to judge the extent to which the included studies are relevant to addressing the research questions.

One further dimension that should also be addressed in developing inclusion criteria is the type of research design employed in a study. Syntheses of between-groups intervention research often limit consideration to research designs that are understood to have strong internal validity—that is, where it is reasonable to interpret the observed results as representing causal effects of the intervention on the outcome. Many syntheses of SCDs limit their scope only to studies that use SCDs (e.g., Heyvaert, et al., 2014), although others include both SCDs and well-designed between-groups studies (e.g., Yoder, Bottema-Beutel, Woynaroski, Chandrasekhar, & Sandbank, 2014). In our opinion, the latter approach is usually preferable because it yields a more complete picture of the evidence. To the extent that both types of research designs can provide internally valid evidence about functional relationships (i.e., causal effects), then there is little reason to limit consideration to only SCDs, at least on an a priori basis.

Syntheses of SCDs also often limit consideration to studies published in peer-reviewed sources (e.g., Briesch & Briesch, 2016; Fowler et al., 2007). This strategy is in marked contrast to typical practice in syntheses of between-groups research, where searching for unpublished research is recognized as critical for obtaining a representative view of the evidence (Rothstein, Sutton, & Borenstein, 2005). The main justification for focusing on peer-reviewed SCD studies is that doing so ensures that studies meet some standard of rigor. However, peer-review is only an indirect indicator of study quality and might itself induce new biases and distortions in a body of evidence (Sham & Smith, 2014). We therefore recommend that synthesis projects not use peer-review status as an inclusion criterion, but instead examine both published and unpublished research (e.g., Losinski et al., 2016).

Searching for and screening literature

Having formulated a set of research questions and articulated clear inclusion criteria, the next stage of the process is to search for and screen research reports for eligibility. The over-arching goal of literature search is to identify as comprehensively as possible the set of studies that fit inclusion criteria. Of course, in the interest of feasibility it is also desirable to limit the amount of irrelevant material that must be screened out.

White (2009) reviews five different modes of literature search that can be used to identify candidate studies: database searches, footnote chasing, consultation, browsing, and citation searches. These search modes commonly involve tasks such as:

- keyword searches in reference databases like PsychInfo®, the Educational Resources Information Center, and PubMed;
- examining the references of previous reviews on relevant topics;
- asking scholars who are active in the field to provide information about completed studies;

- examining the table of contents of subject-area journals;
- examining the references of studies already identified as meeting inclusion criteria; and
- searching for articles that cite included studies.

Research synthesis projects often use several—or even all—of these approaches to identify literature.

Of the five modes, reference database search is often the first and biggest source of candidate studies. In constructing searches, it is crucial to understand the coverage of each database and to use multiple databases in order to improve coverage and reduce the potential for bias in the set of retrieved literature (Reed & Baxter, 2009). The Campbell Collaboration provides useful guidance about developing search strategies for systematic reviews in the social sciences (Kugley et al., 2016).

Once a set of candidate studies has been identified, they must be screened to determine whether they meet the specified inclusion criteria. This time-consuming task is typically carried out in multiple stages, starting with review of titles and abstracts to screen out clearly ineligible studies. Full-text review of potentially eligible studies is then used to make final eligibility decisions. Throughout the process, multiple reviewers should be used to ensure reliability, and reasons for exclusion should be documented.

Extracting data and coding study characteristics

After identifying studies for inclusion in the synthesis, the next step is to extract results and code characteristics of the studies. The results of single-case studies are typically presented graphically in single-case diagrams. Raw data can be extracted reliably from these graphs (Moeyaert, Maggin, & Verkuilen, 2016; Shadish et al., 2009) and then used to calculate effect size indices or used for raw-data meta-analysis, as we describe in later sections. Study characteristics to extract might include information about the research setting, participants, intervention approach, methodological procedures, or even the research team who conducted the study. This information serves two purposes. First, it is used descriptively to characterize the body of research that is being synthesized—much as a primary study would always report descriptive statistics on the demographic characteristics of participants. Second, coded study characteristics become moderators in a meta-analysis of study results, potentially explaining variation in the observed effect sizes.

Wilson (2009) provides practical guidance on how to conduct systematic coding of studies, emphasizing the importance of developing a protocol. The process of coding single-case research is complicated by the hierarchical structure of the information to be encoded. Data to be extracted might describe study-level details, individual case characteristics (nested within the study), or session-level information (i.e., outcome measurements, phase in which the session falls, nested within cases). For dealing with information structured in hierarchical fashion, Wilson (2009) recommends using a relational database so as to avoid unnecessary repetition and ensure accuracy of data entry.

Assessing evidence quality

In any research synthesis, an important class of data to be extracted from identified studies is characteristics that relate to evidence quality—or more precisely, the extent to which a study provides internally valid evidence about the functional relationships of interest. Information about evidence quality is crucial because the validity of one's ultimate conclusions rests on the validity of the evidence that is synthesized. Judgments about evidence quality in SCDs involve dimensions that are distinct from evidence quality in between-groups designs (Odom et al., 2005). Consequently, research syntheses of SCDs should use a quality appraisal tool that is developed specifically for SCDs. Wendt and Miller (2012) identified and critically reviewed seven such tools, finding that they range in scope of application and rigor of their criteria. Two tools that have received considerable attention are the What Works Clearinghouse (WWC) Pilot Standards and the Council for Exceptional Children (CEC) Standards, both of which draw extensively on a conceptualization of evidence quality proposed by Horner and colleagues (Horner et al., 2005).

The WWC Pilot Standards (Kratochwill et al., 2013) were developed as inclusion criteria for meta-analyses conducted by the WWC. To satisfy the design standards without reservations, a study must meet the following criteria:

- the intervention must be systematically manipulated by the researcher;
- each outcome must be measured systematically over time by more than one assessor, with inter-assessor agreement exceeding a specified threshold;
- the design must include at least three attempts to demonstrate an effect; and
- phase designs (e.g., multiple-baseline, ABAB) must have a minimum of five data points per phase, while alternating treatment designs need at least five repetitions of the alternating sequence.

Only SCDs meeting all of the design standards are included the synthesis of evidence.

Wolery (2012) criticizes the WWC Pilot Standards for being overly influenced by concepts from between-groups research; for missing important dimensions of single-case research, such as evaluation of measurement procedures and assessment of procedural fidelity; and for not covering the full range of available single-case designs (e.g., adapted alternating treatment, parallel treatment designs). In a response to Wolery's critiques, the authors of the Pilot Standards emphasized the importance of recognizing that they were designed specifically for use as part of the WWC review process and that (as indicated by their “pilot” status) their development is ongoing (Hitchcock et al., 2014).

The Council for Exceptional Children has put forth standards for evidenced-based practices in special education (Council for Exceptional Children Working Group, 2014), which incorporate quality indicators for both between-groups designs and SCDs. Some quality indicators apply to both types of design, such as assessing and reporting implementation fidelity, whereas some indicators are design-specific. SCD-specific quality indicators include the design providing at least three demonstrations of the effect at at least three different times, the baseline phases having at least three data points (unless measuring dangerous or zero baseline behaviors), the design controlling typical threats to internal validity, and the study providing an appopriate graph of the outcome data.

Once collected, evidence quality information can be used in several different ways. First, study quality variables can be used to define inclusion criteria, so that poorly conducted studies are screened out. Second, the quality of included studies can be summarized so that others can judge the strengths and weaknesses of the evidence base being synthesized. Third, this information can be incorporated into a meta-analysis of study results, such as by examining whether quality-related variables predict variation in effect size magnitude or by sequentially combining study results according to the quality of evidence that they provide (Detsky, Naylor, O’Rourke, McGeer, & L’Abbé, 1992; Higgins et al., 2011).

Further stages of synthesis

After identifying eligible studies, extracting results, coding study characteristics, and assessing the quality of evidence that they provide, the active “field work” involved in a research synthesis is complete. The focus then shifts to synthesizing and drawing inferences from the evidence that has been assembled. These stages of the process are somewhat more technical and involve methods that are more specialized to SCDs, and so we discuss them in subsequent sections. The next section discusses effect size measures for quantifying functional relationships between interventions and outcomes; these indices become the main dependent variable in some approaches to meta-analysis. The following section discusses several approaches for meta-analysis of data from SCDs.

Although the remainder of the chapter focuses on effect sizes and meta-analysis, it is important to note that many reviews of SCDs do not use meta-analytic methods for drawing inferences. Instead, researchers draw conclusions about the evidentiary base for an intervention or practice on the basis of rules derived from professional conventions (Council for Exceptional Children Working Group, 2014; Hitchcock, Kratochwill, & Chezan, 2015; Kratochwill et al., 2013). For example, the WWC Pilot Standards propose that, in order to classify a practice as "evidence-based," the body of research must include at least five single-case studies that meet minimum design standards, were conducted by at least three independent research teams, and include a total of 20 or more participants.

We see some degree of incongruence between use of these criteria and use of meta-analytic approaches to synthesizing evidence from SCDs. Whereas the WWC evidence criteria are based on judgements about the *presence or absence* of effects across participants and contexts, the meta-analytic perspective focuses instead on the *magnitude* of effects, including average magnitude and the extent of variation. It seems to us that the latter perspective provides a more direct way to separate systematic patterns from chance findings, to reconcile conflicting evidence, and to draw generalizations. Still, given that methods for meta-analysis of SCDs are still developing rapidly and that methodological consensus has not yet been reached, it seems likely that both approaches will remain in use for some time to come. Further, employing both perspectives side-by-side, while examining areas of agreement and discrepancies between them, represents an important way to move the field forward.

The final stage in the synthesis process is to report the findings. As in any research report, manuscripts that report research syntheses of SCDs should describe the methods and results in sufficient detail to allow for independent replication. In writing up a study for publication, authors (as well as reviewers) will find it helpful to consult guidelines such as the Meta-Analysis Reporting Standards (APA Publications and Communications Board Working Group on Journal Article Reporting Standards, 2008) or the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Moher, Liberati, Tetzlaff, Altman, & PRISMA Group, 2009).

Effect Size Measures for SCDs

On a broad level, Hedges (2008) defines effect sizes as “quantitative indexes of relations among variables” (p. 167). In the context of SCDs, the main relationship of interest is the functional relationship (i.e., causal effect) between an intervention (independent variable) and an outcome (dependent variable); effect sizes are thus indices that quantify the direction and magnitude of a functional relationship. Both direction and magnitude need to be encoded so that interventions that are highly effective can be distinguished from those that are actively harmful (i.e., a strong but *negative* relationship) as well as those that are inconsequential (i.e., have no or small effects on an outcome).

A wide array of effect size indices have been proposed for use with SCDs. Available effect sizes can be classified broadly into three families: non-overlap measures, within-case parametric measures, and between-case parametric measures. The non-overlap measures include one of the oldest and most widely used SCD effect sizes—the percentage of non-overlapping data (Scruggs, Mastropieri, & Casto, 1987) —and a variety of more recent proposals derived from non-parametric statistical methods. Non-overlap measures quantify functional relationships for each case in the study. The family of within-case parametric measures also involve quantifying functional relationships for each case, but do so based on a parametric model. Between-case parametric measures are a recent innovation designed to quantify average functional relationships at the level of the study, on a scale that facilitates comparisons to effect sizes from between-groups designs.

In this section, we review effect size measures from each family. Given the large number of effect size measures that have been proposed for use with SCDs, our review is necessarily selective, focused on five effect size indices that either are widely applied in practice or have particular strengths that we believe warrant special consideration. Before diving into the details of these indices, it is useful to first review properties that an effect size index should have if it is to be used for synthesizing study results. Understanding these effect size desiderata will help researchers to select one or more appropriate effect size indices from among the many available candidates, as well as to critically assess the effect sizes employed in research syntheses that appear in the literature.

Desirable properties of effect size indices

To be useful in a research synthesis, an effect size index should satisfy several criteria (Lipsey & Wilson, 2001). First and most fundamental is that the effect size index should quantify functional relationships in a way that can be validly compared across studies (Hedges, 2008). Research syntheses often involve combining results across studies that examine similar variables, but which use different procedures to operationalize those variables. For example, studies included in a synthesis of the effects of choice-making opportunities on student problem behavior (Shogren, Faggella-Luby, Bae, & Wehmeyer, 2004) varied in the type of design (multiple baseline or ABAB), the number of sessions per case (ranging from 12 to 40 sessions), and the procedures used to measure problem behavior (interval, continuous, or frequency recording). For an effect size index to be interpretable as a measure of functional relationship, its magnitude should not be strongly influenced by incidental details like these. Effect sizes metrics without this property will create interpretational problems for meta-analysis by introducing artifactual bias and obscuring substantive variation in study results.

Second, to be useful in meta-analysis, effect size measures must be calculable from the information typically available in research reports. This criterion can be a major constraint for effect size indices used in between-groups research, where only summary statistics and selected quantitative results may be available. However, it is less of a constraint with single-case research, where raw data can be extracted reliably from graphs (Moeyaert, Maggin, et al., 2016; Shadish et al., 2009). This makes it feasible to calculate effect size indices that involve data patterns more complicated than basic summary statistics.

Third, effect size measures should ideally be accompanied by measures of sampling uncertainty, such as standard errors or confidence intervals (Hedges, 2008). Standard errors quantify the precision of an effect size estimate—the extent to which the estimate would change if the study were replicated under identical conditions. Measures of uncertainty are used in meta-analysis to determine how much weight to accord each effect size estimate, with more precise estimates being assigned larger weight.

This final criterion presents a particular challenge for SCD effect sizes because SCDs involve repeated measurement of an outcome on each case, which suggests the need to account for serial dependence (auto-correlation) in the outcome data. The presence of serial dependence in SCD data has long been a subject of debate (Busk & Marascuilo, 1988; Huitema & McKean, 1998; Huitema, 1985; Matyas & Greenwood, 1996). A recent meta-analysis of auto-correlation levels in SCD data series indicated near-zero average levels, but also substantial variability from study to study in the extent of auto-correlation. This suggests that, at least in some circumstances, ignoring the possibility of auto-correlation may be imprudent, and there appears to be a growing consensus that statistical methods for SCDs should account for the possibility of some form of serial dependence (Horner, Swaminathan, Sugai, & Smolkowski, 2012; Wolery, Busick, Reichow, & Barton, 2010). Unfortunately, standard errors for some effect size indices are only available under the assumption that the measurements are independent, rather than serially dependent. Although this is a short-coming of existing effect size methods, it need not be a fatal flaw. Rather, the problem can be mitigated by using certain meta-analysis techniques that are robust to mis-estimation of standard errors, as discussed further in a later section.

As will be seen, available effect size indices satisfy these criteria to varying degrees, and none meet all of them completely. Effect sizes also vary in how well they account for specific features of data from SCDs, such as time trends and non-normal outcome distributions. For sake of simplicity, our presentation focuses on effect sizes designed for the more basic case in which the dependent variable does not follow a systematic time trend, while providing references to further extensions that better handle time trends. Also for sake of simplicity, we describe the effect sizes with respect to a comparison between a case's outcomes during an initial baseline phase (A) and the outcomes in a subsequent intervention phase (B). We discuss methods for handling more complicated comparisons at the end of this section.

Non-overlap measures

Non-overlap measures are the oldest and most widely used effect size indices for SCDs (Maggin, O'Keeffe, et al., 2011). This family of measures is sometimes characterized as non-parametric, in that their definitions are not predicated upon distributional assumptions about the outcome measures (Parker, Vannest, & Davis, 2011). Their development was motivated by a search for indices that are relatively easy to calculate, widely applicable, and intuitively interpretable (Scruggs & Mastropieri, 2013). Here, we focus on two measures: the widely-used percentage of non-overlapping data (Scruggs et al., 1987) and the recently proposed non-overlap of all pairs (Parker & Vannest, 2009).

PND. The percentage of non-overlapping data (PND) was the first non-overlap measure to appear in the literature. For an outcome where increase is desirable, PND is defined as the percentage of measurements in the B phase that exceed the highest measurement from the A phase; for an outcome where decrease is desirable, one would instead calculate the percentage of B phase measurements that are lower than the minimum measurement in the A phase (Scruggs et al., 1987). PND can take on values between 0 and 100%. Scruggs and Mastropieri (1998) offered general guidelines for the interpretation of PND, suggesting that a PND value of 90% or greater could be interpreted as indicating a "very effective" intervention; a PND between 70% and 90% as indicating an "effective" one; a PND between 50% and 70% as indicating a "questionable" effect; and a PND of less than 50% as indicating an "ineffective" intervention (p. 224).

Since it was first proposed, PND has been widely criticized (Shadish, Rindskopf, & Hedges, 2008; O. R. White, 1987; Wolery et al., 2010). Using simulation methods, Allison and Gorman (1994) demonstrated that the expected value of the PND statistic is strongly influenced by the number of sessions in the A phase, with more sessions leading to smaller values of PND, even when the intervention has no effect at all (see also Pustejovsky, 2015b). This procedural sensitivity makes it more difficult to compare PND values across cases with different baseline phase lengths. Researchers have also criticized PND as not aligning well with visual inspection of study results (Wolery et al., 2010), for lacking methods to quantify their sampling uncertainty (Shadish et al., 2008), and for lacking discriminatory power (Campbell, 2012). Despite these objections, PND remains by far the most commonly applied effect size in systematic reviews of SCDs (Maggin, O'Keeffe, et al., 2011). For example, Bellini and colleagues (2007) used PND in a synthesis of research on school-based social skills interventions for children with autism. Schlosser, Lee, and Wendt (2008) reviewed how PND has been used in systematic reviews of SCDs and provided guidance on its application.

NAP. Parker and Vannest (2009) proposed the non-overlap of all pairs (NAP) statistic, which involves pairwise comparisons between each point in the B phase and each point in the A phase. For an outcome where increase (decrease) is desirable, NAP is defined as the percentage of all such pairwise comparisons where the measurement from the B phase exceeds (is less than) the measurement from the A phase. Pairs of data points that are exactly tied are counted with a weight of 0.5. The logical range of NAP is from 0 to 100%. When the intervention has no effect, the expected magnitude of NAP is 50%. In contrast to PND, the magnitude of NAP is not affected by the number of sessions in the A phase or the B phase, although it will be sensitive to other procedural factors that affect the variability of the outcomes (Pustejovsky, 2015b).

Parker and Vannest (2009) argued that NAP has several advantages over other non-overlap measures, including ease of calculation, better discrimination among effects in published SCDs, and the availability of valid standard errors and confidence intervals. As they also noted, NAP has been proposed as an effect size measure (under a variety of different names) in many other areas of application (e.g., Acion, Peterson, Temple, & Arndt, 2006; Vargha & Delaney, 2000). Based on visual assessment of a corpus of SCD studies, Parker and Vannest (2009) characterized NAP values between 0 and 65% as “weak,” values between 66% and 92% as “medium,” and values between 93% and 100% as “large” (p. 364). Gaskin, McVilly, and McGillivray (2013) employed NAP (along with several other non-overlap measures) in a synthesis of research on restraint reduction interventions for individuals with developmental disabilities.

An approximate standard error for NAP can be calculated if the outcome measurements are mutually independent. Suppose that there are *m* sessions in phase A and *n* sessions in phase B, and let us denote the outcome data in each phase as $y_1^A, ..., y_m^A$ and $y_1^B, ..., y_n^B$, respectively. The NAP statistic and its standard error both involve comparisons of all $m \times n$ pairs of outcomes, denoted q_{ij} as for $i = 1,\ldots,m$ and $j = 1,\ldots,n$. Let $q_{ij} = 1$ if $y_j^B > y_i^A$, $q_{ij} = 0.5$ if $y_j^B = y_i^A$, and $q_{ij} = 0$ if $y_j^B < y_i^A$. The NAP effect size index is then calculated as

$$\text{NAP} = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} q_{ij} , \tag{1}$$

Hanley and McNeil (1982) proposed a method for estimating the sampling variance of a statistic that is equivalent to NAP. As originally developed, the estimator assumed that the outcome scores contained no ties; here, we present a small, ad hoc modification to account for ties. Calculate

$$Q_1 = \frac{1}{mn^2} \sum_{i=1}^{m} \left[\sum_{j=1}^{n} q_{ij} \right]^2 \quad \text{and} \quad Q_2 = \frac{1}{m^2 n} \sum_{j=1}^{n} \left[\sum_{i=1}^{m} q_{ij} \right]^2 . \tag{2}$$

The sampling variance of NAP can then be estimated as

$$V_{\text{NAP}} = \frac{1}{mn} \left[\text{NAP}\left(1 - \text{NAP}\right) + \left(n-1\right)\left(Q_1 - \text{NAP}^2\right) + \left(m-1\right)\left(Q_2 - \text{NAP}^2\right) \right], \tag{3}$$

with standard error of NAP given by $SE_{\text{NAP}} = \sqrt{V_{\text{NAP}}}$. This standard error is based on the assumption that the outcomes within each phase are mutually independent; it will not be accurate in the presence of serial dependence.

Other non-overlap measures. Several other indices in the family of non-overlap measures have developed for use with SCDs. Parker, Vannest, and Davis (2014) provide an expansive review, including worked examples of how to calculate each index based on graphed data. As an extension to NAP, Parker, Vannest, Davis, and Sauber (2011) proposed a set of related non-overlap measures called Tau-U, which can be used to adjust for time trends in the A phase, B, phase, or both. Other recently proposed indices that account for time trends include the percentage of non-overlapping corrected data (Manolov & Solanas, 2009) and the percentage exceeding median trend (Wolery et al., 2010). However, we believe that these proposed effect sizes are best treated as experimental, insofar as their properties have not be studied extensively and it is not always clear whether they can reasonably be interpreted as measures of effect magnitude.

Within-case parametric measures

In contrast to the non-overlap measures, other effect size indices for SCDs are defined in parametric terms, based on distributional assumptions about the process that generated the observed data for a given case. This approach has the advantage that the definition of the effect size index is clearly separated from the statistics used to estimate it (a distinction that is less clear for many of the non-overlap measures). The major challenge with using parametric effect sizes comes in assessing whether their distributional assumptions are reasonable for the data under analysis. We again limit consideration to one well-known effect size index (the within-case standardized mean difference) and one more recent, promising index (the log-response ratio).

Within-case standardized mean difference. The standardized mean difference (SMD) is one of the most commonly used effect size measures in syntheses of between-groups intervention research. In that setting, the SMD is defined as the difference in mean outcomes between a treatment population and a control population, scaled by the standard deviation (SD) of the outcome in the control population (or in both populations, if assumed to have equal variance). Scaling by the SD leads to an effect size index that is invariant to the scale of the outcome measure, so that it is meaningful to compare SMD indexes across studies that use different instruments to measure a common outcome construct (Hedges, 2008).

Drawing analogies to how the SMD is used with between-groups designs, Gingerich (1984) and Busk and Serlin (1992) proposed a version of the SMD as an effect size for SCDs. This within-case SMD is defined as

$$\delta = \frac{\mu_B - \mu_A}{\sigma_A}, \tag{4}$$

where μ_A is the expected level of the outcome in the A phase, μ_B is the expected level of the outcome in the B phase, and σ_A is the SD of the outcome in the A phase. Although the within-case SMD is similar in form to the SMD for between-groups design, there is a crucial difference. The SD used to scale the within-case measure represents within-individual variability only, whereas the SD used to scale the between-groups SMD represents both between- and within-individual variability in the outcome. As a result, the two effect size measures are on quite different scales and are not directly comparable (Shadish, Hedges, & Pustejovsky, 2014; Van den Noortgate & Onghena, 2008).

Estimates of the within-case SMD and its sampling variance are available under the assumption that the outcome measures from each phase are mutually independent. Let $\bar{y}_A$ and

$\bar{y}_B$ denote the sample means, s_A and s_B denote the sample SDs, and *m* and *n* denote the number of sessions from phases A and B, respectively. Both Gingerich (1984) and Busk and Serlin (1992) suggested estimating the within-case SMD by plugging in sample quantities for the corresponding parameters, so that the within-case SMD is estimated as $d = \left(\bar{y}_B - \bar{y}_A\right) / s_A$. However, this plug-in estimator will have a non-negligible bias when the baseline phase consists of only a few sessions. An approximately unbiased estimator of the within-case SMD is given by

$$g = \left(1 - \frac{3}{4m-5}\right)\left(\frac{\bar{y}_B - \bar{y}_A}{s_A}\right) \tag{5}$$

(cf. Hedges, 1981). An approximate standard error for *g* can be calculated as

$$SE_g = \left(1 - \frac{3}{4m-5}\right)\sqrt{\frac{1}{m} + \frac{s_B^s}{ns_A^2} + \frac{d^2}{2(m-1)}}\ . \tag{6}$$

Randolph (2007) used the bias-corrected estimator (*g*) of the within-case SMD in a meta-analysis of research examining the effects of response cards on academic achievement. Ugille and colleagues (2014) examined several alternative approaches to estimating the within-case SD and correcting the small-sample bias of the estimator.

It is important to emphasize that both the within-case SMD effect size estimator and its standard error are only valid when the outcome measurements are independent. This represents a crucial drawback of this effect size—not only will its variability be estimated incorrectly, but the effect size index itself will also be biased in the presence of serial dependence. (The bias in *g* arises because s_A^2 is a biased estimator for σ_A^2 when the outcomes are serially dependent.) Extensions to the within-case SMD have been developed recently that do account for certain forms of serial dependence, as well as time trends in the A and B phase (Maggin, Swaminathan, et al., 2011).

Log response ratio. In both primary studies and systematic reviews of SCDs, it is common to characterize functional relationships in proportionate terms—i.e., as a percentage change in the level of the outcome from baseline to intervention (e.g., Gaskin et al., 2013). Percentage change measures have also occasionally been applied in syntheses of SCDs, although usually without supporting statistical development (Campbell, 2003; Kahng, Iwata, & Lewin, 2002; Marquis et al., 2000). However, a recently proposed effect size index known as the log response ratio (Pustejovsky, 2015a) quantifies the magnitude of functional relationships in a way that is closely related to proportionate change, and does have a formal statistical grounding. Again letting μ_A and μ_B denote the expected values of the outcome in phases A and B, respectively, and letting ln(.) denote the natural logarithm function, the LRR parameter is defined as

$$\psi = \ln\left(\mu_B / \mu_A\right). \tag{7}$$

This index is appropriate for outcomes measured on a ratio scale, such as frequency counts of behavior or behaviors measured using percentage duration; it would not be appropriate for outcomes such as rating scales, where a score of zero does not correspond to the absence of the outcome. The natural logarithm transformation is used because it makes the range of the index less restricted. When the intervention has no effect on the outcome, then $\mu_B / \mu_A = 1$ and so the index will be equal to zero.

A basic plug-in estimator for the LRR can be calculated by replacing the expected values with the corresponding sample means, yielding:

$$R_1 = \ln\left(\bar{y}_B / \bar{y}_A\right). \tag{8}$$

However, this basic estimator has a small-sample bias. A bias-corrected estimator can be calculated as

$$R_2 = \ln\left(\bar{y}_B\right) + \frac{s_B^2}{2n\bar{y}_B^2} - \ln\left(\bar{y}_A\right) - \frac{s_A^2}{2m\bar{y}_A^2} \tag{9}$$

and should be used when either phase contains only a small number of observations (Pustejovsky, 2015a). Under the assumption that the outcomes in each phase are mutually independent, an approximate standard error for R_2 is given by

$$SE_R = \sqrt{\frac{s_A^2}{m\bar{y}_A^2} + \frac{s_B^2}{n\bar{y}_B^2}}\,. \tag{10}$$

However, just as with standard errors for the other effect sizes described in this section, formula (10) is not a valid estimator if the outcomes are serially correlated; in the presence of positive auto-correlation it will tend to under-estimate the sampling variability of the effect size index.

Pustejovsky (2015a) argues that the LRR is a particularly appropriate effect size index for SCDs that use behavioral outcome measures measured through direct observation. One of its advantages is that its magnitude remains stable when outcomes are measured using different operational procedures, such as use of longer or shorter observation sessions or use of momentary time sampling instead of continuous recording. Furthermore, under certain circumstances, LRR effect sizes based on different dimensional constructs can nonetheless be directly compared. Finally, the LRR is directly related to percentage change measures of effect size; the latter can be calculated from the former as

$$\text{percentage change} = 100\% \times \left[\exp(\psi) - 1\right], \tag{11}$$

where exp(.) denotes exponentiation. As a result of this algebraic relationship, meta-analysis based on LRR effect sizes can be translated into conceptually appealing terms of percentage change. One limitation of the LRR is that, as currently developed, it is based on the assumption that the level of the outcome is stable within each phase. Extensions for handling time trends appear possible in principle, but have yet to be investigated.

Other within-case parametric measures. Center, Skiba, and Casey (1985) proposed another distinct approach to defining a within-case parametric effect size based on a piece-wise linear regression model. In their approach, magnitude of effect is quantified in terms of changes in the proportion of variance explained by the model, compared to a model that assumes no differences across phases. Beretvas and Chung (2008) provide a critical review of this and related approaches (e.g., Allison & Gorman, 1993; Faith, Franklin, Allison, & Gorman, 1996). In our view, this category of effect size indices is less useful than other within-case parametric measures because they conflate different dimensions of change in response to treatment (i.e., changes in level, changes in slope), which makes them less directly interpretable as measures of effect magnitude.

Between-case parametric measures

Hedges, Pustejovsky, and Shadish (2012, 2013) proposed a novel approach to defining and estimating effect size indices for SCDs that are directly comparable to the SMD indices from between-groups designs. These between-case effect size indices involve modeling and summarizing the data across multiple participants simultaneous, rather than estimating separate effect sizes for each case. Broadly, the approach is based upon a hierarchical model that describes both the functional relationship for each case and how the pattern of results varies across the individual cases in the study. This model is then used to consider a hypothetical scenario: what would have happened—and how big an effect size would have been observed—if a between-groups experiment had been performed on the same population of participants?

The between-case SMD effect size index is premised on a certain statistical model for the data, and it is important to be aware of the modeling assumptions involved. The original methods proposed for treatment reversal designs (Hedges et al., 2012) and multiple baseline designs

(Hedges et al., 2013) involve the following assumptions: (a) the baseline is stable (i.e., no baseline trend); (b) the intervention leads to an immediate change in level (i.e., no intervention-phase trend); (c) the intervention effect is constant across cases; (d) the outcome is normally distributed about case- and phase-specific mean levels; and (e) the errors follow a first-order auto-regressive process. The last assumption means that the between-case SMD effect size estimate allows for a certain type of serial dependence, rather than assuming independence of the outcome measurements. The approach has also been extended to accommodate a variety of more general models, including those with time trends or heterogeneity of effects across cases (Pustejovsky, Hedges, & Shadish, 2014). Swaminathan, Rogers, and Horner (2014) proposed extensions that use Bayesian estimation methods.

The calculations involved in estimating between-case SMDs and their sampling variances are too involved to describe here. Software for carrying them out is available in the form of an SPSS macro (Marso & Shadish, 2015) or a package for the R statistical computing environment (Pustejovsky, 2015c). Shadish, Hedges, and Pustejovsky (2014) provided detailed examples demonstrating how to apply the SPSS macro and interpret its output. Losinski and colleagues (2014) used between-case SMDs in a synthesis of the effects of self-regulated strategy development.

Shadish, Hedges, Horner, and Odom (2015) argued that between-case SMD effect sizes have two key advantages. This first advantage is translational: these indices describe the results of SCD studies in a metric that is familiar to researchers who work primarily with between-groups designs, making it more likely that SCDs will be considered for evidence-based practice reviews. Second, the between-case indices allow researchers to compare the results from SCDs to the results from between-groups designs, which may promote a stronger understanding of the utility and limitations of each type of design.

Between-case SMDs are also limited in several respects. The more basic indices (Hedges et al., 2012, 2013) are only available for treatment reversal (e.g., ABAB) designs, multiple baselines across participants, or multiple probes across participants, and the study must include at least three individual participants. The more flexible models (Pustejovsky et al., 2014) generally require data from more than three individuals. More fundamentally, between-case effect sizes have the limitation that they describe *average* effects across cases, and thus potentially conceal individual heterogeneity. This limitation is an inherent consequence of seeking comparability with between-groups effect sizes—because between-groups designs only provide information about average effects—and may make the approach less congruent with visual assessments of SCDs (cf. Kratochwill & Levin, 2014).

Handling multiple phase-contrasts

Our discussion of effect sizes for single-case designs has mostly focused on indices comparing a single A phase with a single B phase. In practice, syntheses will often include SCDs that involve more elaborate designs, such as treatment reversals in which there are multiple AB replications for a given case. Unfortunately, there is not currently consensus on the best approach for dealing with studies that involve multiple phase replications. Rather, researchers have followed a variety of different strategies.

The simplest strategy is to calculate an effect size estimate based on a comparison between only the two phases that best capture the functional relationship of interest. For example, Heath and colleagues (2015) computed effect sizes comparing the initial baseline and the initial intervention phase only, arguing that this comparison was most compatible with the comparison of phases from multiple baseline designs. In contrast, Heyvaert and colleagues (2014) computed effect sizes comparing the initial baseline and the final treatment phase. While simple, these approaches have the drawback that they do not use all available data.

Another strategy is to pool the data across phases from common conditions, then calculate an effect size estimate based on the pooled samples (cf. D. M. White et al., 1989). For instance, in an ABAB design, data from the initial baseline (A1) and return to baseline (A2) would be treated as a single phase, as would data from the initial treatment phase (B1) and the re-introduction of treatment (B2). This approach is computationally straight-forward, but may be less appropriate if the outcome does not immediately return to baseline levels upon removal of the treatment.

A further possibility is to calculate an effect size estimate for each phase contrast of interest, then average those estimates together to obtain a single effect size for each case (cf. Maggin, Chafouleas, Goddard, & Johnson, 2011). For example, an ABAB design might yield two effect size estimates, based on the comparison between phase A1 and B1 and the comparison between A2 and B2, or three estimates, if the comparison between A1 and B2 is included as well. This approach has the advantages of using all of the available data, being more consistent with the logic of treatment reversal designs, and yielding more precise estimates of treatment effects than those based on comparisons between just two phases (Maggin, Chafouleas, et al., 2011).

Given that several options exist, researchers conducting a synthesis will need to choose and justify a strategy for handling multiple phase replications. We would recommend that researchers take into account the context and features of the SCDs to be included in the synthesis when selecting a strategy. In some instances, all of the options may yield very similar results. In instances where there are discrepancies, researchers should examine the included studies to evaluate which strategy best represents the logic of the studies' designs.

Meta-Analysis Methods

In syntheses of between-groups designs, the results of each study are typically summarized in the form of an effect size index, and these effect size indices are then combined or compared using meta-analysis techniques. Meta-analysis of SCDs differs from the conventional approach used with between-groups designs in two crucial respects. First, data from SCDs has a multi-level structure and it is important to take this structure into account for both substantive and technical reasons. On a substantive level, single-case researchers are often interested in the extent to which the efficacy of a treatment varies across individuals, and this heterogeneity can be characterized using a multi-level structure. On a technical level, the outcomes for two cases within the same study are likely to be more closely related than for two cases drawn from different studies, due to common contextual and operational features. This creates dependence among cases within the same study, which must be taken into account for valid statistical inferences to be drawn.

The second distinctive feature of meta-analysis of SCDs is that, compared to meta-analysis of group designs, it is relatively feasible to use meta-analytic techniques for individual participant data, rather than being limited to meta-analyzing summary effect size indices. Although individual-participant data has several advantages for meta-analysis of between-groups designs (Cooper & Patall, 2009), its use is still relatively uncommon there because raw data are not always readily available. With single-case studies, the raw data are usually available in the form of graphs, making meta-analysis of individual participant data a feasible option. In this section, we first discuss multilevel meta-analysis techniques for synthesizing case-level effect size indices, and then turn to multilevel modeling of individual participant data.

Multilevel Meta-Analysis Model for Effect Sizes

Meta-analysis of effect sizes takes place in two distinct stages. In the first stage, effect size indices are obtained for each case, and in the second stage these effect sizes are analyzed. Descriptive analyses can be done to summarize the distribution of any of the effect size estimates reviewed in the previous section, yielding statements about the average observed effect size and the range of observed effect size estimates. However, researchers often want to move beyond description of the observed effect sizes to make inferences about the magnitude of functional relationships, such as constructing a confidence interval for the average effect size or testing a hypothesis about whether case level characteristics (e.g., age, gender, treatment fidelity) or study-level characteristics (e.g., setting, treatment protocol) moderate the magnitude of effect. Multilevel meta-analysis allows these kinds of inferences to be made, while accounting for the nesting of case-level effect size indices within studies (Ugille, Moeyaert, Beretvas, Ferron, & Van den Noortgate, 2012; Van den Noortgate & Onghena, 2003a, 2008).

We shall now describe the basic multi-level meta-analysis model for case-level effect sizes. Let B_{jk} denote the estimated effect size for case *j* in study *k.* At the first level of the multilevel meta-analysis model, the estimated effect size is modeled as the sum of the *true* effect size β_{jk} and a sampling error:

$$B_{jk} = \beta_{jk} + e_{jk} \qquad (12)$$

where the sampling error e_{jk} is assumed to be normally distributed with mean zero and variance V_{jk} equal to the squared standard error of the effect size estimate. The second level of model describes variation in the true treatment effects across cases within a given study:

$$\beta_{jk} = \theta_k + u_{jk} \qquad (13)$$

where θ_k is the average (mean) true effect size for cases in study *k* and u_{jk} is a random deviation for case *j* in study *k*, which is typically assumed to be normally distributed with mean zero and variance ω^2. At the third level, variation in the average effect size across studies is modeled as:

$$\theta_k = \gamma + v_k \tag{14}$$

Where γ is the average true effect size across all studies and v_k is a random deviation for study *k*, which is typically assumed to be normally distributed with mean zero and variance τ^2.

The multilevel meta-analysis model can be further expanded to examine the effects of moderator variables by including case characteristics or study characteristics as predictors in the level two or level three model, respectively (Van den Noortgate & Onghena, 2008). Thus, the parameters of the model align well with the inferential goals of meta-analysis. Specifically, the average effect size across studies (γ) is estimated along with its standard error, which can be used to create a confidence interval or test a hypothesis about the mean effect. In addition, variation in true effect sizes is estimated both across studies and across cases within studies (τ^2 and ω^2, respectively), both of which characterize the degree of heterogeneity in the magnitude of the functional relationship of interest. Finally, adding moderators into the level-2 and level-3 equations provides a way to examine the extent to which case- or study-level characteristics explain variation in the magnitude of the functional relationship.

Meta-analysis of multi-variate effect sizes. The multilevel meta-analytic model we presented here is a univariate model, but multivariate extensions can be made. Consider for example an analyst that is using the regression approach and conceptualizes two of the regression coefficients as effect indicators, one corresponding to the shift in level that occurs with intervention and the other corresponding to the shift in the slope that occurs with intervention (Van den Noortgate & Onghena, 2003b, 2008). The analyst could analyze each of these standardized regression coefficients separately with the previously described univariate model, or the analyst could analyze the two standardized regression coefficients simultaneously with a multivariate model (Van den Noortgate & Onghena, 2003a, 2008).

Estimating the model. The most common approach to estimating the multi-level meta-analysis model is via restricted maximum likelihood, as implemented in SAS PROC MIXED (Littell, Milliken, Stroup, Wolfinger, & Schabenberber, 2006) or the metafor package in R (Viechtbauer, 2010), although other approaches are also possible. It is important to note that the default standard errors, confidence intervals, and hypothesis tests based on the model estimates will only be valid if the sampling variances of the effect size estimates (i.e., the V_{jk}'s) are accurate.

Robust variance estimation. As noted in the previous section, valid standard errors for most within-case effect sizes are only available under the assumption that the outcome measurements are independent. If the data are serially dependent, then the standard errors will be incorrect and inferences based on the multi-level meta-analysis model may become inaccurate. However, a technique called robust variance estimation (Hedges, Tipton, & Johnson, 2010) can be used to generate valid standard errors, confidence intervals, and hypothesis tests for average effect sizes and meta-regression coefficients (i.e., moderators) in the multi-level meta-analysis model, even if the standard errors of the effect size estimates are incorrect. The main drawback to this technique is that it requires more independent studies to achieve adequate power; also, when only a small number of independent studies are included in the meta-analysis, finite-sample corrections are required (Tipton & Pustejovsky, 2015; Tipton, 2015). Tanner-Smith and Tipton (2014) provide an accessible introduction to robust variance estimation and review available software.

Multilevel Model for Individual Participant Data

Multilevel meta-analytic models have also been developed for both raw score and standardized individual participant data from SCDs. Raw score models are appropriate if in each study the outcome measurement is comparable, such as when all studies being synthesized operationalize reading fluency as words read correct per minute or problem behavior as the proportion of 30 s intervals where the problem behavior was exhibited. Standardized score models become the choice when the outcome measurement is not comparable across studies. To standardize the raw scores, Van den Noortgate and Onghena (2008) proposed to divide the raw scores by the root mean-square error of a case-specific regression model (see also Moeyaert et al., 2013).

Let Y_{ijk} be the outcome at time point *i* for case *j* in study *k*, measured in either raw or standardized units. At level-1 of the multilevel model, Y_{ijk} is specified to be a function of the phase of the design. The model is further specified based on the analyst's assumptions about the structure of the relationship between the time of observation and the outcome and about the distribution of errors over time (e.g., independent, normally distributed, and homogeneous; serially dependent, or heterogeneous). The simplest level-1 model specification would be:

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} D_{ijk} + e_{ijk} \tag{15}$$

where D_{ijk} indicates whether a given observation was in baseline (coded 0) or intervention (coded 1), β_{0jk} is the mean outcome for case *j* in study *k* during the baseline phase, β_{1jk} is the shift in level for case *j* in study *k* (i.e., mean difference between the intervention and baseline phase), and the errors (e_{ijk}) are assumed to be independent and normally distributed with a mean of 0 and variance of σ^2 (Owens & Ferron, 2012; Van den Noortgate & Onghena, 2003a). In this model, the magnitude of the functional relationship for case *j* in study *k* is represented by β_{1jk} and is assumed to be constant throughout the intervention phase.

This level-1 model can be expanded in a variety of different ways to account for time trends in the baseline phase, treatment phase, or both (Rindskopf & Ferron, 2014). One specification of particular interest allows for a linear time trend during both phases (Huitema & McKean, 2000; Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2014). Let $Time_{ijk}$ denote the session number of observation *i* for case *j* in study *k*. The level-1 model is then

$$Y_{ijk} = \beta_{0jk} + \beta_{1jk} D_{ijk} + \beta_{2jk} Time_{ijk} + \beta_{3jk} D_{ijk} Time_{ijk} + e_{ijk} \qquad (16)$$

In this model, the treatment effect is assumed to change linearly with time as $\beta_{1jk} + \beta_{3jk} Time_{ijk}$. Judicious centering of the *Time* variable allows β_{1jk} to be interpreted as the treatment effect at a point in time that is of focal interest, such as the first or final session of the intervention phase. Further modifications to the level-1 model have been suggested to accommodate non-linear trends (Hembry, Bunuan, Beretvas, Ferron, & Van den Noortgate, 2015) or designs with more than two phases (Rindskopf & Ferron, 2014; Van den Noortgate & Onghena, 2007). In addition, a variety of alternative error structures have been considered, including first order autoregressive models (Petit-Bois, Baek, Van den Noortgate, Beretvas, & Ferron, 2015) and models with heterogeneous variances across phases (Ferron, Moeyaert, Van den Noortgate, & Beretvas, 2014).

After specifying a level-1 model, the analyst specifies a level-2 model to account for variation in the coefficients across cases within a study. For example, if Equation (15) is chosen for the level-1 model, the level-2 model would have two equations, one to model variation in the baseline level (β_{0jk}) across cases and one to model variation in the treatment effect (β_{1jk}) across cases in the study:

$$\begin{aligned}\beta_{0jk} &= \theta_{00k} + u_{0jk} \\ \beta_{1jk} &= \theta_{10k} + u_{1jk}\end{aligned} \tag{17}$$

where θ_{00k} is the mean baseline level across cases in study *k*, θ_{10k} is the mean treatment effect across cases in study *k*, and the case specific errors (u_{0jk} and u_{1jk}) are assumed to be multivariate normal with means of zero and covariance $\mathbf{\Omega}_u$, where $\mathbf{\Omega}_u$ may be diagonal or unstructured (Moeyaert, Ugille, Ferron, Beretvas, & Van den Noortgate, 2016). To examine potential case-level moderators of the treatment effect, variables describing case-level characteristics could be included as predictors in Equation (17).

In addition to specifying level-2 equations to allow for variability between cases within a study, level-3 equations are specified to account for variation between studies. For example, if Equation (15) is specified as the level-1 model and Equation (17) as the level-2 model, then the level-3 model would be:

$$\begin{aligned}\theta_{00k} &= \gamma_{000} + v_{00k} \\ \theta_{10k} &= \gamma_{100} + v_{10k}\end{aligned} \tag{18}$$

where γ_{000} is the overall mean baseline level, γ_{100} is the overall mean treatment effect, and the study specific errors (v_{00k} and v_{10k}) are assumed to be distributed multivariate normal with means of zero and covariance $\mathbf{\Omega}_v$, where $\mathbf{\Omega}_v$ may be diagonal or unstructured (Moeyaert, Ugille, et al., 2016). Study characteristics could be included as predictors in Equation (18) in order to examine potential moderating effects.

Just as with the meta-analysis model for effect sizes, the multilevel modeling of individual participant data yields parameter estimates that align well with the objectives of meta-analysts. Estimates of the average treatment effect across studies (e.g., γ_{100}) are provided along with corresponding standard errors, as well as estimates of the variation in the treatment effect both across studies and across cases within studies. This variation can be further explored by adding moderators into the level-2 and level-3 equations, which provides estimates of the differences in the magnitude of effects for varying case- and study-level characteristics.

The multilevel modeling of individual data is flexible enough to allow for a variety of assumptions about the data, including alternative assumptions about the structural relationship of time with the outcome and alternative assumptions about the errors. As with any statistical model, care must be taken to specify a meta-analysis model that is conceptually appropriate to the research studies being synthesized and consistent with the data being analyzed. Finally, when sample sizes are small, inferences can be made more accurately when degrees of freedom are estimated using either Kenward-Roger or Satterthwaite approaches (Ferron, Bell, Hess, Rendina-Gobioff, & Hibbard, 2009).

Outstanding Issues

The previous sections have reviewed an array of effect size indices for characterizing the magnitude of functional relationships, as well as for synthesizing the results of SCDs using meta-analysis. Despite recent advances, there remain a number of limitations to existing methods, which create challenges for synthesizing SCDs. In this section, we comment briefly on some of the major outstanding issues and areas for further methodological research.

Publication and reporting biases

The validity of generalizations from a research synthesis—whether based on between-groups designs, SCDs, or both—is contingent on the assumption that the studies in the synthesis are representative of the full body of research relevant to the topic of investigation. In areas of research that rely on between-groups designs and statistical inference, there is strong evidence that published findings represent an incomplete and biased view of the full population of research (Rothstein et al., 2005). Biases arise due to a preference—on the part of both authors and journals—to publish statistically significant results, coupled with analytic flexibility, preference for novelty, and pressures created by outside interests (Ioannidis, 2005, 2008). Concern about reporting bias is the main reason that many research syntheses expend great effort in searching for unpublished results, as we noted in a previous section. Besides searching for unpublished studies, a number of other meta-analytic tools have been developed for detecting and adjusting for reporting biases in between-groups research (e.g., Duval & Tweedie, 2000; Egger, Smith, Schneider, & Minder, 1997; Vevea & Hedges, 1995).

There is good reason to expect that publication and reporting biases are likely to operate in single-case research as well, particularly due to the emphasis placed on visually detectable evidence for functional relationships (Kazdin, 2011; Kratochwill, Levin, Horner, & Swoboda, 2014). Furthermore, Sham and Smith (2014) provided initial evidence that published SCDs depict larger effects than unpublished studies, and Shadish and colleagues (2016) found that single-case researchers tend to selection (or censor) studies for publication based on visual assessments of the magnitude of functional relationships. However, there are currently few methods for detecting or addressing reporting biases in meta-analysis of SCDs. Some recent research syntheses of SCDs have applied publication bias methods from between-groups research (Dart, Collins, Klingbeil, & McKinley, 2014; Shadish et al., 2013), but this approach is unlikely to be adequate because the between-groups tools are premised on the assumption that reporting biases arise from statistical significance testing. If syntheses of SCDs are to be used to inform evidence-based practice, there is a critical need to develop better tools for understanding and addressing reporting biases.

Matching the analysis to the design

Most effect sizes that have been proposed for use with SCDs are defined in terms of comparisons between a baseline (A) phase and a treatment (B) phase. These effect sizes work well for multiple baseline designs, in which each case is assessed in just one baseline and one treatment phase; using the extensions described in a previous section, they can also be applied to treatment reversal designs that contain multiple phase comparisons. However, it is less clear how they can be applied to other types of SCDs, such as alternating treatment designs, changing criterion designs, or hybrid designs that combine multiple strategies for experimental control. The flexible and creative application of design elements is viewed as a strength of single-case research (Gast & Ledford, 2014), and so it is clearly not ideal to exclude these designs from syntheses. Some initial work has been done on estimating and meta-analyzing effects from different types of single-case designs (Moeyaert et al., 2015), but further, broader investigation is warranted. Further work will need to focus on developing estimators for both within-case and between-case effect size indices in a way that captures the key features of these more complex types of designs (Horner et al., 2012; Parker & Vannest, 2012), as well as how to incorporate effect sizes or raw data from these designs into a meta-analysis.

More broadly, further research is needed on how to develop analysis procedures that better match the logic of how SCDs are used in practice. For example, the key feature of multiple baseline designs is staggering the introduction of treatment across cases, which provides a way to examine internal validity threats created by outside influences that are common across cases (Gast & Ledford, 2014; Horner et al., 2005). However, effect sizes based on within-case comparisons can still be biased by these outside influences. There is recent work on detecting such biases by using effect estimates that involve between-case (vertical) comparisons (Ferron et al., 2014), but more research is needed to determine the extent and magnitude of the biases and to develop methods to adjust for them.

In practice, the design of single-case studies is often guided by ongoing visual analysis of the data. For example, in a response-guided multiple baseline study, the baseline phase might be extended until the baseline data are deemed stable and the intervention staggers between cases might be extended to allow each successive case to demonstrate a response to the intervention prior to intervening with the next case (Ferron & Jones, 2006). Response-guided experimentation is seen as an integral feature of single-case research (Baer, Wolf, & Risley, 1968; Barlow, Nock, & Hersen, 2008; Kazdin, 2011), but few statistical methods for analyzing data from SCDs take its consequences into account. Concerns have been raised that response-guided experimentation may lead to biased estimates of functional relationships (Dugard, File, & Todman, 2012; Ferron et al., 2014). More research is needed to understand its consequences for effect size estimation and to develop statistical methods that take into account how single-case designs are applied in practice.

Evaluating model quality

As we have sought to demonstrate in the previous sections, researchers interested in synthesizing SCDs now have a diverse range of statistical tools available. Much of the methodological research effort over the past decade or more has focused on expanding and improving upon previous methods, such as developing new non-overlap measures (e.g., Parker, Vannest, & Davis, 2011) and more flexible multi-level models for analyzing or meta-analyzing SCDs (Moeyaert, Ferron, Beretvas, & Van den Noortgate, 2014; Rindskopf & Ferron, 2014). As the range of options continues to expand, however, researchers will need guidance about how to determine which of the many available methods are best suited for a given set of data to be synthesized.

This issue can be seen as the broad question of how to evaluate model quality, where the model encompasses the choice of effect size index or approach to scaling the raw data and the meta-analytic model used to synthesize study results. In absolute terms, guidance is needed about how to assess whether the critical assumptions of an effect size and a meta-analytic model are adequate. In relative terms, methods are needed for determining which of several possible approaches provides the best summary description of the data. Very little work has addressed these questions in the context of single-case research, although it strikes us that finding ways to more closely integrate visual inspection with statistical modeling holds promise as a way to make progress (e.g., Baek, Petit-Bois, Van den Noortgate, Beretvas, & Ferron, 2014; Davis et al., 2013).

Conclusion

In this chapter, we have described the process of conducting a synthesis of single-case research and reviewed a selection of available methods—including effect size indices and approaches to meta-analytic modeling—for combining, contrasting, and examining study results. The volume of recent methodological developments in this area is exciting. However, it may also seem overwhelming to researchers (and perhaps especially to students) seeking to complete a synthesis project, particularly given the current lack of consensus guidance about some aspects of the process. Regarding the choice of effect sizes, researchers can be guided to some extent by the abstract criteria for what makes an effect size index useful for synthesis, as discussed in a previous section. Beyond this, though, we would suggest several actions that researchers can take to conduct successful syntheses of single-case research, even as the methodology continues to evolve.

First, researchers can use multiple methods for meta-analyzing the data and can examine the extent to which their main findings hold across methods (cf. Kratochwill et al., 2013). Of course, using every method that has been proposed is neither practical nor desirable; sensitivity analysis is most compelling when based on a judicious selection of complementary analytic methods. In a meta-analysis of SCDs, this might involve using both a non-overlap measure and a within-case parametric effect size, or using a meta-analysis of effect size indices in addition to a meta-analysis of individual participant data.

Second, researchers can seek feedback or collaborate with methodologists. As several field leaders have emphasized (e.g., Campbell, 2012; Fisher & Lerman, 2014; Parker & Vannest, 2012; Shadish, 2014), dialogue between methodologists and researchers with substantive expertise (and practical experience carrying out single-case studies) is crucial for improving the relevance and usability of statistical methods for single-case data. Speaking as methodologists, we have found collaborations to be particularly valuable because applied researchers help us to identify and scrutinize our underlying assumptions, while also pushing us to clarify how we explain things.

Third, researchers can enable replication and re-use of their synthesis projects by making the underlying data available (including study characteristics and raw outcome data). Data can be posted either as supplementary materials on a journal website, or through an open data repository such as the Open Science Framework (https://osf.io/) or Harvard Dataverse (https://dataverse.harvard.edu/). Providing the underlying data makes it easier for future syntheses to build upon existing work, thereby strengthening the contribution of the project (cf.

Albarracín, 2015). Re-analysis and revisiting of existing syntheses is particularly important as methods continue to evolve. Moreover, as research synthesists, we rely on the availability of data from primary study reports; we should thus hold ourselves to our own highest standard by providing a full, complete, and readily accessible record of our work.

Research synthesis projects are not easy or simple undertakings, yet the potential contributions of SCD research synthesis efforts are also considerable, in terms of providing a sound basis for identifying evidence-based practices, characterizing and identifying sources of heterogeneity in intervention effects, and even strengthening the methodology of the discipline. We would encourage researchers to keep these aims in mind as they conduct reviews, explore techniques for meta-analysis of SCDs, and further develop the methodology of single-case research synthesis.

References

Acion, L., Peterson, J. J., Temple, S., & Arndt, S. (2006). Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine, 25*(4), 591–602. doi:10.1002/sim.2256

Albarracín, D. (2015). Editorial. *Psychological Bulletin*, *141*(1), 1–5. doi:http://dx.doi.org/10.1037/bul0000007

Allison, D. B., Faith, M. S., & Franklin, R. D. (1995). Antecedent exercise in the treatment of disruptive behavior: A meta-analytic review. *Clinical Psychology: Science and Practice*, *2*(3), 279–303. Retrieved from http://onlinelibrary.wiley.com/doi/10.1111/j.1468-2850.1995.tb00045.x/full

Allison, D. B., & Gorman, B. S. (1993). Calculating effect sizes for meta-analysis: The case of the single case. *Behaviour Research and Therapy*, *31*(6), 621–31.

Allison, D. B., & Gorman, B. S. (1994). “Make things as simple as possible, but no simpler.” A rejoinder to Scruggs and Mastropieri. *Behaviour Research and Therapy*, *32*(8), 885–890. doi:10.1016/0005-7967(94)90170-8

APA Publications and Communications Board Working Group on Journal Article Reporting Standards. (2008). Reporting standards for research in psychology: why do we need them? What might they be? *The American Psychologist*, *63*(9), 839–51. doi:10.1037/0003-066X.63.9.839

Baek, E. K., Petit-Bois, M., Van den Noortgate, W., Beretvas, S. N., & Ferron, J. M. (2014). Using visual analysis to evaluate and refine multilevel models of single-case studies. *The Journal of Special Education, 50*(1), 18–26. doi:10.1177/0022466914565367

Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis. *Journal of Applied Behavior Analysis*, *1*(1), 91–7.

Barlow, D. H., Nock, M., & Hersen, M. (2008). *Single case research designs: Strategies for studying behavior change*. New York, NY: Allyn and Bacon.

Bellini, S., Peters, J. K., Benner, L., & Hopf, A. (2007). A meta-analysis of school-based social skills interventions for children with autism spectrum disorders. *Remedial and Special Education*, *28*(3), 153–162. doi:10.1177/07419325070280030401

Beretvas, S. N., & Chung, H. (2008). An evaluation of modified R2-change effect size indices for single-subject experimental designs. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 120–128. doi:10.1080/17489530802446328

Briesch, A. M., & Briesch, J. M. (2016). Meta-analysis of behavioral self-management interventions in single-case research. *School Psychology Review*, *45*(1), 3–18. doi:10.17105/SPR45-1.3-18

Busk, P. L., & Marascuilo, L. A. (1988). Autocorrelation in single-subject research: A counterargument to the myth of no autocorrelation. *Behavioral Assessment*, *10*(3), 229–242.

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

Campbell, 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*(2), 120–138. doi:10.1016/S0891-4222(03)00014-3

Campbell, J. M. (2012). Commentary on PND at 25. *Remedial and Special Education*, *34*(1), 20–25. doi:10.1177/0741932512454725

Center, B. A., Skiba, R. J., & Casey, A. (1985). A methodology for the quantitative synthesis of intra-subject design research. *The Journal of Special Education*, *19*(4), 387–400. doi:10.1177/002246698501900404

Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *Journal of Consulting and Clinical Psychology*, *66*(1), 7–18. doi:10.1037/0022-006X.66.1.7

Chambless, D. L., & Ollendick, T. H. (2001). Empirically supported psychological interventions: Controversies and evidence. *Annual Review of Psychology*, *52*, 685–716. doi:10.1146/annurev.psych.52.1.685

Conroy, M. A., Dunlap, G., Clarke, S., & Alter, P. J. (2005). A descriptive analysis of positive behavioral intervention research With young children with challenging behavior. *Topics in Early Childhood Special Education*, *25*(3), 157–166. doi:10.1177/02711214050250030301

Cooper, H. M. (1982). Scientific guidelines for conducting integrative research reviews. *Review of Educational Research*, *52*(2), 291–302. doi:10.3102/00346543052002291

Cooper, H. M. (2009). Hypotheses and problems in research synthesis. In H. M. Cooper, L. V Hedges, & J. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (2nd ed., pp. 19–35). New York, NY: Russell Sage Foundation.

Cooper, H. M. (2010). *Research Synthesis and Meta-Analysis* (4th ed.). Thousand Oaks, CA: SAGE Publications.

Cooper, H. M., Hedges, L. V, & Valentine, J. C. (2009). *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation Publications.

Cooper, H. M., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14*(2), 165–176. doi:10.1037/a0015565

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. doi:10.1177/0040059914531389

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*(4), 367–384. Retrieved from https://www.scopus.com/inward/record.uri?eid=2-s2.0-84922544346&partnerID=40&md5=6c3cba4f53576bd998cc04690741a54d

Davis, D. H., Gagné, P., Fredrick, L. D., Alberto, P. a, Waugh, R. E., & Haardörfer, R. (2013). Augmenting visual analysis in single-case research with hierarchical linear modeling. *Behavior Modification, 37*(1), 62–89. doi:10.1177/0145445512453734

Detsky, A. S., Naylor, C. D., O’Rourke, K., McGeer, A. J., & L’Abbé, K. A. (1992). Incorporating variations in the quality of individual randomized trials into meta-analysis. *Journal of Clinical Epidemiology, 45*(3), 255–265. doi:10.1016/0895-4356(92)90085-2

Dugard, P., File, P., & Todman, J. (2012). *Single-case and small-n experimental designs: A practical guide to randomization tests* (2nd ed.). New York, NY: Routledge.

Duval, S., & Tweedie, R. (2000). A nonparametric “trim and fill” method of accounting for publication bias in meta-analysis. *Journal of the American Statistical Association, 95*(449), 89. doi:10.2307/2669529

Egger, M., Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *BMJ: British Medical Journal, 315*(7109), 629–634.

Faith, M. S., Franklin, R. D., Allison, D. B., & Gorman, B. S. (1996). Meta-analysis of single-case research. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and Analysis of Single-Case Research* (pp. 245–277). Mahwah, NJ: Lawrence Erlbaum.

Ferron, J. M., Bell, B. A., Hess, M. R., Rendina-Gobioff, G., & Hibbard, S. T. (2009). Making treatment effect inferences from multiple-baseline data: The utility of multilevel modeling approaches. *Behavior Research Methods*, *41*(2), 372–84. doi:10.3758/BRM.41.2.372

Ferron, J. M., & Jones, P. K. (2006). Tests for the visual analysis of response-guided multiple-baseline data. *The Journal of Experimental Education*, *75*(1), 66–81. doi:10.3200/JEXE.75.1.66-81

Ferron, J. M., Moeyaert, M., Van den Noortgate, W., & Beretvas, S. N. (2014). Estimating causal effects from multiple-baseline studies: Implications for design and analysis. *Psychological Methods*, *19*(4), 493–510. doi:10.1037/a0037038

Fisher, W. W., & Lerman, D. C. (2014). It has been said that, “There are three degrees of falsehoods: Lies, damn lies, and statistics.” *Journal of School Psychology,* 1–6. doi:10.1016/j.jsp.2014.01.001

Fowler, C., Konrad, M., & Walker, A. (2007). Self-determination interventions’ effects on the academic performance of students with developmental disabilities. *Education and Training in Developmental Disabilities*, *42*(3), 270–285. Retrieved from http://daddcec.org/Portals/0/CEC/Autism_Disabilities/Research/Publications/Education_Training_Development_Disabilities/Full_Journals/ETDD200709V42n3.pdf#page=34

Gaskin, C. J., McVilly, K. R., & McGillivray, J. A. (2013). Initiatives to reduce the use of seclusion and restraints on people with developmental disabilities: A systematic review and quantitative synthesis. *Research in Developmental Disabilities, 34*(11), 3946–3961. doi:10.1016/j.ridd.2013.08.010

Gast, D. L., & Ledford, J. R. (2014). *Single case research methodology: Applications in special education and behavioral sciences*. New York, NY: Routledge.

Gingerich, W. J. (1984). Meta-analysis of applied time-series data. *Journal of Applied*

Behavioral Science, *20*(1), 71–79. doi:10.1177/002188638402000113

Hammond, D., & Gast, D. L. (2010). Descriptive analysis of single subject research designs: 1983-2007. *Education and Training in Autism and Developmental Disabilities*, *45*(2), 187–202.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*, 29–36. doi:10.1148/radiology.143.1.7063747

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. doi:10.1007/s40489-014-0044-3

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics*, *6*(2), 107–128. doi:10.3102/10769986006002107

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

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. doi:10.1002/jrsm.1052

Hedges, L. V, Pustejovsky, J. E., & Shadish, W. R. (2013). A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, *4*(4), 324–341. doi:10.1002/jrsm.1086

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. doi:10.1002/jrsm.5

Hembry, I., Bunuan, R., Beretvas, S. N., Ferron, J. M., & Van den Noortgate, W. (2015). Estimation of a nonlinear intervention phase trajectory for multiple-baseline design data. *The Journal of Experimental Education*, *83*(4), 514–546. doi:10.1080/00220973.2014.907231

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. doi:10.1016/j.ridd.2014.06.017

Heyvaert, M., Saenen, L., Maes, B., & Onghena, P. (2014). Systematic Review of Restraint Interventions for Challenging Behaviour Among Persons with Intellectual Disabilities: Focus on Effectiveness in Single-Case Experiments. *Journal of Applied Research in Intellectual Disabilities : JARID*. doi:10.1111/jar.12094

Higgins, J. P. T., Altman, D. G., Gotzsche, P. C., Juni, P., Moher, D., Oxman, A. D., … Sterne, J. A. C. (2011). The Cochrane Collaboration’s tool for assessing risk of bias in randomised trials. *BMJ*, *343*(oct18 2), d5928–d5928. doi:10.1136/bmj.d5928

Hitchcock, J. H., Horner, R. H., Kratochwill, T. R., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2014). The What Works Clearinghouse Single-Case Design Pilot Standards: Who Will Guard the Guards? *Remedial and Special Education*. doi:10.1177/0741932513518979

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. doi:10.1007/s10864-015-9224-1

Horner, R. H., Carr, E. G., Halle, J., McGee, G., Odom, S. L., & Wolery, M. (2005). The use of single-subject research to identify evidence-based practice in special education. *Exceptional Children*, *71*(2), 165–179. doi:10.1177/001440290507100203

Horner, R. H., Swaminathan, H., Sugai, G., & Smolkowski, K. (2012). Considerations for the systematic analysis and use of single-case research. *Education and Treatment of Children*, *35*(2), 269–290. doi:10.1353/etc.2012.0011

Huitema, B. E. (1985). Autocorrelation in applied behavior analysis: A myth. *Behavioral Assessment*, *7*(2), 107–118.

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

Huitema, B. E., & McKean, J. W. (2000). Design specification issues in time-series intervention models. *Educational and Psychological Measurement*, *60*(1), 38–58. doi:10.1177/00131640021970358

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), 0696–0701. doi:10.1371/journal.pmed.0020124

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology (Cambridge, Mass.)*, *19*(5), 640–8. doi:10.1097/EDE.0b013e31818131e7

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. doi:10.1352/0895-8017(2002)107<0212:BTOSIT>2.0.CO;2

Kazdin, A. E. (2011). *Single-Case Research Designs: Methods for Clinical and Applied Settings*. New York, NY: Oxford University Press.

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. doi:10.1177/0741932512452794

Kratochwill, T. R., & Levin, J. R. (2014). Meta- and statistical analysis of single-case intervention research data: Quantitative gifts and a wish list. *Journal of School Psychology*, 1–5. doi:10.1016/j.jsp.2014.01.003

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.

Kratochwill, T. R., & Stoiber, K. C. (2002). Evidence-based interventions in school psychology: Conceptual foundations of the Procedural and Coding Manual of Division 16 and the Society for the Study of School Psychology Task Force. *School Psychology Quarterly, 17*(4), 341–389. doi:10.1521/scpq.17.4.341.20872

Kugley, S., Wade, A., Thomas, J., Mahood, Q., Jørgensen, A.-M. K., Hammerstrøm, K., & Sathe, N. A. (2016). *Searching for studies: A guide to information retrieval for Campbell Systematic Reviews*. *Campbell Systematic Reviews* (Vol. 2016).

Lane, K. L., & Carter, E. W. (2012). Reflections on the Special Issue: Issues and Advances in the Meta-Analysis of Single-Case Research. *Remedial and Special Education*, *34*(1), 59–61. doi:10.1177/0741932512454582

Lipsey, M. W., & Wilson, D. B. (2001). *Practical Meta-Analysis*. Thousand Oaks, CA: Sage Publications, Inc.

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.

Losinski, M., Cuenca-Carlino, Y., Zablocki, M., & Teagarden, J. (2014). Examining the efficacy of self-regulated strategy development for students with emotional or behavioral disorders: A meta-analysis. *Behavioral Disorders*, *40*(1), 52–67.

Losinski, M., Sanders, S. A., & Wiseman, N. M. (2016). Examining the use of deep touch pressure to improve the educational performance of students with disabilities: A meta-analysis. *Research and Practice for Persons with Severe Disabilities*, *41*(1), 3–18. doi:10.1177/1540796915624889

Maggin, D. M., Chafouleas, S. M., Goddard, K. M., & Johnson, A. H. (2011). A systematic evaluation of token economies as a classroom management tool for students with challenging behavior. *Journal of School Psychology*, *49*(5), 529–54. doi:10.1016/j.jsp.2011.05.001

Maggin, D. M., O'Keeffe, B. V, & Johnson, A. H. (2011). A quantitative synthesis of methodology in the meta-analysis of single-subject research for students with disabilities: 1985-2009. *Exceptionality*, *19*(2), 109–135. doi:10.1080/09362835.2011.565725

Maggin, D. M., Swaminathan, H., Rogers, H. J., O'Keeffe, B. V, Sugai, G., & Horner, R. H. (2011). A generalized least squares regression approach for computing effect sizes in single-case research: Application examples. *Journal of School Psychology*, *49*(3), 301–321. doi:10.1016/j.jsp.2011.03.004

Manolov, R., & Solanas, A. (2009). Percentage of nonoverlapping corrected data. *Behavior Research Methods*, *41*(4), 1262–1271. doi:10.3758/BRM.41.4.1262

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.

Marso, D., & Shadish, W. R. (2015). Software for meta-analysis of single-case design: DHPS macro. Retrieved from http://faculty.ucmerced.edu/wshadish/software/software-meta-analysis-single-case-design/dhps-version-march-7-2015

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.

Moeyaert, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2014). From a single-level analysis to a multilevel analysis of single-case experimental designs. *Journal of School Psychology*, *52*(2), 191–211. doi:10.1016/j.jsp.2013.11.003

Moeyaert, M., Maggin, D. M., & Verkuilen, J. (2016). Reliability, validity, and usability of data extraction programs for single-case research designs. *Behavior Modification*. doi:10.1177/0145445516645763

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2013). The three-level synthesis of standardized single-subject experimental data: A monte carlo simulation study. *Multivariate Behavioral Research*, *48*(5), 719–748. doi:10.1080/00273171.2013.816621

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2014). Three-level analysis of single-case experimental data: Empirical validation. *The Journal of Experimental Education*, *82*(1), 1–21. doi:10.1080/00220973.2012.745470

Moeyaert, M., Ugille, M., Ferron, J. M., Beretvas, S. N., & Van den Noortgate, W. (2016). The misspecification of the covariance structures in multilevel models for single-case data: A monte carlo simulation study. *The Journal of Experimental Education*, *84*(3), 473–509. doi:10.1080/00220973.2015.1065216

Moeyaert, M., Ugille, M., Ferron, J. M., Onghena, P., Heyvaert, M., Beretvas, S. N., & Van den Noortgate, W. (2015). Estimating intervention effects across different types of single-subject experimental designs: Empirical illustration. *School Psychology Quarterly, 30*(1), 50–63. doi:10.1037/spq0000068

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *Annals of Internal Medicine, 151*(4), 264–269. doi:10.7326/0003-4819-151-4-200908180-00135

Odom, S. L., Brantlinger, E., Gersten, R., Horner, R. H., Thompson, B., & Harris, K. R. (2005). Research in special education: Scientific methods and evidence-based practices. *Exceptional Children*, *71*(2), 137–148. doi:10.1177/001440290507100201

Odom, S. L., Collet-Klingenberg, L., Rogers, S. J., & Hatton, D. D. (2010). Evidence-based practices in interventions for children and youth with autism spectrum disorders. *Preventing School Failure: Alternative Education for Children and Youth*, *54*(4), 275–282. doi:10.1080/10459881003785506

Owens, C. M., & Ferron, J. M. (2012). Synthesizing single-case studies: a Monte Carlo examination of a three-level meta-analytic model. *Behavior Research Methods*, *44*(3), 795–805. doi:10.3758/s13428-011-0180-y

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. doi:10.1016/j.beth.2008.10.006

Parker, R. I., & Vannest, K. J. (2012). Bottom-up analysis of single-case research designs.

Journal of Behavioral Education, *21*(3), 254–265. doi:10.1007/s10864-012-9153-1

Parker, R. I., Vannest, K. J., & Davis, J. L. (2011). Effect size in single-case research: A review of nine nonoverlap techniques. *Behavior Modification*, *35*(4), 303–22. doi:10.1177/0145445511399147

Parker, R. I., Vannest, K. J., & Davis, J. L. (2014). Non-overlap analysis for single-case research. In T. R. Kratochwill & J. R. Levin (Eds.), *Single-Case Intervention Research: Methodological and Statistical Advances* (pp. 127–151). Washington, DC: American Psychological Association. doi:10.1037/14376-005

Parker, R. I., Vannest, K. J., Davis, J. L., & Sauber, S. B. (2011). Combining nonoverlap and trend for single-case research: Tau-U. *Behavior Therapy*, *42*(2), 284–299. doi:10.1016/j.beth.2010.08.006

Petit-Bois, M., Baek, E. K., Van den Noortgate, W., Beretvas, S. N., & Ferron, J. M. (2015). The consequences of modeling autocorrelation when synthesizing single-case studies using a three-level model. *Behavior Research Methods*. doi:10.3758/s13428-015-0612-1

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

Pustejovsky, J. E. (2015b). *Operational sensitivities of non-overlap effect sizes for single-case designs*. Poster presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Pustejovsky, J. E. (2015c). scdhlm: Estimating hierarchical linear models for single-case designs. R package version 0.2.1. Retrieved from http://github.com/jepusto/scdhlm

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. doi:10.3102/1076998614547577

Randolph, J. J. (2007). Meta-Analysis of the Research on Response Cards: Effects on test achievement, quiz achievement, participation, and off-task behavior. *Journal of Positive Behavior Interventions, 9*(2), 113–128.

Reed, J. G., & Baxter, P. M. (2009). Using reference databases. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (pp. 73–101). Thousand Oaks, CA: Sage Publications.

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. doi:10.1037/14376-008

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication Bias in Meta-Analysis: Prevention, Assessment, and Adjustments* (pp. 1–7). West Sussex, England: John Wiley & Sons. doi:10.1002/0470870168

Schlosser, R. W., Lee, D. L., & Wendt, O. (2008). Application of the percentage of non-overlapping data (PND) in systematic reviews and meta-analyses: A systematic review of reporting characteristics. *Evidence-Based Communication Assessment and Intervention, 2*(3), 163–187. doi:10.1080/17489530802505412

Scruggs, T. E., & Mastropieri, M. A. (1998). Summarizing single-subject research: Issues and applications. *Behavior Modification, 22*(3), 221–242. doi:10.1177/01454455980223001

Scruggs, T. E., & Mastropieri, M. A. (2013). PND at 25: Past, present, and future trends in summarizing single-subject research. *Remedial and Special Education, 34*(1), 9–19.

doi:10.1177/0741932512440730

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

Scruggs, T. E., Mastropieri, M. A., Cook, S. . B., & Escobar, C. (1986). Early intervention for children with conduct disorders: A quantitative synthesis of single-subject research. *Behavioral Disorders*, *11*(4), 260–271.

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

Shadish, W. R., Brasil, I. C. C., Illingworth, D. A., White, K. D., Galindo, R., Nagler, E. D., & Rindskopf, D. M. (2009). Using UnGraph to extract data from image files: Verification of reliability and validity. *Behavior Research Methods*, *41*(1), 177–83. doi:10.3758/BRM.41.1.177

Shadish, W. R., 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. doi:10.1016/j.jsp.2013.11.005

Shadish, W. R., Hedges, L. V, Pustejovsky, J. E., Boyajian, J. G., Sullivan, K. J., Andrade, A., & Barrientos, J. L. (2013). A d-statistic for single-case designs that is equivalent to the usual between-groups d-statistic. *Neuropsychological Rehabilitation: An International Journal*, 1–26. doi:10.1080/09602011.2013.819021

Shadish, W. R., & Rindskopf, D. M. (2007). Methods for evidence-based practice: Quantitative synthesis of single-subject designs. *New Directions for Evaluation*, *113*(113), 95–109. doi:10.1002/ev.217

Shadish, W. R., Rindskopf, D. M., & Hedges, L. V. (2008). The state of the science in the meta-analysis of single-case experimental designs. *Evidence-Based Communication Assessment and Intervention*, *2*(3), 188–196. doi:10.1080/17489530802581603

Shadish, W. R., & Sullivan, K. J. (2011). Characteristics of single-case designs used to assess intervention effects in 2008. *Behavior Research Methods*, *43*(4), 971–980. doi:10.3758/s13428-011-0111-y

Shadish, W. R., Zelinsky, N. A. M., Vevea, J. L., & Kratochwill, T. R. (2016). A survey of publication practices of single-case design researchers when treatments have small or large effects. *Journal of Applied Behavior Analysis*, *49*(3), 1–18. doi:10.1002/jaba.308

Sham, E., & Smith, T. (2014). Publication bias in studies of an applied behavior-analytic intervention: An initial analysis. *Journal of Applied Behavior Analysis*, *47*(3), 663–678. doi:10.1002/jaba.146

Shogren, K. A., Faggella-Luby, M. N., Bae, S. J., & Wehmeyer, M. L. (2004). The effect of choice-making as an intervention for problem behavior. *Journal of Positive Behavior Interventions*, *6*(4), 228–237.

Smith, J. D. (2012). Single-case experimental designs: A systematic review of published research and current standards. *Psychological Methods*, *17*(4), 510–550. doi:10.1037/a0029312

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. doi:10.1177/0145445513510931

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology*. doi:10.1016/j.jsp.2013.12.002

Tanner-Smith, E. E., & Tipton, E. (2014). Robust variance estimation with dependent effect sizes: Practical considerations including a software tutorial in Stata and SPSS. *Research Synthesis Methods*, *5*(1), 13–30. doi:10.1002/jrsm.1091

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*(3), 375–393. doi:10.1037/met0000011

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*(6), 604–634.

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J. M., & Van den Noortgate, W. (2012). Multilevel meta-analysis of single-subject experimental designs: A simulation study. *Behavior Research Methods*. doi:10.3758/s13428-012-0213-1

Ugille, M., Moeyaert, M., Beretvas, S. N., Ferron, J. M., & Van den Noortgate, W. (2014). Bias corrections for standardized effect size estimates used with single-subject experimental designs. *The Journal of Experimental Education*, *82*(3), 358–374. doi:10.1080/00220973.2013.813366

Van den Noortgate, W., & Onghena, P. (2003a). Combining single-case experimental data using hierarchical linear models. *School Psychology Quarterly*, *18*(3), 325–346.

Van den Noortgate, W., & Onghena, P. (2003b). Hierarchical linear models for the quantitative integration of effect sizes in single-case research. *Behavior Research Methods, Instruments, & Computers*, *35*(1), 1–10.

Van den Noortgate, W., & Onghena, P. (2007). The aggregation of single-case results using hierarchical linear models. *Behavior Analyst Today, 8*(2), 196–209.

Van den Noortgate, W., & Onghena, P. (2008). A multilevel meta-analysis of single-subject experimental design studies. *Evidence-Based Communication Assessment and Intervention, 2*(3), 142–151. doi:10.1080/17489530802505362

Vargha, A., & Delaney, H. D. (2000). A critique and improvement of the "CL" common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics, 25*(2), 101–132. doi:10.2307/1165329

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika, 60*(3), 419–435. doi:10.1007/BF02294384

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

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*(2), 235–268. doi:10.1353/etc.2012.0010

White, D. M., Rusch, F. R., Kazdin, A. E., & Hartmann, D. P. (1989). Applications of meta analysis in individual-subject research. *Behavioral Assessment, 11*(3), 281–296.

White, H. (2009). Scientific communication and literature retrieval. In H. M. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (pp. 51–71). Thousand Oaks, CA: Sage Publications.

White, O. R. (1987). Some comments concerning "The quantitative synthesis of single-subject research." *Remedial and Special Education*, *8*(2), 34–39. doi:10.1177/074193258700800207

Wilson, D. B. (2009). Systematic coding. In H. M. Cooper, L. V Hedges, & J. C. Valentine (Eds.), *The Handbook of Research Synthesis and Meta-Analysis* (pp. 159–176). Thousand Oaks, CA: Sage Publications.

Wolery, M. (2012). A Commentary: Single-Case Design Technical Document of the What Works Clearinghouse. *Remedial and Special Education*, *34*(1), 39–43. doi:10.1177/0741932512468038

Wolery, M., Busick, M., Reichow, B., & Barton, E. E. (2010). Comparison of overlap methods for quantitatively synthesizing single-subject data. *The Journal of Special Education*, *44*(1), 18–28. doi:10.1177/0022466908328009

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. doi:10.1007/s10803-014-2351-z

Yoder, P. J., Bottema-Beutel, K., Woynaroski, T., Chandrasekhar, R., & Sandbank, M. (2014). Social communication intervention effects vary by dependent variable type in preschoolers with autism spectrum disorders. *Evidence-Based Communication Assessment and Intervention*, (June), 1–25. doi:10.1080/17489539.2014.917780